Algorithmic Foundations of Computational Biology
(CSCI 1820)

Sorin Istrail
January 21, 2015

SYLLABUS
Contents

1 Sequence Alignment \hspace{1cm} 3
2 Substitutions Matrices and Information Theory \hspace{1cm} 4
3 Genome Sequencing and Genome Assembly \hspace{1cm} 5
4 Gene Finding and Hidden Markov Models \hspace{1cm} 6
5 BLAST and Database Search \hspace{1cm} 7
6 Regulatory Genomics \hspace{1cm} 8
7 Suffix Trees and the Burrows-Wheeler Transform \hspace{1cm} 8
1 Sequence Alignment

Class 1
Topics covered:

1. Course Chapters
   (a) Sequence Alignment
   (b) Substitution Matrices and Information Theory
   (c) Genome Assembly
   (d) BLAST: Algorithms and Statistical Theory
   (e) Hidden Markov Models and Gene Prediction
   (f) Regulatory Genomics
   (g) Suffix Trees and the Burrows-Wheeler Transform

2. A first glance at Genome Assembly algorithms (PPT)
3. The Gold-Bug: A metaphor for Bioinformatics (PPT)

Class 2
Topics covered:

1. A first glance at Regulatory Genomics (PPT)
2. Chapter 1: Sequence Alignment – Overview
   (a) The three famous alignment papers
   (b) Local Pairwise Alignment: The Smith-Waterman Algorithm
   (c) Scoring schemes for alignment as a heuristic for maximum likelihood alignment

3. Introduction to Alignment (PPT)

Class 3
Topics covered:

1. Dynamic Programming algorithms and the principle of optimality (graph theory)
2. Interpretation of alignment score as maximum likelihood heuristic
3. Local Alignment: The Smith-Waterman Algorithm
4. Suffixes and prefixes of strings and the Suffix Tree data structure by example
5. Intro to statistical significance in sequence alignment
Class 4 Topics covered:

1. Local Alignment: The Smith-Waterman Algorithm (continued): proof of optimality
2. Waterman notation for alignment
3. Two probability problems about DNA: (1) Primers, PCR, and hybridization and (2) Herpes virus genomic fragments

Class 5 Topics covered:

1. Dijkstra’s Shortest Paths algorithm: an example
2. Dijkstra’s Shortest Paths algorithm with trace-back
3. Similarity alignment vs. distance alignment
4. The Monty Hall Problem

Class 6 Topics covered:

1. The Monty Hall Problem: Solutions
2. Substitution matrices: BLOSUM and DAYHOFF
3. How to lie with statistics: Statistics biases (continuous math) and voting theory (discrete math)
4. DAYHOFF Substitution Matrices theory: Definition and calibration of 1 PAM unit
5. Similarity vs. distance alignment: a theorem on when the optimal minimization and maximization coincide
6. A Prisoner’s Dilemma puzzle

2 Substitutions Matrices and Information Theory

Class 7 Topics covered:

1. A first look at the Karlin-Altschul statistical theory of local alignment
2. Information theory and substitution matrices
3. Substitutions Matrices, Information Theory and Statistical Theory of BLAST

Class 8 Topics covered:
1. Biological intuition for gaps inspiring gap scoring functions
2. cDNA alignment to the genome: A gap story
3. Computing the dynamic programming matrix and the alignment in linear space: The Hirshberg Algorithm

Class 9
Topics covered:
1. The Hirshberg Algorithm
2. Affine Gap alignment

Class 10
Topics covered:
1. Local alignment and phase transition: intuition for alignment parameter values choices
2. The Hirshberg Algorithm - in depth and time complexity analysis
3. Affine Gap alignment

Class 11
Topics covered:
1. Relative entropy and PAM matrices

3 Genome Sequencing and Genome Assembly

Class 12
Topics covered:
1. Introduction to Genome Assembly – The Sequence of the Human Genome – an overview of the Celera Assembly algorithm and software engineering aspects

Class 13 Topics covered:
1. Fundamental Assembly Algorithms: The Idury-Waterman Algorithm and Eulerian Graphs
2. Poisson statistics of k-mers

Class 14
Topics covered:
1. Fundamental Assembly Algorithms: The Idury-Waterman Algorithm II
Class 15 Topics covered:

1. Fundamental Assembly Algorithms: The Idury-Waterman Algorithm III
   1. PHRAP
   2. Tigr Assembly
   3. Celera Assembler
   4. Arachne Assembler

Class 16 Topics covered:

1. Statistical Theory of Genome Assembly: The Lander-Waterman Formulas
2. Three fundamental problems and the derivation of their solution formulas
   (a) Problem 1: What is the mean portion of the genome covered by the contigs of the assembly?
   (b) Problem 2: What is the mean number of contigs of the assembly?
   (c) Problem 3: What is the mean contig size of the assembly?

Class 17 Topics covered:

1. Statistical Theory of Genome Assembly: The Lander-Waterman Formulas II

Class 18 Topics covered:

1. Statistical Theory of Genome Assembly: The Lander-Waterman Formulas III
2. Applications to Sanger sequencing projects, and NGS projects

4 Gene Finding and Hidden Markov Models

Class 19 Topics covered:

1. An overview of the DNA Structure of a Gene
(a) Problem 1: THE MODEL EVALUATION PROBLEM
(b) Problem 2: THE DECODING/REVEALING "THE HIDDEN" PROBLEM
(c) Problem 3: THE LEARNING PROBLEM

Class 20 Topics covered:
1. Problem 1: THE MODEL EVALUATION PROBLEM; ALGORITHMIC SOLUTION TO PROBLEM 1 - the Forward Algorithm.
2. Problem 2: THE DECODING/REVEALING "THE HIDDEN" PROBLEM; ALGORITHMIC SOLUTION TO PROBLEM 2 - The Viterbi Algorithm.

Class 21 Topics covered:
2. CpG islands and an algorithm for finding them
3. Baum’s Q function and faster convergence in the EM algorithm
4. an informal proof of local optimality using Lagrange multipliers for the EM algorithm

5  BLAST and Database Search

Class 23 Topics covered:
1. BLAST and database searches
2. The BLAST statistical theory of Karlin and Altschul

Class 24 Topics covered:
1. The BLAST algorithm phases: Seeding, Extension and Evaluation
2. Probability versus Expectation: E-values and P-values
3. BLAST and Finite automata and data structures for BLAST hits

Class 25
1. BLAST statistical theory: Sequential analysis and random walks
2. Random walks and difference equations and their solutions
3. Random walks and difference equations and their solutions II
4. The concept of support for statistical testing
5. The sequential probability ratio test

Class 26
1. Ladder points, excursions and HSTs
2. The test statistics of BLAST
3. Karlin-Astschul equation
4. P-value and E-value in BLAST

6 Regulatory Genomics

Class 27 Topics covered:
1. Motifs and cis-regulatory sequence structure
2. Position Weight Matrices algorithms and statistical theory

7 Suffix Trees and the Burrows-Wheeler Transform

Class 28
Topics covered:
1. A first glance at algorithms for data compression
2. The Suffix Tree construction in linear time
3. The Burrows-Wheeler transform