Algorithmic Foundations of Computational Biology

CIS On-line Research Program 2019

Professor Sorin Istrail
Department of Computer Science, Brown University, USA

1 Course Description

The aim of this course is to provide an introduction to Computational Molecular Biology. The course is organized into four chapters: (1) Sequence Alignment, (2) Combinatorial Pattern Matching, (3) Phylogenetics Trees, and (4) Machine Learning Methods - Hidden Markov Models. Each chapter is devoted to a class of basic computational problems related to the analysis of DNA, RNA and protein sequences and their molecular biology function. Our journey in each chapter is driven by a set of most beautiful algorithms. "Beautiful" algorithm here refers to an algorithm that is rigorous, practical and with elegant simplicity that makes it also easy to implement. These algorithms are among those presenting the state of the art of the theory and practice of solving the computational problems presented in the corresponding chapter. In addition to the beautiful algorithms, each chapter contains a *Foundations* section that presents in detail the biological problems discussed and theoretical computer science and statatistical results that led to the invention of the algorithms that resolve the modeled biological problems. The algorithms are presetned together with their underlying data structures, mathematical analysis of their performace, mathematical puzzles that highlight the computational challanges, and at times, the exciting story of the researchers quest for algorithm optimality (speed). The overall work in the class will help in providing a comprehensive introduction to the field for both potential concetrators and those who may take only a single course.

2 Lecture Topics

1. Chapter 1: **Sequence Alignment**
   Algorithms:
   - Needleman-Wunsch Algorithm (global alignment)
   - Smith-Waterman Algorithm (local alignment)
   - De Brujin Genome Assembly Algorithms (de Brujin graphs and Eulerian paths)

Foundations: Margaret Dayhoff the *"mother and father of Bioinformatics"* – pioneer of statistical methods, similarity matrices statistics (intro); dynamic programming; protein structure alignment as gold standard for sequence alignment; k-mers and Poisson statistics; DNA, RNA and protein sequence alignment; gaps in alignment; multiple alignment, graph theory, information theory. Open problems.

2. Chapter 2: **Combinatorial Pattern Matching**
   Algorithms:

   - Knuth-Morris-Pratt Algorithm (finding a string pattern in a text)
   - BLAST Algorithm
   - Weiner Algorithm (Position Trees and Suffix Trees)

   Foundations: Eric Davidson – master of the universe of the developmental gene regulatory networks, sea urchin, and "the regulatory genome and the computer"; transcription factors and their DNA binding sites motifs; the transcriptome of the sea urchin embryo; finite-automata and regular expressions; approximate string matching; patterns in DNA, RNA and protein sequences. Open problems.

3. Chapter 3: **Phylogenetic Trees**
   Algorithms:

   - UPGMA Algorithm (evolutionary distance matrices with uniform clock)
   - Neighbour-Joining Algorithm (general evolutionary distance matrices)

   Foundations: Ronald Fisher – "the Einstein of Statistical Science" pioneer of mathematical models of evolution, *"Nothing in Biology Makes Sense Except in the Light of Evolution,"* (Theodosius Dobzhansky, 1973); tree (multiple) alignment, maximum likelihood phylogeny and other probabilistic models. Open problems.

4. Chapter 4: **Machine Learning Methods: Hidden Markov Models (HMMs)**
   Algorithms:

   - Forward Algorithm for PB 1. "Computing the probability" (model scoring)
   - Viterbi Algorithm for PB 2. "Best Explanation" (Viterbi maximum likelihood)

   Foundations: Andrew Viterbi – pioneer of coding and decoding theory algorithms; probabilistic finite automata; finding genes in genomes.

5. **Good textbooks:**

   (a) Michael Waterman, *Introduction to Computational Biology: Maps, Sequences, and Genomes*, Chapman and Hall, 1995

   (b) Richard Durbin, Sean Eddy, Anders Krogh, and Graeme Mitchison, *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids*, Cambridge University Press, 2012

(c) Neil Jones and Pavel Pevzner, *An Introduction to Bioinforamtics Algorithms*, MIT Press, 2004

(d) Dan Gusfield, *Algorithms on Strings, Trees and Sequences*, Cambridge University Press, 1997

(e) Marketa J Zvelebil, Jeremy O. Baum, *Understanding Bioinforamatics*, Garland Science, 2007

(f) Maxime Crochemore, Wojciech Rytter, *Jewels of Stringology: Text Algorithms*, World Scientific, 2003

(g) Alberto Apostolico, Zvi Galil, *Pattern Matching Algorithms*, Oxford University Press, 1997

(h) Barry Hall, *Phylogenetic Trees Made Easy: A How-To Manual*, Sinauer Associates, 2008

(i) Roderic Page, Edward Holmes, *Molecular Evolution: A Phylogenetic Approach*, Blackwell Science, 1998

(j) Joseph Felsenstein, *Inferring Phylogenies*, Sinauer Associates, 2004

(k) Masatoshi Nei, Sudhir Kumar, *Molecular Evolution and Phylogenetics*, Oxford University Press, 2000

(l) M. Vidyasagar, *Hidden Markov Processes: Theory and Applications to Biology*, Princeton University Press, 2014

(m) Lawrence Rabiner, *A Tutorial on Hidden Markov Models and Selected Applications to Speech Recognition*, Proceedings of the IEEE, vol 77, No. 2, 1989

6. **Three Team Research Projects: Team DNA, Team RNA, Team PROTEINS**

(a) **Team DNA**: *Project: The Human Genome and Human Disease*
Pioneers of the field:

- Ronald Fisher
- Craig Venter
- Edsger Dijskstra

(b) **Team RNA**: *Project:The Regulatory Genome and the Computer*
Pioneers of the field:

- John von Neumann
- Eric Davidson

(c) **Team PROTEINS**: *Project: The Protein Folding Problem*
Pioneers of the field:

- Christian Anfinsen, Nobel Laureate
- Ken Dill