Syllabus: CS 5124
Algorithms in Bioinformatics
Fall, 2004

`http://courses.cs.vt.edu/~algnbio/index.php`

Instructor: Lenwood Heath

- **Office:** 2160J Torgersen Hall
- **Office Hours:** 9:30–11 AM Tuesdays and Thursdays; 9-10:30AM Wednesdays
- **Email:** `heath@vt.edu`

Graduate Teaching Assistant: Raghavendra Nyamagoudar

- **Office:** 133 McBryde Hall
- **Office Hours:** To be announced on the course web site
- **Email:** `raghavgn@vt.edu`

Class Meets: McBryde 226, 11:15-12:05, MWF

Exams

Midterm Exam	Monday, October 18, 11:15–12:05
Final Exam	Monday, December 13, 10:05-12:05

Index Number: 91475

Prerequisites:

- Data Structures (CS 2604) required
- CS 4104, Data and Algorithm Analysis, highly desirable
- **Corequisite:** PPWS 5314 — Biological Paradigms for Bioinformatics — or equivalent coursework in genetics and molecular cell biology

Required Textbook:

Algorithms on Strings, Trees, and Sequences: Computer Science and Computational Biology, Dan Gusfield, Cambridge University Press, 1997.

On Reserve:

For current list, see class web site.

Description

This course emphasizes algorithms to solve problems found in biology, especially molecular biology. A variety of current problems in computational molecular biology will be introduced, investigated, analyzed for computational complexity, and solved with efficient algorithms, when feasible. A number of such problems will be shown to be NP-complete or other evidence of their difficulty will be presented.

Grading Policy

Grading for the course is on a 1000-point scale, with the points distributed as follows:

Homework assignments: 12 at about 50 points each	600
Midterm exam: October 18, 11:15–12:05	100
Final exam: December 13, 10:05-12:05	300

A typical homework assignment consists of 2 or 3 problems or exercises, posted on the web site. All homework must be prepared with LaTeX or other word processing system and submitted as a stapled printout to a box outside the instructor's office (2160J Torgersen Hall). Homework is due at 4:00 PM on the due date (see course calendar). **No late homework will be accepted.**

Ethics

The Honor Code applies. All work submitted must be the student's own work. Students may solicit help only from the instructor or the GTA.

Announcement

If any student needs special accommodations because of a disability, please contact the instructor during the first week of classes.

Intended Topics and Approximate Order

SOURCE[1]	**TOPIC**
	Course overview
Gusfield 1	Exact matching: first algorithms; fundamental preprocessing
Gusfield 2.1–2.3	Exact matching: classical algorithms; Boyer-Moore and Knuth-Morris-Pratt
Gusfield 3.4	Exact set matching; keyword trees
Gusfield 5	Suffix trees
Gusfield 6.1–6.2	Linear-time construction of suffix trees
Gusfield 7.2, 7.4–5, 7.11-12	Selected applications of suffix trees; exact set matching again, longest common substring, DNA contamination, and finding repeats
Gusfield 10	The importance of sequence comparison in molecular biology
Gusfield 11	Core string edits, alignments, and dynamic programming
Gusfield 14; Durbin, et al., 6	Multiple string comparison and multiple sequence alignment
Papers from the literature; Durbin, et al., 2.2, 2.7–2.8	Probability in bioinformatics; the statistical basis for scoring matrices; PAM and BLOSUM matrices
Gusfield 15; Durbin, et al., 2.3–2.6	Sequence databases and searching — BLAST, PSI-BLAST, and FASTA
Durbin, et al., 3–5	Hidden Markov models in bioinformatics
Gusfield 17; Durbin, et al., 7–8	Evolutionary or phylogenetic trees; survey of algorithms for constructing phylogenetic trees; bootstrapping
Gusfield 16	Selected sections on mapping and sequencing, if there is time

End of Syllabus

[1] A variety of sources other than the textbook will be employed. Some research papers will be made available on the course web site, but a number of books will be cited as well. Durbin, et al., is *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids,* Richard Durbin, Sean Eddy, Anders Krogh, and Graeme Mitchison. Cambridge University Press, 1998.