# CMSC 423: Bioinformatics

Fall 2011

**Course objectives:** Study interesting algorithms and methods for the analysis of biological data. We will cover string matching algorithms, string searching, string pattern finding (e.g. gene finding, discovery of protein binding sites), genome assembly, phylogenetics, protein structure prediction, and several topics of current research interest in bioinformatics.

**Professor:** Carl Kingsford, Office: CBCB 3113. Email: `carlkcs.umd.edu`. Office hours are posted on the course webpage. If you cannot attend office hours at this time, email me about scheduling a different time.

**Web page:** `http://www.cs.umd.edu/class/fall2011/cmsc423/`. Grades will be posted on `http://grades.cs.umd.edu/`.

**Class time:** Tue/Thr 12:30-1:45pm in CSIC 2107.

**Textbook:** Unfortunately, because bioinformatics is a broad, new, and raplidly changing field, there is no great textbook. The material in will be presented in slides and lecture notes.

**Course work:** There will be ∼ 6 homework sets of 3–5 problems each. There will be an in-class midterm and a comprehensive final. There will be a two part project. Approximate grading weight: 20% for homeworks, 30% for the midterm, 30% for the final, and 20% for the project. The class will be graded on a curve.

**Homework policies:**

- Written problem sets are due at the start of class. **No late homework will be accepted** — turn in what you have completed. If you will miss class, turn in the homework early.
- Answers to homework problems should be written concisely and clearly. **Messy or poorly written homeworks will not be graded.** Typesetting homeworks with LaTeX is encouraged (but not required).
- Homework problems that ask for an algorithm should present: a clear English description or pseudocode of the algorithm, a convincing argument for why the algorithm is correct, and an estimate of the running time.
- Graded homeworks should be picked up in class; if you miss the class when the homework is returned, please pick it up during office hours.
- Regrade requests should be made in writing within 1 week of the homework being returned.
- You may discuss homework problems with classmates. **You must list the names of the class members with whom you worked at the top of your homework. You must write up your own solution independently!**

**Exam policies:** Exams and the final will be closed book, closed note. The final exam will be in-class at the time set by the official university exam schedule.

**Project policies:** Projects will be completed in small groups and will involve programming in Java. **You may NOT copy or give code to other groups.** Providing code and using code from other groups are both academic integrity violations that generally receive the same punishment. **You cannot incorporate code from the internet into your projects.** Submitted projects may be checked automatically for inappropriate code use. More details about the projects will be available in a few weeks.

The maximum possible score of a project will be reduced by 10% for every day it is late. After 5 late days, the project will no longer be accepted.

*The fine print:*

**Excused absences:** Any student who needs to be excused for an absence from a single lecture, recitation, or lab due to a medically necessitated absence shall: (a) Make a reasonable attempt to inform the instructor of his/her illness prior to the class. (b) Upon returning to the class, present their instructor with a self-signed note attesting to the date of their illness. Each note must contain an acknowledgment by the student that the information provided is true and correct. Providing false information to University officials is prohibited under Part 9(h) of the Code of Student Conduct (V-1.00(B) University of Maryland Code of Student Conduct) and may result in disciplinary action. (c) This self-documentation may not be used for the Major Scheduled Grading Events as defined below and it may only be used for only 1 class meeting during the semester.

Any student who needs to be excused for a prolonged absence (2 or more consecutive class meetings) or for a Major Scheduled Grading Event must provide written documentation of the illness from the Health Center or from an outside health care provider. This documentation must verify dates of treatment and indicate the timeframe that the student was unable to meet academic responsibilities. No diagnostic information needs to be given. The Major Scheduled Grading Events for this course include: (a) Midterm: October 20 during the lecture period and (b) The final exam, as scheduled by the offical university exam schedule. Additional Major Grading Events can be added with 2 weeks notice.

Absences for religious observances must be submitted in writing to the instructor within two weeks of the start of the semester. The instructor is not under obligation to offer a substitute assignment or to give a student a make-up assessment unless the failure to perform was due to an excused absence. An excused absence for an individual typically does not translate into an extension for team deliverables on a project.

**Academic accommodations:** Any student eligible for and requesting reasonable academic accommodations due to a disability is requested to provide, to the instructor in office hours, a letter of accommodation from the Office of Disability Support Services (DSS) within the first two weeks of the semester.

**Course evaluations:** At the end of the semester, please fill out a course evaluation at `http://www.courseevalum.umd.edu`. Course evaluations are read and taken seriously.

**Academic honesty:** All class work should be done independently unless explicitly indicated on the assignment handout. You may discuss homework problems with classmates, but you must write your solution by yourself. If you do discuss assignments with other classmates, you must supply their names at the top of your homework. Projects may be completed in teams as specified on the project handout.

No excuses will be accepted for copying others work (from the current or past semesters), and violations will be dealt with harshly. Every year, many CS students are referred to the honor board, which is an unpleasant experience for everyone and can seriously impact plans for graduate school, graduation, etc. Getting a bad grade is much preferable to cheating.

To quote the honor council: "The University of Maryland, College Park has a nationally recognized Code of Academic Integrity, administered by the Student Honor Council. This Code sets standards for academic integrity at Maryland for all undergraduate and graduate students. As a student you are responsible for upholding these standards for this course. It is very important for you to be aware of the consequences of cheating, fabrication, facilitation, and plagiarism. For more information on the Code of Academic Integrity or the Student Honor Council, please visit `http://www.shc.umd.edu`.

To further exhibit your commitment to academic integrity, remember to sign the Honor Pledge on all examinations and assignments: 'I pledge on my honor that I have not given or received any unauthorized assistance on this examination (assignment).'"

## Tentative Schedule

*Sequence Comparison & Dynamic Programming* (3 weeks)

- Dynamic programming
- Longest common subsequence
- Sequence alignment (local, global, semiglobal)
- Space efficient sequence alignment
- Multiple sequence alignment
- RNA folding

*Sequence Search & String Data Structures* (2 weeks)

- Fast string searching algorithms
- Suffix trees
- Suffix arrays
- Burrows-Wheeler transform

*Pattern Finding with Hidden Markov Models & EM/Gibbs Sampling* (2 weeks)

- Hidden Markov models
- HMMs for gene finding
- HMMs for motif-finding
- EM/Gibbs sampling for motif finding

*Gene Expression & Clustering* (1.5 weeks)

- Gene expression matrices
- K-means clustering
- Gene association studies, genotyping, SNPs

*Phylogenetics* (2 weeks)

- Building evolutionary trees
- Neighbor-joining / UPGMA
- Fitch's algorithm
- Maximum likelihood / parsimonious trees
- Genome rearrangements

*Protein Structure* (1.5 weeks)

- Protein structure prediction
- Secondary structure prediction
- Side-chain positioning
- Threading

*Current Research Topics* (1.5 weeks)

- The shape of the genome
- Network alignment
- Genome assembly
- ...