# Course Sysllabus
# CS 238: Algorithmic Techniques in Computational Biology
# Spring, 2020

**Introduction:** Due to the revolutionary development of genomics, the sheer volume of digital genetic information now available is surpassed only by the potential for biological and medical discovery. The exploration of this information is critically dependent upon the development of advanced computational methods for data analysis. From this dependency, a new field of research, *Computational Molecular Biology* (or simply *Computational Biology*), emerged in recent decades.

The primary reason that molecular biology is of great interest to computer scientists is that genes, proteins, chromosomes, and genomes can all be viewed, at one level, as simply strings of symbols from a finite alphabet (the alphabet $\{C, G, A, T\}$ of the four nucleotides or the alphabet of 20 amino acids). At this level, they are similar to textual documents with a different alphabet; thus, many of the techniques of "stringology" are applicable to them (indeed, genomic issues were responsible for much of the development of this subfield of Computer Science). The analogy can be carried at least one step further. Just as a document, such as this syllabus, has a structure, so too does a chromosome; a chromosome is a sequence of genes and noncoding regions and a gene is a sequence of exons and introns. A second reason for computer scientists' interest is that various completed (or ongoing) large-scale sequencing, mapping and profiling projects projects in life sciences such as the Human Genome Project, ENCODE, modENCODE, HapMap Project, 1000 Genome Project, Human Microbiome Propject, etc. have produced an enormous amount of digital data and have raised many *intractable* computational questions of an optimization flavor. While many biologists have intuitively realized that a problem such as multiple sequence alignment or phylogenetic inference is intractable due to its combinatorial nature and, therefore, developed heuristic algorithms, few proofs of intractability and even fewer proofs of the quality of the results of the heuristic algorithms have been established for these combinatorial optimization problems.

**Course format:** The course will be focused on the design and analysis of efficient (combinatorial) algorithms for important problems in computational molecular biology. The format of the course will include lectures by the instructior, class discussion, directed reading, and student presentations or projects. The exact format will depend on the size of enrolment and student background. We emphasize mathematics, algorithms, and data structures instead of biological implications and applications, although some relevant biological background and motivations will be discussed. We may also have some guest speakers to talk about their research problems.

**Schedule:** MWF 1:00-1:50pm, MSE 113; but the class will be moved to Zoom with meeting ID 890 757 141 at least initially. Register in advance for lectures on Zoom:
https://ucr.zoom.us/meeting/register/tZ0tcOyvrz8rxuxWypENK32sh9AaTz0ykA

**Textbooks:** *An Introduction to Bioinformatics Algorithms* (required), Neil C. Jones and Pavel Pevzner, The *MIT Press Series on Computational Molecular Biology*, 2004, 4th Ed. ISBN 0-262-10106-8. Available at the Bookstore with the same discount as on Amazon.

*Bioinformatics Algorithms: An Active Learning Approach* (optional), Phillip Compeau and Pavel Pevzner, Active Learning Publishers, 2018. ISBN: 978-0-9903746-3-3

**Lecture Notes:** The slides used in class (in PPT or PDF or web formats) as well as some supplementary material can be found on the class homepage `https://www.cs.ucr.edu/~jiang/238-homepage.html`

**Optional reference books:**

*Genome-Scale Algorithm Design*, Veli Makinen, Djamal Belazzougui, Fabio Cunial and Alexandru I. Tomescu. Cambridge University Press, 2015. ISBN 978-1-107-07853-6

*Algorithms for Strings, Trees, and Sequences: Computer Science and Computational Biology*, Dan Gusfield, Cambridge University Press, 1997. ISBN: 0-521-58519-8

*Fundamental Concepts of Bioinformatics*, Dan Krane and Michael Raymer, Benjamin Cummings, 2003. ISBN: 0-8053-4722-4

*Rosalind*: An online platform for learning bioinformatics through problem solving. See `www.rosalind.info`.

**Instructor:** Tao Jiang, WCH 336, phone: x22991, email: `jiang@cs.ucr.edu`. Office hours: MW 2-3pm. Zoom meeting ID 886-998-0294.

**TA:** Xizhe Yin, email: `xyin014@ucr.edu`. Office hours: Tu 1-2pm. Zoom meeting ID 426-417-7502.

**Prerequisite:** CS218 (Design and Analysis of Algorithms) **or** CS141 (Algorithms and Data Structures) **and** CS/Math 111 (Discrete Math), **or** equivalent knowledge. No background in biology is required.

**Topics covered:** introduction to molecular biology (2 lectures), physical (restriction) mapping (2 lectures), motif finding, regulatory signal recognition, and probabilistic algorithms (3 lectures), genome rearrangement by greedy and approximation methods (3 lectures), sequence alignment by dynamic programming and divide-and-conquer methods (3 lectures), multiple sequence alignment (2 lectures), gene prediction (2 lectures), reconstruction of evolutionary trees (3 lectures), fragment assembly by graph methods and shortest common superstrings (1 lecture), gene expression analysis and clustering methods (1 lecture), and selected topics (*e.g.* haplotype inference, comparative genomics). The actual topics may change according to progress.

**Homework assignment:** There will be three assignments to help digest the material learned in lectures. Most of the assignments will involve analytical work (*i.e.*, design and analysis of algorithms), but some may require serious programming effort.

**Term presentation:** Each student is required to give a presentation on a topic selected from a list of topics provided by the instructor. Examples of possible topics include follow-up discussion on a topic covered in lectures, survey on something not covered in lectures, original solution of a technical problem posed in class, nontrivial improvement on a known result, proposal of new methods, practical considerations of theoretical results, etc.

**Reading assignment:** The students will be expected to review, in advance, the material to be covered in class. In addition to the text and reference books, there will be handouts of papers and book chapters from time to time.

**Grading:** Homeworks 50%, term presentation 40%, and class participation 10%.

**Academic dishonesty:** You are basically expected to work alone on your assignments and presentaton. However, it is a common practice to rehearse your presentation in front of friends and classmates and obtain their feedback. For a detailed UCRpolicy on academic dishonesty, see `https://conduct.ucr.edu/policies/academic-integrity-policies-and-procedures`.

**Class Mailing List:** Please subscribe to the class mailing through the class homepage ASAP and remember to confirm the subscription.