Bioinformatics Algorithms
CS 579, 3 Credits, Fall 2021
Tues, Thurs 1:30 - 2:45 P.M., Rec 122

Prof. Alex Pothen
Dept. of Computer Science
apothen at purdue dot edu

This course is intended to be a first course in Bioinformatics algorithms for graduate students in computer science, computer engineering, mathematics, statistics, computational life sciences, and related fields. This is suitable as a foundational course for students planning to do research in bioinformatics with a focus on algorithms; it is also suitable for those who wish to acquire some breadth in algorithms during graduate study and learn about important recent developments in bioinformatics.

New high through-put technologies make it possible to study the genome (all the genes of an organism) and the proteome (all of the proteins in an organism), making biology an information-rich science. New technologies such as RNA-Seq measure quantitatively gene expression, and metagenomics is able to identify genes from multiple organisms present in an environment. Computational algorithms and software tools are critical to process the massive, error-plagued, data to understand the processes of life at the molecular level. This increased understanding advances life sciences, and has practical benefits too: personalized medicine (the design of drugs sensitive to one's genetic make-up), the engineering of microorganisms for industrial uses, an understanding of how all organisms are related, etc. Bioinformatics or computational biology is the moniker for this area of research, and it involves the development of string algorithms, combinatorial algorithms, databases, statistical methods, high performance computing, and software.

Purdue has made large investments in two life sciences pillars of excellence, PI4D, in the Purdue Institute in Inflammation, Immunology, and Infectious Diseases; and another in neurosciences. Purdue is also investing in a Big Data in life sciences initiative. In all of these computer science algorithms, software, databases, visualization, etc. should play an important role, and these provide exciting research opportunities for students and faculty.

A National Science Foundation study on Simulations based Engineering and Science (SBES) states that the shortage of trained students is the biggest bottleneck to progress in bioinformatics and computational biology. Students who have a working knowledge of bioinformatics concepts should find that their employment prospects are improved by their skills in bioinformatics. There are many excellent opportunities in academia, industry, and the national labs available to students who brave the challenge.

Students from computer science, the engineering disciplines, and mathematical sciences may be concerned about how much previous exposure to biological sciences they need to study bioinformatics. For this introductory course the answer is, not much—students do not need any previous knowledge of bioinformatics or biology beyond high school. *But you will need a willingness to learn biological concepts needed to understand the computational problems we study.* We will begin with a quick review of some of the biological concepts needed for bioinformatics, and will learn more as needed. Pre-requisites include at least an undergraduate course in algorithms (a graduate course in algorithms is preferred). You will also need some programming experience. We will program in Python and use functions available in Python.

What about students from the biological sciences? The same pre-requisites apply; hence you would benefit the most from this course if you have the computational and mathematical knowledge corresponding to an undergraduate algorithms course. However, I make up a separate set of HW problems for students from the life sciences who do not have the algorithmic background. You will be graded in the course based on home work and a project. In case of questions, please talk to me before registering for the course.

The programming assignments will involve Python. You could complete many of these problems using the Matlab Bioinformatics toolbox, or other software. But I would recommend that you complete your assignments in Python.

The course will provide an introduction to the basic techniques and algorithms in bioinformatics. The topics discussed will include some of the following:

- Biological Sequences: pairwise global alignments of genes and proteins;
  pairwise local alignment of genes and proteins;
  scoring matrices;
  multiple alignment of proteins

- Database search for sequences
  BLAST and variants

- Whole Genome Alignment: suffix trees and algorithms

- Genome Assembly

- Phylogenetic Trees and networks:
  distance based, parsimony, and maximum likelihood methods.

- Algorithms for analyzing RNA-Seq data

- Overview of Systems Biology and Biological Networks
  Clustering and module discovery in biological networks. Network alignment

- Metagenomics

There are two ways to take this course.

Students in computer science and related fields will be graded on regular homework problems, programming assignments, and two in-class exams: a midterm exam and a final exam. Students in the life sciences will be graded on the basis of homework problems, programming assignments, and a class project, for which students would need to write a report and present the results in class. The project will require you to work on a problem to be decided after discussion with me, and you will need to submit regular reports at about two week intervals. You can expect to put in about five hours of work per week for ten to twelve weeks of the semester on the project.

Students will find the first of the books listed below accessible for the material in the first half of the course. The first book focuses more on algorithms, but the writing style (especially the proofs) can be hard to follow in places. The second book (now available in two volumes) is the basis for a Bioinformatics Algorithms course taught through Coursera by Compeau and Pevzner. This is a good source for self-study and for open-ended projects which introduce research problems. The third book is focused on students in the life sciences, and emphasizes the use of bioinformatics databases and software tools to solve problems.This is the book I would recommend to students who are in the biological sciences.

Other lectures will be based on material from some of the books listed below and from the research literature. The last book is a brief introduction to the molecular biology background needed for a mathematical or computational scientist who wishes to read the research literature in the biological sciences.

Wing-Kin Sung, *Algorithms in Bioinformatics: A Practical Introduction*, Chapman and Hall/CRC Press, 2009. (ISBN 978-1-420070330) **Required for students with an algorithmic background.** An e-version of the book is available.

Phillip Compeau and Pavel Pevzner, *Bioinformatics Algorithms: An Active Learning Approach*, Active Learning Publishers, third edition, 2018. (ISBN 978-0990374633) **Optional.** The first five chapters of this book are available on line at https://www.bioinformaticsalgorithms.org/

Caroline St. Clair and Jonathan E. Visick, *Exploring Bioinformatics: A project-based approach*, Second edition, Jones and Bartlett Learning, 2015. (ISBN 978-1-284-02344-2) **Recommended for students who do not have an algorithmic background.**

Dan Gusfield, *Algorithms on Strings, Trees, and Sequences*, Cambridge University Press, 1997. **Optional.**

Tandy Warnow, *Computational Phylogenetics: An introduction to designing methods for phylogeny estimation*, Cambridge University Press, 2017. **Optional.**

Peter Clote and Rolf Backhofen, *Computational Molecular Biology: An Introduction*, John Wiley, 2000. **Optional.**

Ingvar Eidhammer, Inge Jonassen, and William R. Taylor, *Protein Bioinformatics: An Algorithmic Approach to Sequence and Structure Analysis*, John Wiley and Sons, 2004. (ISBN: 978-0470848395) **Optional.**

Srinivas Aluru, *Handbook of Computational Molecular Biology*, Chapman and Hall/CRC, 2006. **Optional.**

Lawrence E. Hunter, *The Processes of Life: An Introduction to Molecular Biology*, MIT Press, 2009. **Optional.**