NEW for FALL 2017

CS 189A Special topics course:

Integer Linear Programming in Computational Biology: An entry-level "how to" and "why do"

T,Th 12:10 - 1:30 in Olson 205
Disc. Th: 4:10 - 5:00
CRN: 62337
4 units

Instructor: Dan Gusfield

Integer (Linear) Programming, abbreviated "ILP", is a versatile modeling and optimization technique that has been increasingly used in *computational biology* in *non-traditional* ways. ILP is often very effective in solving *instances* of hard computational problems in biology on *realistic* data of current importance, despite the fact that many of those problems lack general algorithmic solutions that are efficient (in a provable, worst-case sense).

Highly engineered, commercial ILP solvers are available (previously very expensive, but now free to academics and researchers) to solve ILP formulations. The improvement of the best solvers has been *spectacular*, with an estimate that (combined with faster computers) benchmark ILP problems can now be solved 200-*billion* times faster than twenty-five years ago. Exploiting ILP, some biological problems of importance can be modeled in a way that allows a solution in seconds on a laptop, while more commonly-used statistical models require days, weeks or months of computation on large clusters.

The effectiveness of the best ILP solvers on problem *instances* of importance in biology opens huge opportunities. The impact of cheaper, faster and easier-to-implement computation could be truly transformative in several subareas of biology.

The course, the first of its kind anywhere, teaches integer programming through the lens of computational biology. It is oriented towards practice with very little theory. The only mathematics required is elementary high-school algebra, although good quantitative reasoning skills are important. Familiarity with some computer programming language (for example, Python) is helpful, but a motivated student without that background can learn enough of it in the class.

The class will be a mix of lectures and hands-on labs, and a term project using the integer programming solvers from Gurobi Optimization and IBM Cplex. Illustrations and problems will be selected from a wide range of areas in biology and types of data analysis. Possibilities include: Clustering and bi-clustering; Building Phylogenetic Trees, RNA folding and Protein structure analysis; Biological network analysis; Co-phylogeny analysis for host-parasite interaction; Maximum-likelihood pedigree reconstruction; use of ILP in the study of cancer development; Analysis of brain connections; Haplotyping in populations and in genome assembly; Genetic applications in diabetes II; Genome sequence assembly; Genomic sequence analysis; Applications in ecology and evolution. We may also develop frivolous non-biological applications, such as an ILP that can almost instantly solve huge Sudoku puzzles, and other annoying logic-game puzzles.

I have developed the course and course materials in the hopes of reaching biology students, but only very basic biological background is needed, and all biological applications will be explained in full. Students from outside of Biology (e.g., CS, Math, Statistics etc.) are also welcome; and the course is appropriate for students who want to learn about ILP, for any application.

Development of this course has been partially funded by a grant from the National Science Foundation.