AdvancedBioinformatics

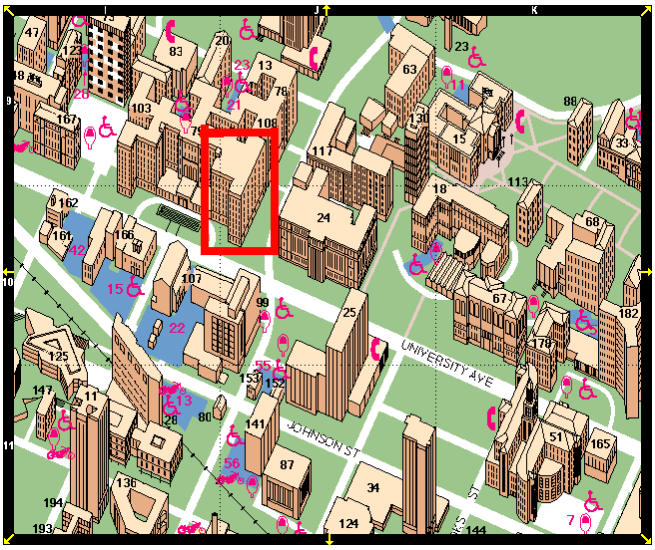
Biostatistics & Medical Informatics 776
Computer Sciences 776
Spring 2002

Mark Craven
Dept. of Biostatistics & Medical Informatics
Dept. of Computer Sciences
craven@biostat.wisc.edu
www.biostat.wisc.edu/~craven/776.html

BSMI/CS 776: Bioinformatics

- Instructor: Prof. Mark Craven
  - craven@biostat.wisc.edu *or*
  - craven@cs.wisc.edu
- Office hours: 2:00-3:00 Tues, 2:30-3:30pm Wed, or by appointment
  - room 6730, Medical Sciences Center
- Course home page: www.biostat.wisc.edu/~craven/776.html
- Course mailing list: TBA

Finding My Office



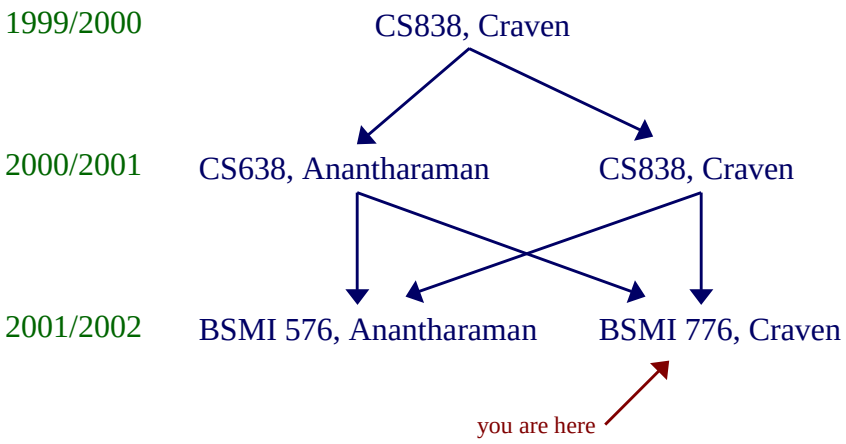
Course TA

- Wei Luo
  - luo@biostat.wisc.edu
  - 6749 Medical Sciences Center
    (across the hall from my office)
  - Office hours: 3:00-4:00pm Tuesday & Thursday

Computing Resources for the Class

- UNIX workstations in Dept. of Biostatistics & Medical Informatics
  - no “lab”, must log in remotely
  - more details later
- CS department offers UNIX orientation sessions
  - 4:00pm in 1325 Computer Sciences
  - January 23, 24, 28, 29, 30

The History of this Course



Expected Background

- technically, BSMI/CS 576
- statistics: good if you've had at least one course, but not required
- molecular biology: no knowledge assumed, but an interest in learning some basic molecular biology is mandatory

Related Courses

- BSMI/CS 576
- Biochemistry 711/712, "Sequence Analysis", taught by Prof. Ann Palmenberg
- not-for-credit evening BioModules on "Sequence Analysis", "Genetics Computing" and "Desktop Molecular Graphics" www.bocklabs.wisc.edu/acp/bnmcdrop/biomodinfo.html
- CS 731, "Advanced Artificial Intelligence with BiomedicalApplications", taught by Prof. David Page

Course Emphases

- Understanding the types and sources of data available for computational biology.
- Understanding the important computational problems in molecular biology.
- ✓ Understanding the most significant & interesting algorithms.

Course Requirements

- homework assignments: ~40%
  - programming
  - computational experiments (e.g. measure the effect of varying parameter *x* in algorithm *y*)
  - some written exercises
- project: ~20%
- final exam: ~ 35%
- class participation: ~ 5%

Course Readings

- required: *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids*. R. Durbin, S. Eddy, A. Krogh, and G. Mitchison. Cambridge University Press, 1998.
- recommended: *Introduction to Computational Molecular Biology*. J. Setubal and J. Meidanis. PWS Publishing, 1997.
- articles from the primary literature (scientific journals, etc.)

Reading Assignment

- for next week read:
  - Molecular Biology for Computer Scientists. L. Hunter
  - DOE Primer on Molecular Genetics
  - Finally, the Book of Life and Instructions for Navigating It. E. Pennisi. *Science*, 2000.
  - All of the above available from course web page
  - Chapter 2 (sections 2.1 to 2.5) from Durbin et al. OR Chapter 3 from Setubal & Meidanis

Student Survey

- name
- taking course for credit or sitting in
- grad/undergrad and year
- major/home department
- CS background
- biology background
- statistics background
- took 638 or 576 w/Prof. Anantharaman

What is Bioinformatics

- representation/storage/retrieval/analysis of biological data concerning
  - sequences
  - structures
  - functions
  - activity levels
  - networks of interactions

  of/among biomolecules
- sometimes used synonymously with *computational biology* or *computational molecular biology*

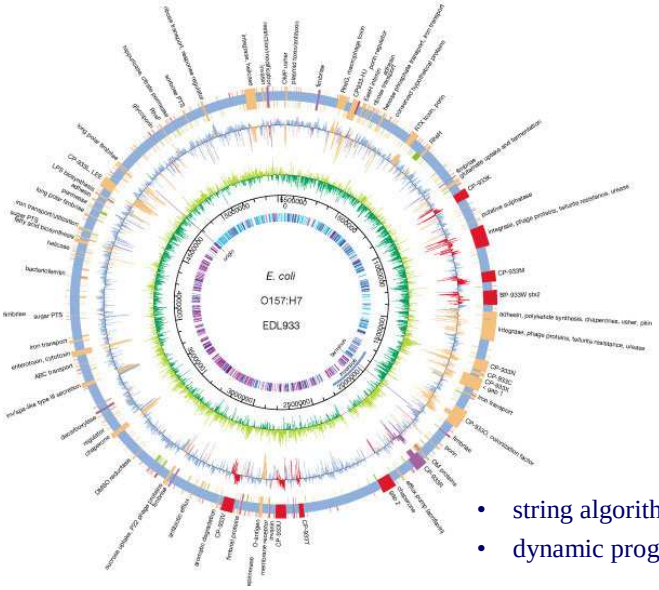
Topics to be Covered: Computational Problems in Molecular Biology

- pairwise sequence alignment
- sequence database searching
- multiple sequence alignment
- whole genome comparisons
- gene recognition
- protein structure and function prediction
- gene expression analysis
- phylogenetic tree construction
- RNA structure modeling
- biomedical text analysis

Topics to be Covered: Computer Science Issues & Algorithms

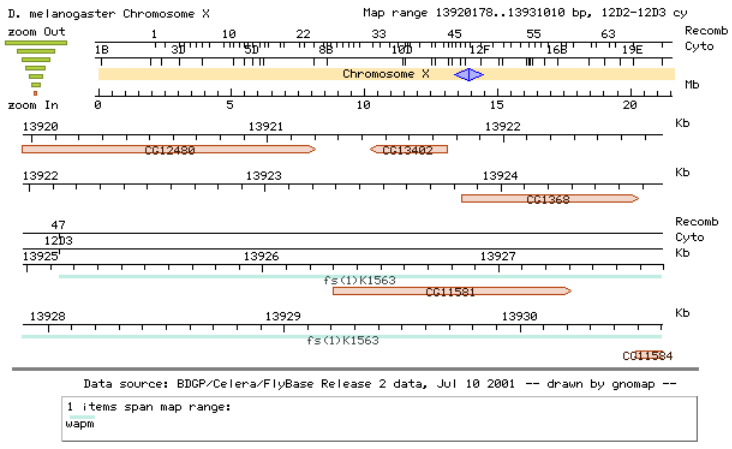
- string algorithms
- dynamic programming
- machine learning
- Markov chain models
- hidden Markov models
- stochastic context free grammars
- EM algorithms
- Gibbs sampling
- clustering
- tree algorithms
- text analysis
- and more…

What do two sequences/genomes have in common?



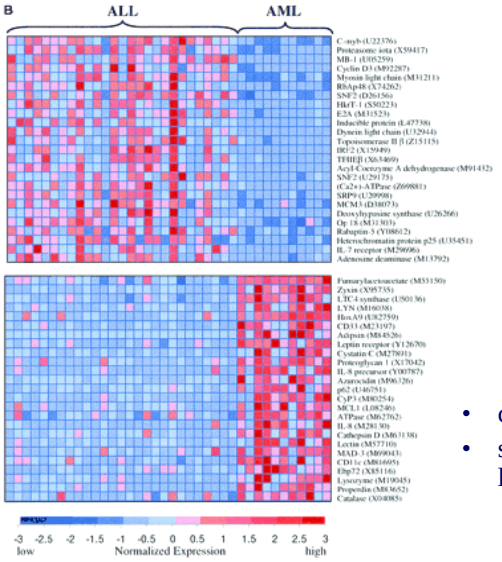
- string algorithms
- dynamic programming

Where are the genes in this genome?



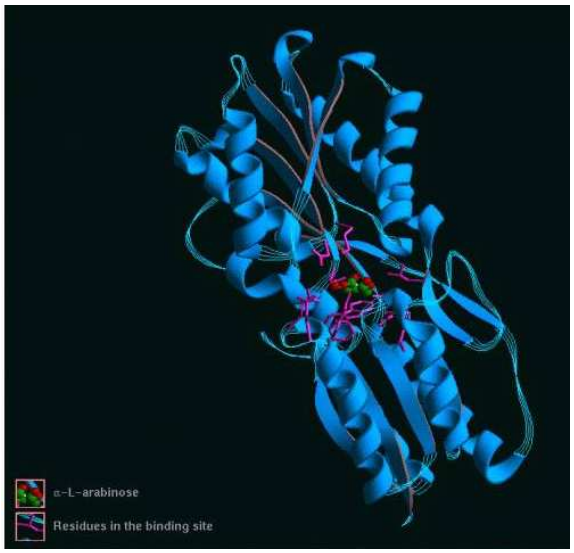
- Markov chain models
- hidden Markov models

Can diseases be characterized by patterns of gene activity?



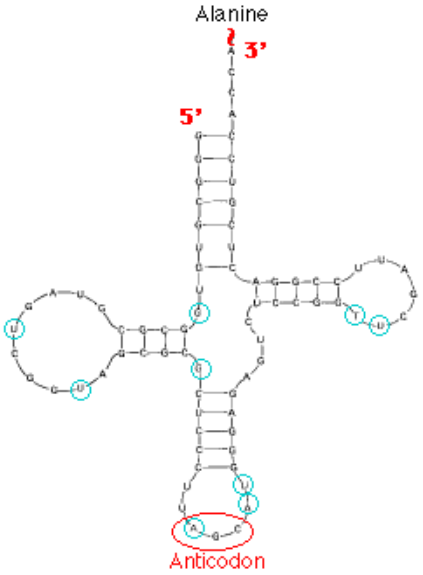
- clustering
- supervised machine learning

What does the protein encoded by this gene look like? What does it do?



- dynamic programming
- branch & bound
- hidden Markov models
- Tarot cards?

What other RNA sequences fold up like this?



- stochastic context free grammars