Advanced Bioinformatics

Biostatistics & Medical Informatics 776
Computer Sciences 776
Spring 2009

Mark Craven
Dept. of Biostatistics & Medical Informatics
Dept. of Computer Sciences
craven@biostat.wisc.edu
www.biostat.wisc.edu/bmi776/

Agenda Today

- introductions
- course information
- overview of topics

Course Web Site

- www.biostat.wisc.edu/bmi776/
- syllabus
- readings
- lecture slides in PDF
- homework
- mailing list archive
- etc.

Your Instructor: Mark Craven

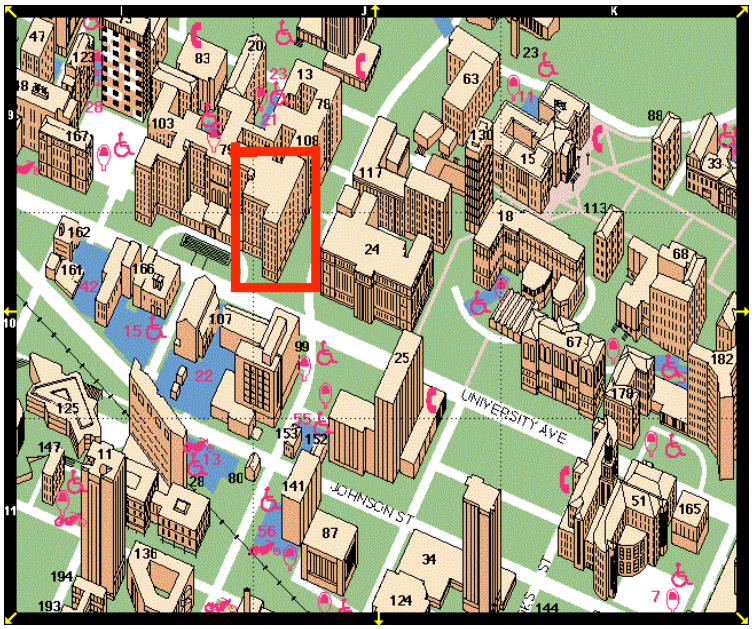
- email:
  craven@biostat.wisc.edu *or*
  craven@cs.wisc.edu

- office hours: TBA
  room 6730, Medical Sciences Center

- my home department is Biostatistics & Medical Informatics, and I have an affiliate appointment in Computer Sciences

- research interests: machine learning, gene regulation and cellular networks, biomedical text mining, probabilistic models, time series

Finding My Office: 6730 Medical Sciences Center



- confusing building
- best bet: enter at door marked *420 North Charter*

Course TA

- Dan Wong
  - dwong@cs.wisc.edu
  - 1301 Computer Sciences
  - office hours: TBA

Course Requirements

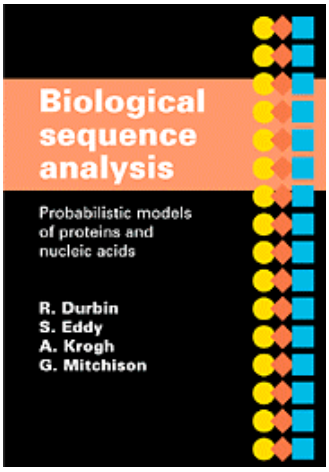
- 6 or so homework assignments: ~30%
  - programming (in Java, C++, C, Perl, Python)
  - computational experiments (e.g. measure the effect of varying parameter *x* in algorithm *y*)
  - written exercises
- project: ~30%
- final exam: ~ 30%
- class participation: ~10%

Participation

- Take advantage of the small class size!
- do the assigned readings
- show up to class
- don't be afraid to ask questions

Course Readings

- mostly articles from the primary literature (scientific journals, etc.)
- *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids*. R. Durbin, S. Eddy, A. Krogh, and G. Mitchison. Cambridge University Press, 1998.



Computing Resources for the Class

- Linux workstations in Dept. of Biostatistics & Medical Informatics
  - no “lab”, must log in remotely
  - accounts will be created later this week
  - two machines
    - mi1.biostat.wisc.edu
    - mi2.biostat.wisc.edu
- CS department usually offers UNIX orientation sessions at beginning of semester
- the “CS 1000” UNIX tutorial
  - online at http://www.cs.wisc.edu/csl/cs1000/

The Class Mailing List

- bmi776-1-s09@lists.wisc.edu
- you will automatically be subscribed
- check your mail daily or have it forwarded to an account where you do

Major Topics to be Covered (the task perspective)

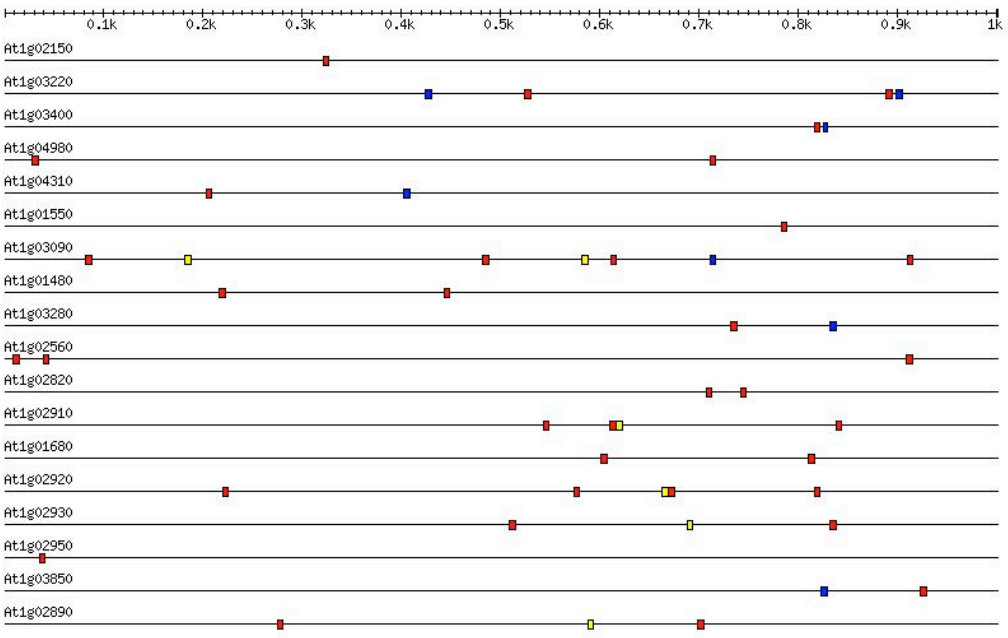
- modeling of motifs and *cis*-regulatory modules
- gene finding
- advanced topics in sequence alignment
- modeling cellular networks
- RNA sequence and structure modeling
- biomedical text mining
- phylogenetic inference
- association studies

Major Topics to be Covered (the algorithms perspective)

- Gibbs sampling and EM
- HMM structure search
- duration modeling and semi-Markov models
- pairwise HMMs
- interpolated Markov models and back-off methods
- parametric alignment
- suffix trees
- sparse dynamic programming
- stochastic context free grammars
- Bayesian networks and module networks
- abduction
- topic models
- etc.

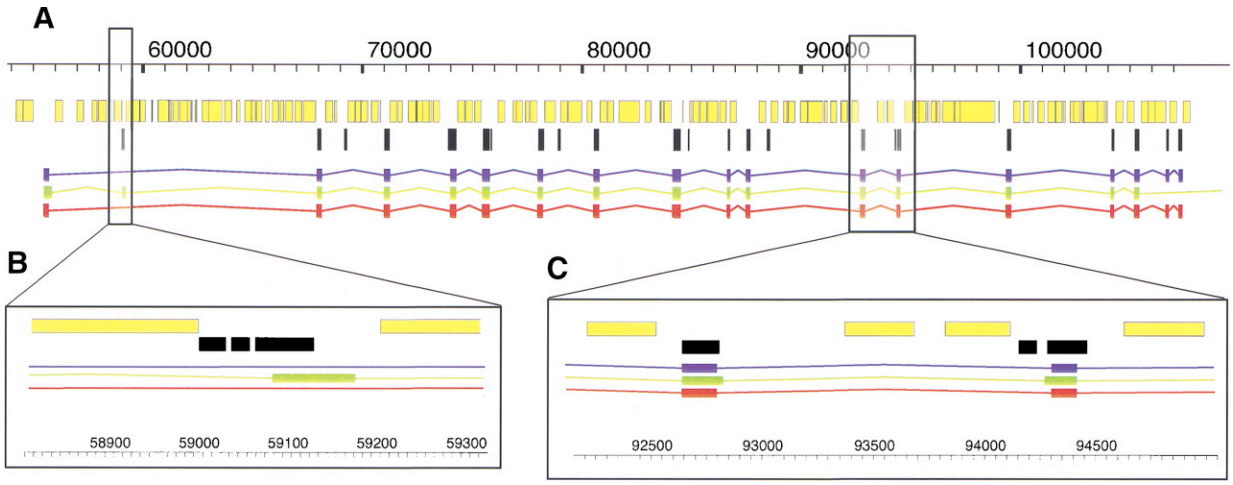
Motif and CRM Modeling

What sequence motifs do these promoter regions have in common?



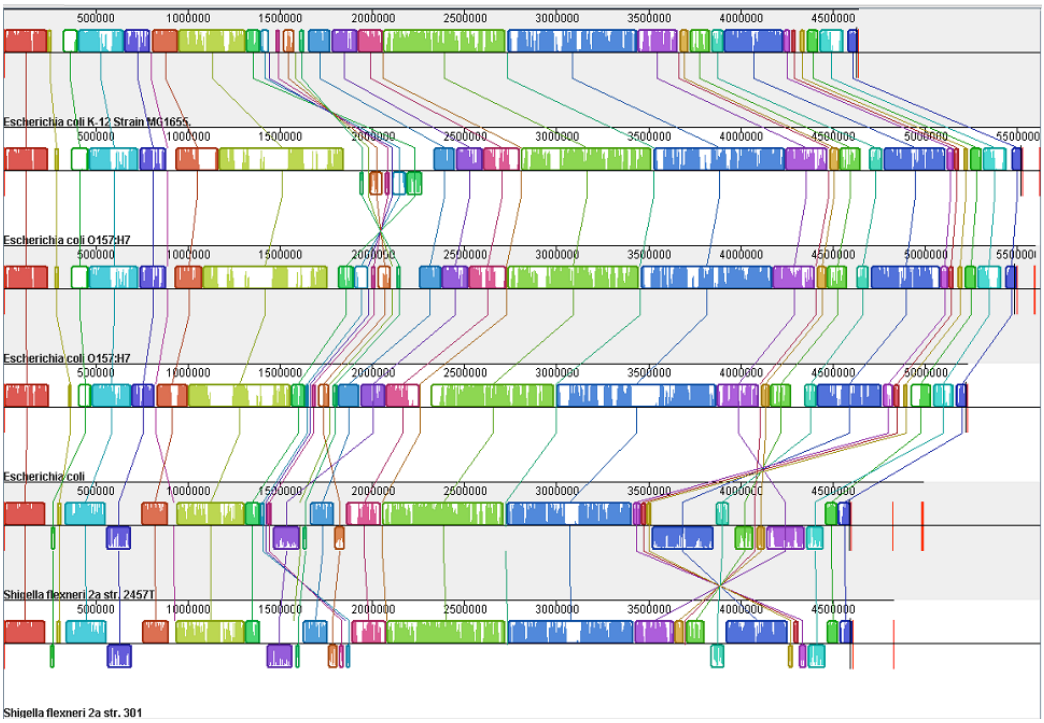
Gene Finding

Where are the genes in this genome, and what is the structure of each gene?



Advanced Topics in Sequence Alignment

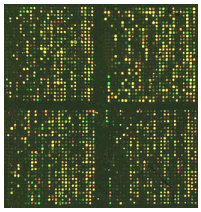
What is the best alignment of these 5 genomes?

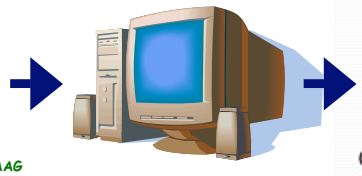


Inferring Cellular Networks

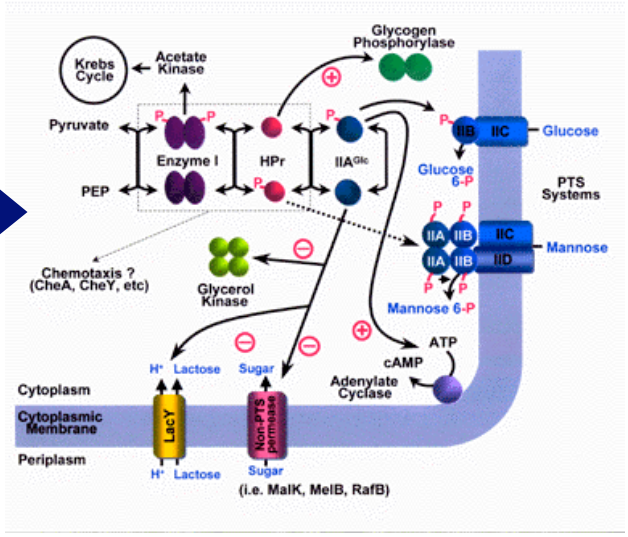
Can we automatically infer models of regulatory/signaling/metabolic networks from data?

high-throughput data



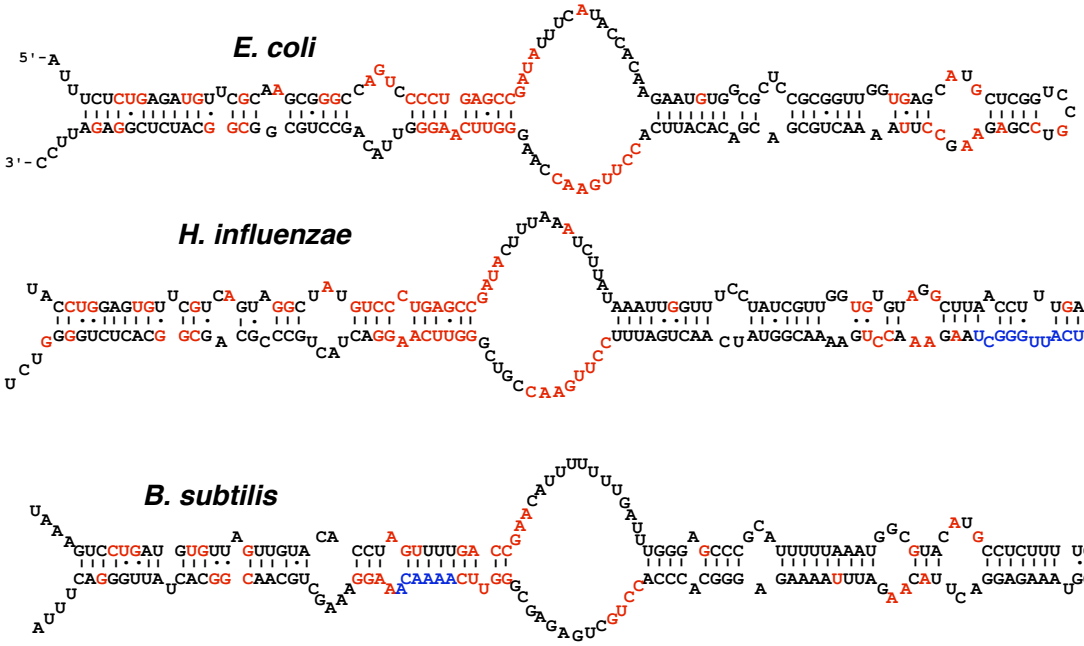


```
AGTGCGATTGACAGATCACAGCTCTGAAG
TAGAACGATCTTCTCTCTCGATCGCTCGC
TCTCGCTGCACACACACACATGGGCGCTG
TGCGATTGACAGATCACAGCTCTGAAGTA
GAACGATCGCTGCAGTGCGATTGACAGAT
CACAGCTCTGAAGTAGAACGATCTTCTCT
CTCGATCGCTCGCTCTTGAAGTAGTCGCT
CGCTCTTGAAGAGTGCGATTGACAGATCA
CAGCTCTGAAGTAGAACGATCTTCTCTCT
CGATCGCTCGCTCTCGCTGCACACACACA
```

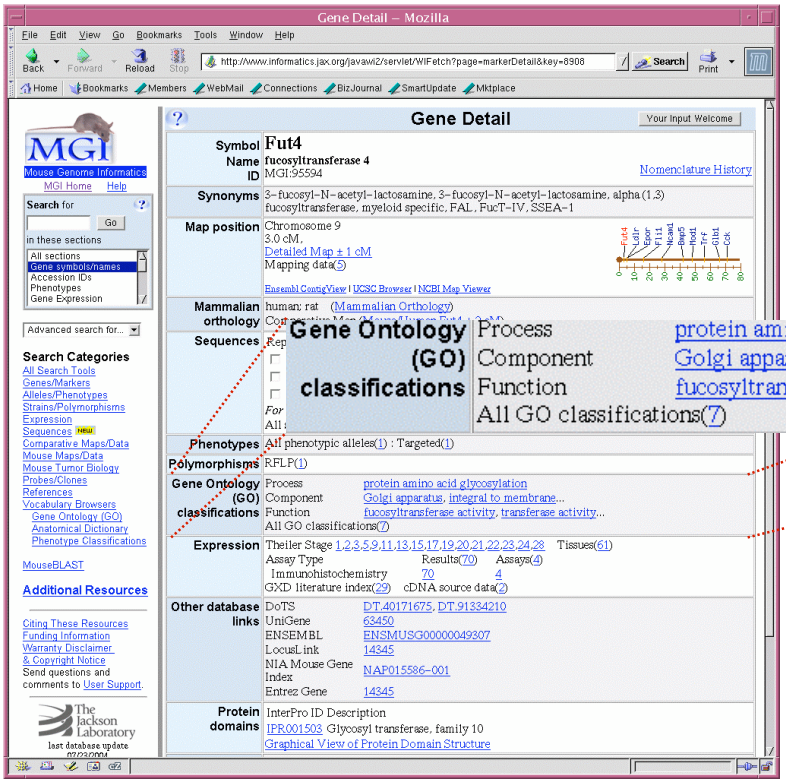


RNA Sequence and Structure Modeling

Given a genome, how can we identify sequences that encode this RNA structure?

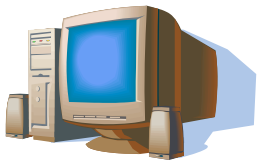


Biomedical Text Mining



Can we partially automate the process of curating genomics databases?

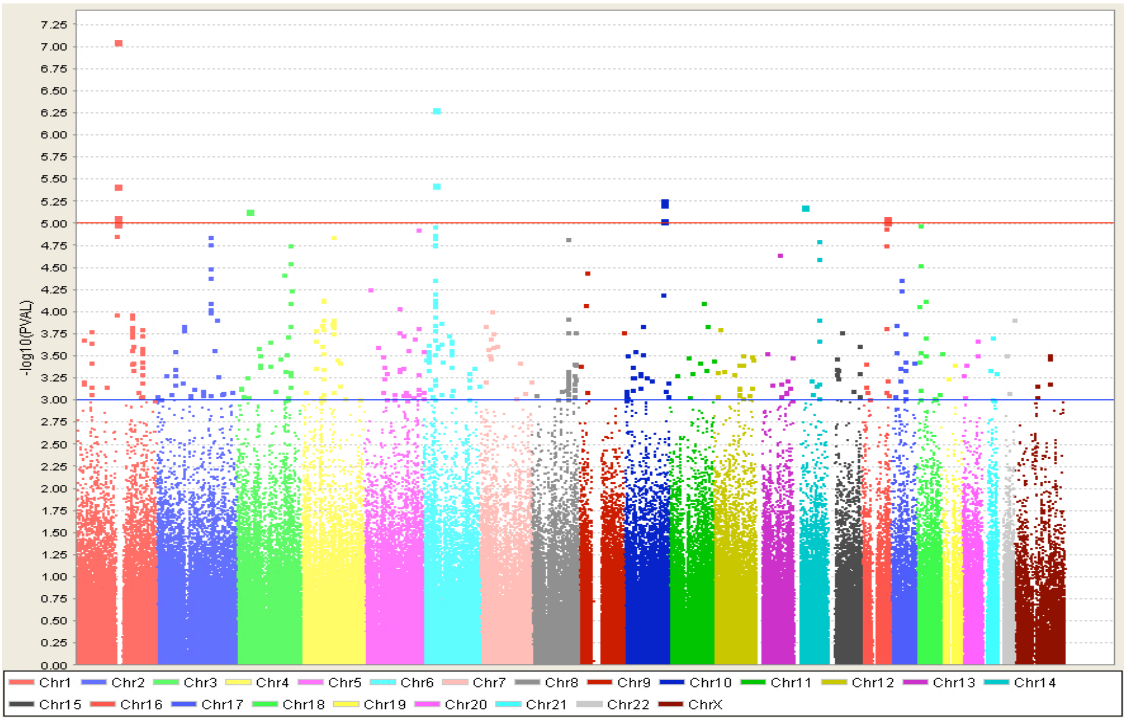
Biomedical Text Mining



gene: FUT 4
GO concept : protein amino-acid glycosylation

Association Studies

Which genes are involved in diabetes?



Type 2 diabetes association P values by chromosome (386,731 markers). The x-axis is the genomic position by chromosome 1-22 and X (by color), and the y-axis is the negative base 10 logarithm of the P value.

Reading Assignment

- Bailey and Elkan, *ISMB* ’95
- Lawrence et al., *Science* ’93
- available on the course web site