

Introduction to Bioinformatics

Biostatistics & Medical Informatics 576

Computer Sciences 576

Fall 2011

Mark Craven

Dept. of Biostatistics & Medical Informatics

Dept. of Computer Sciences

craven@biostat.wisc.edu

www.biostat.wisc.edu/bmi576/

BMI/CS 576: Bioinformatics

- instructor: Prof. Mark Craven
craven@biostat.wisc.edu
- office hours:
Tuesday 2:30-3:30
Friday 11-12
or by appointment
room 6730, Medical Sciences Center
- course home page: <http://www.biostat.wisc.edu/bmi576/>

Finding my office

- 6730 Medical Sciences Center
- easiest to enter from Charter St. and take elevator immediately to your right



Course TA

Sharad Akshar Pununganti

akshar@cs.wisc.edu

1302 Computer Sciences (1210 W. Dayton St.)

office hours: Monday, Wednesday, Friday 1-2pm

Finding the TA's office

1302 Computer Sciences



Expected background

- CS 367 (Intro to Data Structures) or equivalent
- 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

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

- 5 or so homework assignments: ~50%
 - programming
 - computational experiments (e.g. measure the effect of varying parameter x in algorithm y)
 - written exercises
- midterm exam: ~20%
- final exam: ~ 30%

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
- must use WiscVPN to access the servers

Linux help

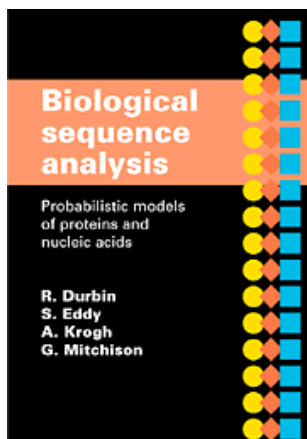
- CS Department Linux orientations
 - Tuesday, September 6 @ 4:30 PM
 - Wednesday, September 7 @ 4:00 PM
 - Thursday, September 8 @ 4:00 PM
- held in room 1325 Computer Sciences

Programming languages

- for the programming assignments, you must use one of
 - C
 - C++
 - Java
 - Perl
 - Python

Course readings

- *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids*. R. Durbin, S. Eddy, A. Krogh, and G. Mitchison. Cambridge University Press, 1998.



- articles from the primary literature (scientific journals, etc.)

The short-term plan

- Thursday (9/8) and Tuesday (9/13)
optional “Molecular Biology 101” lectures
- next Thursday (9/15)
start on “Sequence Assembly”

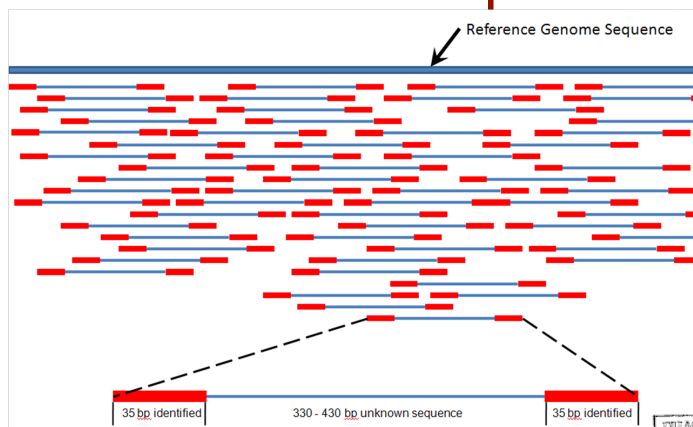
Reading assignment

- Life and Its Molecules: A Brief Introduction. L. Hunter
(available from web page)

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*

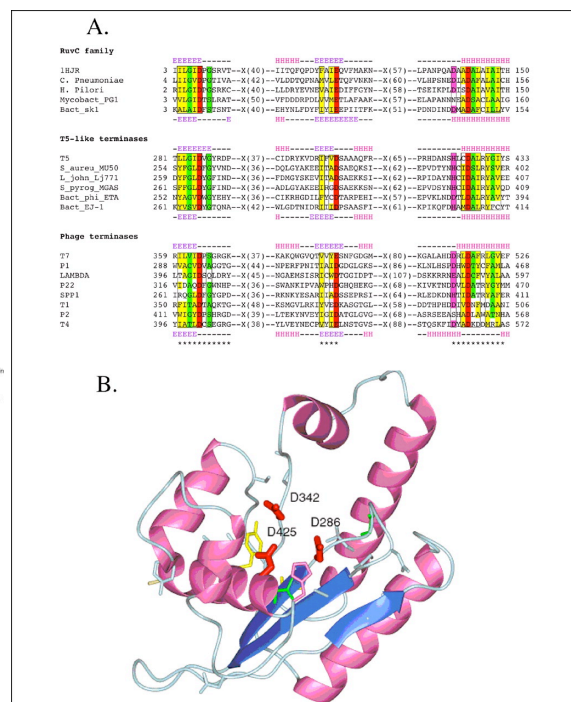
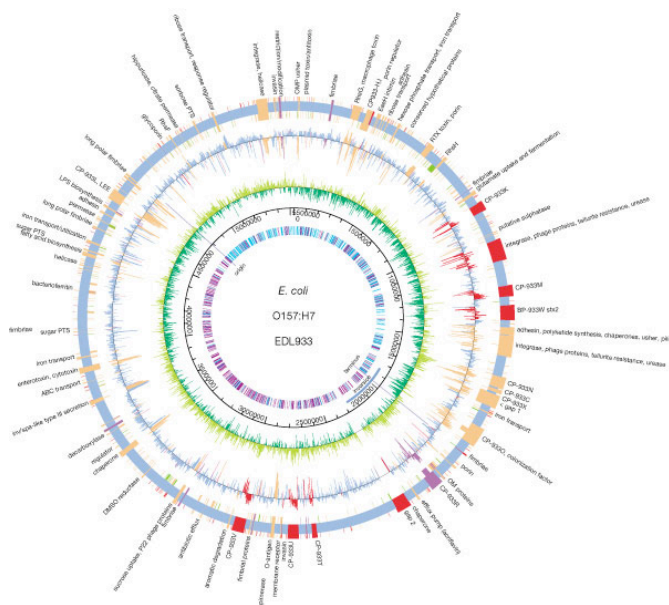
How can we assemble a genome from sequence fragments?



Sequence assembly topics

- fragment assembly
- spectral assembly
- shortest superstring problem
- finding Hamiltonian paths in overlap graphs
- finding Eulerian paths in k-mer graphs

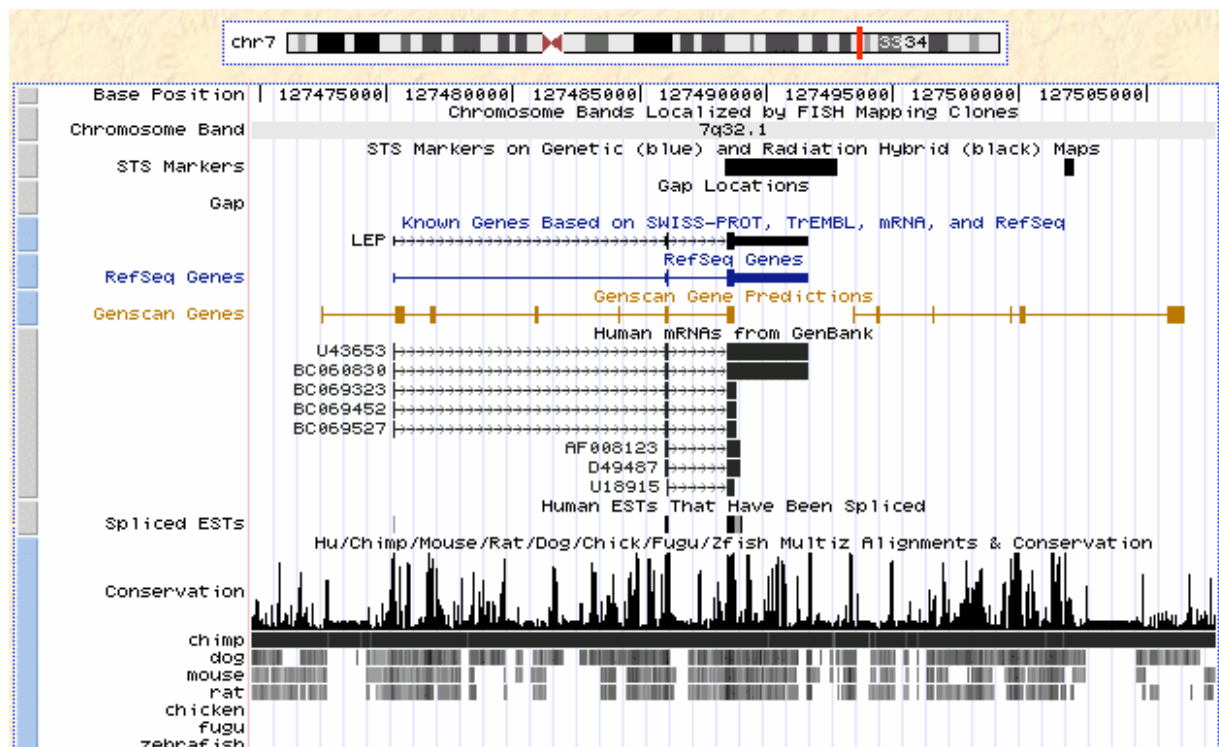
What do these DNA/protein sequences have in common?



Sequence alignment topics

- pairwise and multiple sequence alignment
- dynamic programming methods for global and local alignments
- linear and affine gap penalty functions
- the BLAST algorithm
- dynamic programming and heuristic methods for multiple sequence alignment
- alignment statistics and substitution matrices

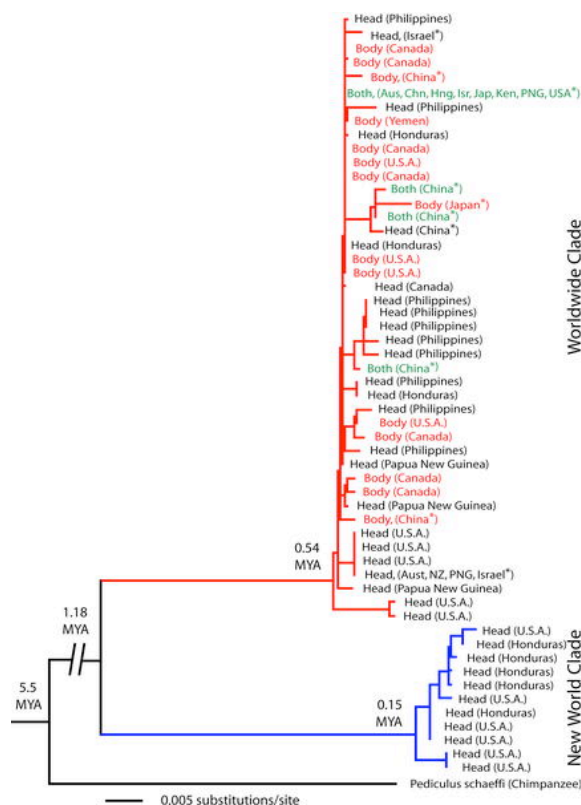
Where are the genes in this genome?



Probabilistic sequence modeling topics

- Markov chains
- high-order Markov models
- inhomogeneous Markov models
- hidden Markov models
- Forward/Backward/Viterbi algorithms
- applications to gene finding and protein family modeling

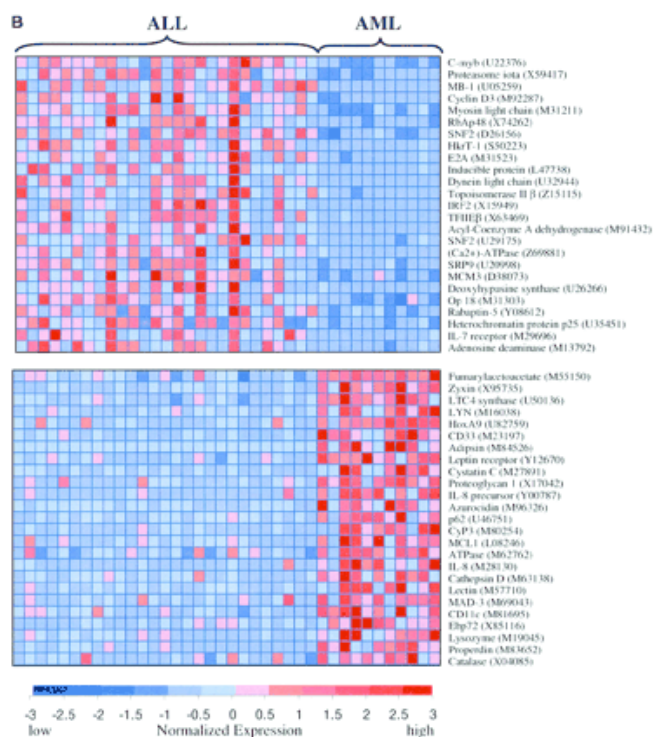
How are these species related?



Phylogenetic tree inference topics

- distance-based approaches
- parsimony-based approaches
- branch-and-bound search

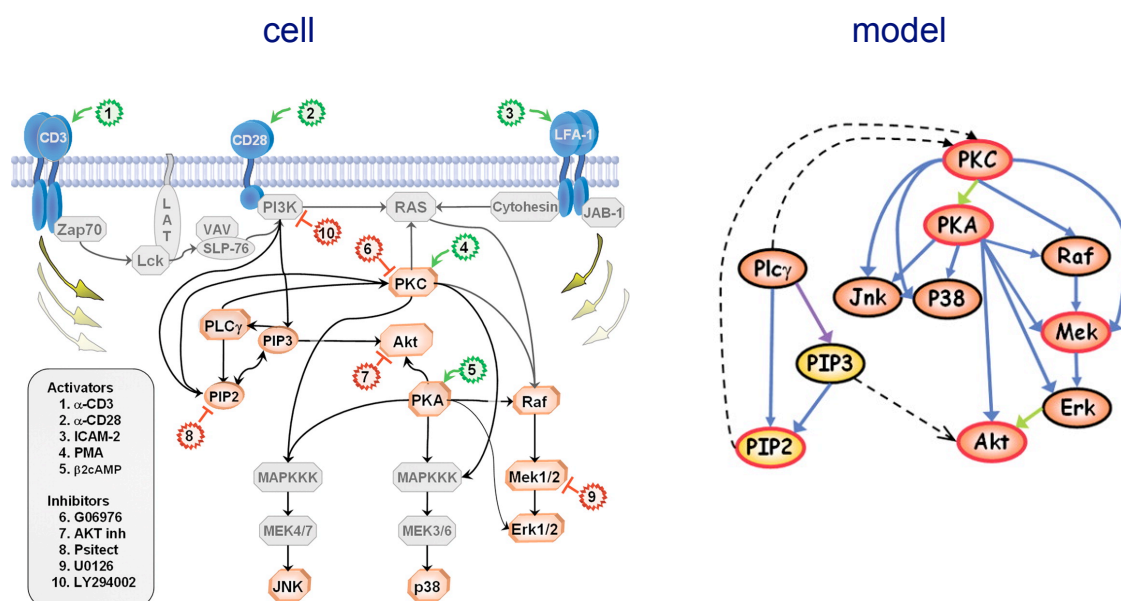
Can diseases be characterized by patterns of gene activity?



Topics in analyzing data from high-throughput experiments

- clustering algorithms
 - hierarchical clustering
 - *k*-means clustering
 - EM-based clustering
- classification algorithms (simple methods for supervised learning)
- multiple hypothesis testing and the false discovery rate

Can we induce models of cellular processes from high-throughput experiments?



Network modeling topics

- Bayesian network representations
- algorithms for exact inference in BNs
- algorithms for structure and parameter learning in BNs
- applications of BNs to inferring subcellular network models