Introduction to Bioinformatics
Biostatistics & Medical Informatics 576
Computer Sciences 576
Fall 2003

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 or
  – craven@cs.wisc.edu
• Office hours: 2:00-3:00 Tuesday and Wednesday
  or by appointment
  – room 6730, Medical Sciences Center
• Course home page: www.biostat.wisc.edu/bmi576/
• Course mailing list: bmi576@biostat.wisc.edu
  – you will automatically be subscribed
  – check your mail daily or have it forwarded to an account
    where you do
Finding My Office

Course TA

- Wei Luo
  - luo@biostat.wisc.edu
  - 6763 Medical Sciences Center
    (across the hall from my office)
  - Office hours: 2-3:00pm Monday, 2-3:00pm Friday
Computing Resources for the Class

- UNIX 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?
- the “CS 1000” UNIX tutorial
  - can buy a hardcopy at the DoIT Tech Store
  - online at http://www.cs.wisc.edu/csl/cs1000/

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.

Related Courses

- Biochemistry 711/712, “Sequence Analysis”, taught by Prof. Ann Palmenberg (next fall)
- BMI/CS 776, “Advanced Bioinformatics” (next spring, 576 is a prerequisite)
- CS 731, “Advanced Artificial Intelligence with Biomedical Applications”, taught by Prof. David Page (this semester)
- some Statistics special topics courses that may become regular courses
  - Statistics 692, “Statistical Methods in Genomics”, taught by Prof. Nicole Perna and Dr. Bob Mau
  - Statistics 692, “Statistical Genomics”, taught by Prof. Brian Yandell
  - Statistics 992, “Microarray Data Analysis” taught by Prof. Christina Kendziorski
Course Requirements

- 5 or so homework assignments: ~40%
  - mostly programming (in Java)
  - computational experiments (e.g. measure the effect of varying parameter $x$ in algorithm $y$)
  - some written exercises
- midterm exam: ~25%
- final exam: ~35%

Course Readings

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

• for this week read:
  – Molecular Biology for Computer Scientists. L. Hunter
  – DOE Primer on Molecular Genetics
  – both are available from course web page
  – additionally the web page lists two recommended readings
• for next week read:
  – Sections 3.1-3.3 in Setubal and Meidanis
  – both are available from course web page
  – additionally the web page lists two recommended readings

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
Major Topics to be Covered

- pairwise and multiple sequence alignment
- probabilistic sequence models
- phylogenetic tree inference
- gene expression analysis
- protein structure prediction
- probably a few others depending on time

What do two sequences/genomes have in common?
Sequence Alignment

- 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 Models

- Markov chains
- high-order Markov models
- inhomogenous Markov models
- hidden Markov models
- Forward/Backward/Viterbi algorithms

How are these species related?
Phylogenetic Tree Inference

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

Can diseases be characterized by patterns of gene activity?
Gene Expression Analysis

- clustering gene-expression profiles
- hierarchical clustering
- $k$-means clustering
- EM-based clustering
- classification of gene-expression profiles
- simple supervised learning methods

What does the protein encoded by a given gene look like? What does it do?
Protein Structure Prediction

- protein threading
- dynamic programming revisited
- hidden Markov models revisited
- branch-and-bound search revisited

The Short-term Plan

- Thursday (9/4) and next Tuesday (9/9)
  - optional “Molecular Biology 101” lectures
- Thursday (9/11)
  - start on “Pairwise Sequence Alignment”
Student Survey

• tell me the following
  – name
  – taking course for credit or sitting in
  – grad/undergrad and year
  – major/home department
  – CS background
  – biology background
  – statistics background
  – why you are taking this course