Today

- Introductions
- Course information
- Overview of course topics
My introduction

• Arrived in August

• Departments of Biostatistics & Medical Informatics and Computer Sciences

• Member of the Genome Center of Wisconsin

• Interests in comparative genomics

• Expertise in multiple whole-genome alignment
Your introductions

- Name
- Department
- Year
- Interests (academic/research)
Web site

- URL: http://www.biostat.wisc.edu/bmi776/
- Syllabus/Readings/Lectures
- Homeworks/Project
- Email list and archive
Office hours

- Times TBA
- 6720 Medical Sciences Center (MSC)
- Very confusing building
- Best bet: Take elevator near 420 N. Charter St. entrance
Prerequisites

- BMI/CS 576
- Computer Science: Graphs, Dynamic Programming (at least CS 367)
- Statistics: Probability, Bayesian networks
- Biology: None required, but you must be interested in learning some!
Related courses

• Botany 563
  • Phylogenetic Analysis of Molecular Data
  • Taught by Professor David Baum

• Computer Science 838
  • Advanced Natural Language Processing
  • Taught by Professor Jerry Zhu
Biological sequence analysis
Probabilistic models of proteins and nucleic acids
R. Durbin
S. Eddy
A. Krogh
G. Mitchison
Reading

• Reading assignments for each lecture
• Types of reading
  • From textbook
  • Research articles
• Notes passed out in class (better come!)
Grading

- Participation: 10%
- Final Exam: 30%
- Project: 30%
- Homework: 30%
- Final Exam: 30%
Participation

• Very small class
• Keys to participation
  • Show up to class!
  • Do the assigned reading
  • Don’t be afraid to ask questions
Homework

- Programming
  - Implement algorithms from course
  - Analyze real data
  - Preferred languages: C, C++, Java, & Python
- Written problems
  - Algorithm simulation
  - Proofs
Project

- Goals:
  - Develop a new model/algorithm
  - Implement it
  - Apply to a meaningful data set

<table>
<thead>
<tr>
<th>Milestones</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>March 9</td>
<td>Proposal</td>
</tr>
<tr>
<td>April 13</td>
<td>Progress Report</td>
</tr>
<tr>
<td>May 4</td>
<td>Final report</td>
</tr>
</tbody>
</table>
Final exam

- Cumulative
- Open book/notes, no memorization required
Computer accounts

• BMI UNIX machines
  • No lab, login remotely via SSH
  • May need VPN if off campus network
  • Machines: mi1.biostat.wisc.edu, mi2.biostat.wisc.edu

• Need UNIX help? Try CS1000 from DoIT
Major topics

- Biology of nucleic acids
- Modeling of nucleotide evolution
- Finding elements in genomes
- Genome alignment
- RNA structure and discovery
- Analysis of cellular networks
Plan of attack

- All things DNA-related before Spring break
- RNA, cellular networks, and other topics after the break
Course goals

• We will have been successful if...

• You are aware of and understand the most important problems in computational molecular biology

• You have an understanding of the models and algorithms that are currently used for these problems.
Course theme

"Nothing in biology makes sense except in the light of evolution."
- Theodosius Dobzhansky

• Evolution as a **tool** in deciphering the genome

• “Comparative” models

• Combine **within** genome models with **between** genome models
Biology of nucleic acids

• How does DNA replicate and mutate?
• How do we define evolutionary relationships between DNA positions?
• How do we represent and classify such relationships?
• Key concepts: homology, tree theory
Modeling of nucleotide evolution

• How do we model the evolution of a set of sequences from an ancestral sequence?
• How can we use such models to infer trees?
• How might we reconstruct ancestral sequences?
• Key concepts: Markov models, Poisson processes, Rate matrices, Maximum likelihood, Bayesian analysis
Motif finding

How can we find common (degenerate) patterns in a set of functionally similar sequences?

Key concepts: Hidden Markov models, Gibbs sampling, Expectation-Maximization

CTATCGTAGCGACTGCTACTCGATACTAGCT
CACTAGTCCATGCTTGCTAGGCGAGTCGTAGC
CGATCGGGATTAAGTCGAAGCCTCGCAAAACCA
CGCAATTTCGATGCTCACATGAGCATTGGGGCC
CATCGTATGGCTCAAGTCGATCCTAGGACGA
Gene finding

- How do we find gene structures in the genomes of... Prokaryotes? Eukaryotes?
- Can we use comparative genomics to increase the accuracy of our predictions?
- Key concepts: Generalized HMMs, higher-order Markov models, Pair HMMs.
Alignment

- What is the meaning of sequence alignment?
- How do we align...
  - a pair of short sequences?
  - a pair of long sequences?
  - multiple sequences?
  - whole genomes?

CGCCTCGGGT
CGCC---GGT
CACCTAGTAC
CGCTACTTGC
CG--TCTTGC
CGTAGGCTTTC
Alignment concepts

• Alignment combinatorics
• Statistical alignment
• Pair Hidden Markov Models
• Local alignment and statistics (Karlin-Altschul theorem)
• Parametric alignment
Pattern matching

- How can we quickly identify highly-similar substrings in sets of large sequences?
- Key concepts: Suffix trees/arrays, locality-sensitive hashing, q-gram filtration, randomized matching
RNA

• How can we predict the secondary structure of RNA?
• How can we locate RNAs of a given structure within a genome?
• Key concepts: Energy minimization, Stochastic context free grammars
Cellular Networks

- How can we infer and represent interactions between cellular components?
- How can we determine important functional modules within networks?

Key concepts: Graph theory, Inference of Bayesian networks
Next time

• Topic: “The trees of life”

• Assignments:
  • Do assigned readings (check Web site)
  • Log in to BMI machines and change password (use command “passwd”)
  • First homework to be assigned on Thursday