Advanced Bioinformatics
Biostatistics & Medical Informatics 776
Computer Sciences 776
Spring 2016

Anthony Gitter
gitter@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 and policies
- Readings
- Tentative schedule
- Lecture slides (posted after each lecture)
- Homework
- Project information
- Link to Piazza discussion board
Your Instructor: Anthony Gitter

• Email: gitter@biostat.wisc.edu

• Office: room 3268, Discovery Building

• Assistant professor in the department of Biostatistics & Medical Informatics with an affiliate appointment in Computer Sciences

• Investigator in the Morgridge Institute for Research

• Research interests: biological networks, time series analysis, computational problems related to cancer and virology
Your TA: Amin Alhashim

- Email: alhashim@wisc.edu
- Office: desk next to 3251, Discovery Building
- Graduate student in the department of Computer Sciences
- Research interests: visualization
Tentative Office Hours

• Instructor: Tuesday and Thursday, 2:30-3:30 PM

• TA: Monday, 10:00-11:30 AM

• Who cannot attend any of these times?
Finding My Office: Discovery Building

- 3<sup>rd</sup> floor has restricted access
- Partial WID access will be enabled in the first week or two
- Stop at visitor desk to call my office if card does not work
So that we can all get to know each other better, please tell us your

- name
- major or graduate program
- research interests and/or topics you’re especially interested in learning about in this class
- favorite programming language
Course Requirements

• 4 or 5 homework assignments: ~40%
  – written exercises
  – programming (Python, Java, C++, C, Perl, etc.) + computational experiments (e.g. measure the effect of varying parameter x in algorithm y)
  – five late days permitted

• Project: ~25%
• Midterm: ~15%
• Final exam: ~15%
• Class participation: ~5%
Exams

• Midterm: March 8\textsuperscript{th}, in class
• Final: May 12\textsuperscript{th}, 2:45-4:45 PM
• Please let me know \textit{immediately} if you have a conflict with either of these exam times
Project

• Design and implement a new computational method for a task in molecular biology
• Improve an existing method
• Perform an evaluation of several existing methods
• Run on real biological data
• Some project suggestions will be listed on website
• Each student works individually
Participation

• Do the assigned readings
• Show up to class
• No one will have the perfect background
  – Ask questions about computational or biological concepts
• Piazza discussion board
Piazza Discussion Board

• Instead of a mailing list
• http://piazza.com/wisc/spring2016/bmics776/home
• Please consider posting your questions to Piazza first before emailing the instructor or TA
• Also consider answering your classmates’ questions
• Announcements will also be posted to Piazza
Course Readings

• Mostly articles from the primary literature
• Must be using a UW IP address to download some of the articles (can use WiscVPN from off campus)
Prerequisites

• BMI/CS 576 or equivalent
• Knowledge of basic biology and methods from that course will be assumed
• May want to go over the material on the 576 website to refresh
• http://www.biostat.wisc.edu/bmi576/
Computing Resources for the Class

- Linux workstations in Dept. of Biostatistics & Medical Informatics
  - no “lab”, must log in remotely (use WiscVPN)
  - accounts created for everyone on course roster
  - two machines
    - mi1.biostat.wisc.edu
    - mi2.biostat.wisc.edu
  - HW0 tests your access to these machines

- CS department usually offers Unix orientation sessions at beginning of semester
What you should get out of this course

• An understanding of some of the major problems in computational molecular biology
• Familiarity with the algorithms and statistical techniques for addressing these problems
• At the end you should be able to
  – Read the bioinformatics literature
  – Apply the methods you have learned to other problems both within and outside of bioinformatics
Major Topics to be Covered (the task perspective)

- Modeling of motifs and cis-regulatory modules
- Identification of transcription factor binding sites
- Genotype analysis and association studies
- Transcriptome quantification
- Mass spectrometry peptide and protein identification
- Modeling evolution and pathways in cellular networks
- Gene finding
- Large-scale and whole-genome sequence alignment
- RNA sequence and structure modeling
- Protein structure prediction
Major Topics to be Covered (the algorithms perspective)

- Expectation Maximization and Gibbs sampling
- Hidden Markov Model structure search
- Duration modeling and semi-Markov models
- Neural networks
- Linear programming
- Pair HMMs
- Interpolated Markov models and back-off methods
- Tries and suffix trees
- Sparse dynamic programming
- Markov random fields
- Stochastic context free grammars
- Branch and bound search
Motif Modeling

What sequence motif do these promoter regions have in common?
cis-Regulatory Modules

What configuration of sequence motifs do these promoter regions have in common?
Genome-wide 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.
Noncoding Variants

How do genetic variants outside protein coding regions impact phenotypes?

Transcriptome Analysis with RNA-Seq

What genes are expressed and at what levels?
Proteomic Analysis with Mass Spectrometry

What proteins are expressed and at what levels?

How do networks align across species?
Gene Finding

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

What is the best alignment of these 6 genomes?
RNA Sequence and Structure Modeling

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

Can we predict the 3D shape of a protein from its sequence?

\[ E = \varepsilon \sum_{i} |c_i - c_i^{\text{target}}| + E_{\text{steric}} \]
Other Topics

• Many other topics we aren’t covering
  – Modeling long reads
  – Metagenomics
  – Epigenomics
  – Etc.

• Any other topics of interest?
Courses of Interest

• Tools for Reproducible Research (BMI 826)
  – Prof. Karl Broman
  – http://kbroman.org/Tools4RR/

• Others?
Reading groups

• Computational Systems Biology Reading Group
  – http://lists.discovery.wisc.edu/mailman/listinfo/compsysbiojc

• AI Reading Group
  – http://lists.cs.wisc.edu/mailman/listinfo/airg