Advanced Bioinformatics
Biostatistics & Medical Informatics 776
Computer Sciences 776
Spring 2020

Daifeng Wang
daifeng.wang@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 (draft posted before lecture)
- Announcements
- Homework
- Project information
- Link to Piazza discussion board
Your Instructor: Daifeng Wang

- Email: daifeng.wang@wisc.edu
- Website: https://daifengwanglab.org/
- Office: Waisman Center 517
- Office for Office Hours: Medical Sciences Center 4750
- Office Hours: Tue 2:30-3:30pm, Thu 2:30-3:30pm

- Assistant Professor in the Department of Biostatistics & Medical Informatics and Investigator in Waisman Center

- Research interests: interpretable machine learning, network biology, functional genomics, comparative genomics, brain diseases, precision medicine
My research in Waisman Center

• Mission of Waisman Center
  – Advance knowledge about human development, developmental disabilities, and neurodegenerative diseases

• Goal of my research

https://www.waisman.wisc.edu/2020/01/07/new-researcher-uses-machine-learning-to-decode-genomic-information/
Finding My Office:
517 Waisman Center

- Far away, most west building
- Take No. 80 Bus or Bike/Walk for exercise
Finding My “Office for Office Hours”:
Medical Sciences Center 4750
Course TA

- Fangzhou Mu
  - fmu2@wisc.edu
  - MSC 6729
  - Graduate student
    - Pharmacy & CS
Finding Fangzhou’s Office: MSC 6729

- **very** confusing building
- best bet: use *420 North Charter St entrance*
Office Hours

• Tue 2:30-3:30pm, Thu 2:30-3:30pm
  – MSC 4750
• Will begin next week
• Free to schedule an individual meeting
  – Waisman Center or other places
• You are encouraged to visit our office hours!
You

• 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
  – favorite programming language
Course Requirements

• 4 homework assignments: ~40%
  – Written exercises
  – Programming (Python)
  – 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: Thursday, March 12, in class
• Final: Monday May 4, 7:45-9:45 AM

• Let me know *immediately* if you have a conflict with either of these exam times
Computing Resources for the Class

• Linux servers in Dept. of Biostatistics & Medical Informatics
  – No “lab”, must log in remotely (use WiscVPN)
  – Will create accounts for everyone on course roster
  – Two machines
    mi1.biostat.wisc.edu
    mi2.biostat.wisc.edu
  – HW0 tests your access to these machines
  – Homework must be able to run on these machines

• CS department usually offers Unix orientation sessions at beginning of semester
Programming Assignments

- All programming assignments require Python
  - Project can be in any language

- Have a Python 3 environment on biostat servers
  - Permitted packages on course website
  - Can request others

- HW0 will be Python introduction

- Use Piazza for Python discussion
  - If you know Python, please help answer questions
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
• Suggestions will be provided
• Not simply your existing research
• Can email me now to discuss ideas
Participation

• Do the assigned readings before class
• Show up to class
• No one will have the perfect background
  – Ask questions about computational or biological concepts
• Correct me when I am wrong
  – Seriously, it will happen
• Piazza discussion board
  – Questions and answers
Piazza Discussion Board

- Instead of a mailing list
- [https://piazza.com/wisc/spring2020/bmics776](https://piazza.com/wisc/spring2020/bmics776)
- Post your questions to Piazza instead of emailing the instructor or TA
  - Unless it is a private issue or project-related
- Answer your classmates’ questions
- Announcements will also be posted to Piazza
- Supplementary material for lecture topics
Course Readings

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

Recommended online reading

• Translational Bioinformatics
  – https://collections.plos.org/translational-bioinformatics
Python references

• https://docs.python.org

• If you want a book:
  – Python 3 for programmers

• Many other good books and online resources

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/
What you should get out of this course

• An understanding of some of the major problems in computational biology and bioinformatics
• Familiarity with the techniques for addressing these problems
  – Computational, statistical, machine learning
• How to think about different data types
• 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
  – Write a short bioinformatics research paper
Major Topics to be Covered (the algorithms perspective)

• Expectation Maximization
• Gibbs sampling
• Mutual information
• Network flow algorithms
• Stochastic context free grammars
• Multiple hypothesis testing correction
• Deep learning (e.g., Convolutional neural networks)
• Linear programming
• Tries and suffix trees
• Markov random fields
Major Topics to be Covered (the task perspective)

- Modeling of motifs and *cis*-regulatory modules
- Identification of transcription factor binding sites
- Transcriptome quantification
- Transcriptome assembly
- RNA sequence and structure modeling
- Regulatory information in epigenomic data
- Genotype analysis and association studies
- Quantitative Trait Locus (QTL) Analysis
- Mass spectrometry peptide and protein identification
- Pathways in cellular networks
- Large-scale sequence alignment
Motif Modeling

What sequence motif do these promoter regions have in common?
What configuration of sequence motifs do these promoter regions have in common?

**cis-Regulatory Modules**
Transcriptome Analysis with RNA-Seq

What genes are expressed and at what levels?
Transcriptome assembly

https://tacorna.github.io/
RNA Sequence and Structure Modeling

How can we identify sequences that encode this RNA structure?
Noncoding Genetic Variants

How do genetic variants outside protein coding regions impact phenotypes?

Genotype to Phenotype

Genotype vs. Phenotype
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.
Quantitative Trait Locus (QTL) analysis
Proteomic Analysis with Mass Spectrometry
What proteins are expressed and at what levels?

Identifying Signaling Pathways

How do proteins coordinate to transmit information?

Yeger-Lotem et al., *Nature Genetics*, 2009
Large Scale Sequence Alignment

What is the best alignment of these 6 genomes?

<table>
<thead>
<tr>
<th>Genomes</th>
<th>Alignment Details</th>
</tr>
</thead>
<tbody>
<tr>
<td>Escherichia coli K-12 Strain MG1655</td>
<td>Alignment 1</td>
</tr>
<tr>
<td>Escherichia coli O157:H7</td>
<td>Alignment 2</td>
</tr>
<tr>
<td>Escherichia coli O157:H7</td>
<td>Alignment 3</td>
</tr>
<tr>
<td>Escherichia coli</td>
<td>Alignment 4</td>
</tr>
<tr>
<td>Shigella flexneri</td>
<td>Alignment 5</td>
</tr>
<tr>
<td>Shigella flexneri 2a str. 301</td>
<td>Alignment 6</td>
</tr>
</tbody>
</table>
Machine learning and integrative analysis

- **Genotype**
- **AGE**
- **BPD**

- **Phenotype**
- **Genes**
- **Modules**

- **Higher-order groupings** (e.g., pathways, circuits)

- **Cell types**
- **Modules**
- **Regulatory elements**
- **Genes**

- **Genomic elements** (e.g., GWAS, differentially expressed genes)

- **Biological Interpretation**

**Interactions between elements**
(e.g., co-expression, regulation)

**Integrative & Predictive model**
(e.g., deep neural network)

**Genomic elements** (e.g., GWAS, differentially expressed genes)

**Raw data**
- Pre-processing

**Clean data**
- AC GT C
- GC GT A
- GT C C G
- TT AG T
- CG TA G
- GA GA A

**Features**

**Model**

**Results**

- Inspired from Ideker & Lauffenburger, Trends in Biotechnology, 2003
Other Topics

• Many topics we aren’t covering
  – Protein structure prediction
  – Protein function annotation
  – Metagenomics
  – Metabolomics
  – Graph genomes
  – Single-cell sequencing
  – Text mining
  – 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

• ComBEE Python Study Group
  – https://combee-uw-madison.github.io/studyGroup/

• Many relevant seminars on campus