Advanced Bioinformatics
Biostatistics & Medical Informatics 776
Computer Sciences 776
Spring 2024

Daifeng Wang
daifeng.wang@wisc.edu
www.biostat.wisc.edu/bmi776/

These slides, excluding third-party material, are licensed under CC BY-NC 4.0 by Mark Craven, Colin Dewey, Anthony Gitter and Daifeng Wang
Agenda Today

• Introductions
• Course information
• Overview of topics
Course Web Site

- [www.biostat.wisc.edu/bmi776/](http://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
- Class Location: 2540 Engineering Hall
- Class Times: Mondays & Wednesdays 9:30 – 10:45 am CST
- Office for Office Hours: Medical Sciences Center 4750
- Office Hours: Mondays & Wednesdays 11:00-12:00 am CST and Zoom (by appointments)
- Associate Professor in the Department of Biostatistics & Medical Informatics and Investigator in Waisman Center
My research

- 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 the Office for in-person Office Hours: MSC 4750

- **very** confusing building
- best bet: use **420 North Charter St entrance**
Course TA:
Haotian Ma

- Email: hma232@wisc.edu
- Ph.D. student
  - Computer Science
- Office Hours
  - Monday: 5 – 6 pm
  - Wednesday: 5 – 6 pm (Zoom)

- Grading student assignments and providing feedback

The Course TA is here to help support you!
Please do reach out with any questions!
Office Hours

• Mon 11:00-12:00am, Wed 11:00-12:00am
  – MSC 4750
• Will begin next week
• Free to schedule an individual meeting
  – Waisman Center or Zoom
• You are encouraged to visit our office hours!
You

• So that we can all get to know each other better, please tell us your (by email)
  – 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: Wednesday, March 20, in class
• Final: Monday May 9, 5:05 PM – 7:05 PM

• 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

• Resources:
  – Previous TA Saniya Khullar prepared a video on working on a remote server (pushing, pulling, running files remotely to Biostat servers, WiscVPN): Video Link along with other helpful videos on Servers
  – 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/spring2024/bmics776](https://piazza.com/wisc/spring2024/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 book


- However, many readings from recent papers and materials
Recommended online reading

• Translational Bioinformatics
  – [https://collections.plos.org/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

Things to be familiar before you start

Biology

- DNA
- RNA
- Protein
- Gene expression
- Cell
- ......

Programming

- Python
- print ("Hello BMI 776!")
- import math
- length = r*math.pi
- ......

Informatics

- Clustering
- Classification
- Markov chain
- Conditional Probability
- ...

22
## Course Schedule

<table>
<thead>
<tr>
<th>Week</th>
<th>Date</th>
<th>Subject</th>
<th>Notes</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1/24</td>
<td>Course Overview</td>
<td>HW0 out</td>
</tr>
<tr>
<td>2</td>
<td>1/29</td>
<td>RNA Analysis &amp; Gene Discovery</td>
<td></td>
</tr>
<tr>
<td></td>
<td>1/31</td>
<td>RNA Analysis &amp; Gene Discovery</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>2/5</td>
<td>Machine Learning in Bioinformatics</td>
<td>HW0 due, HW1 out</td>
</tr>
<tr>
<td></td>
<td>2/7</td>
<td>Machine Learning in Bioinformatics</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>2/12</td>
<td>Single-cell RNA-seq Analysis</td>
<td></td>
</tr>
<tr>
<td></td>
<td>2/14</td>
<td>Single-cell RNA-seq Analysis</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>2/19</td>
<td>Single-cell RNA-seq Analysis</td>
<td></td>
</tr>
<tr>
<td></td>
<td>2/21</td>
<td>Network biology</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>2/26</td>
<td>Network biology</td>
<td>HW1 due, HW2 out</td>
</tr>
<tr>
<td></td>
<td>2/28</td>
<td>Epigenomics</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>3/4</td>
<td>Epigenomics</td>
<td></td>
</tr>
<tr>
<td></td>
<td>3/6</td>
<td>Epigenomics</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>3/11</td>
<td>Deep learning</td>
<td></td>
</tr>
<tr>
<td></td>
<td>3/13</td>
<td>Deep learning</td>
<td>HW2 due, HW3 out</td>
</tr>
<tr>
<td>9</td>
<td>3/18</td>
<td>Deep learning</td>
<td></td>
</tr>
<tr>
<td></td>
<td>3/20</td>
<td>Midterm exam</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Spring Break (3/23-31)</td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>4/1</td>
<td>Protein Structure &amp; AlphaFold</td>
<td></td>
</tr>
<tr>
<td></td>
<td>4/3</td>
<td>Protein Structure &amp; AlphaFold</td>
<td>proposal due</td>
</tr>
<tr>
<td>12</td>
<td>4/8</td>
<td>Learning Motif models</td>
<td></td>
</tr>
<tr>
<td></td>
<td>4/10</td>
<td>Learning Motif models</td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>4/15</td>
<td>Genotype Analysis</td>
<td></td>
</tr>
<tr>
<td></td>
<td>4/17</td>
<td>Genotype Analysis</td>
<td>HW3 due, HW4 out</td>
</tr>
<tr>
<td>14</td>
<td>4/22</td>
<td>Genotype Analysis</td>
<td>Progress report due</td>
</tr>
<tr>
<td></td>
<td>4/24</td>
<td>Advanced topics</td>
<td></td>
</tr>
<tr>
<td>15</td>
<td>4/29</td>
<td>Project Presentations</td>
<td></td>
</tr>
<tr>
<td>16</td>
<td>5/1</td>
<td>Project Presentations</td>
<td>HW4 due</td>
</tr>
<tr>
<td></td>
<td>5/9</td>
<td>Final exam</td>
<td>5:05 PM - 7:05 PM</td>
</tr>
<tr>
<td></td>
<td>5/11</td>
<td>Project report due</td>
<td>23:59 PM</td>
</tr>
</tbody>
</table>
Prerequisites

• BMI/CS 576 or equivalent
• Knowledge of basic biology and methods from that course will be assumed
• Knowledge of basic machine learning background
• May want to go over the material on the 576 website to refresh
• [http://www.biostat.wisc.edu/bmi576/](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
- Network flow algorithms
- Multiple hypothesis testing correction
- Classification
- Deep learning
- AlphaFold
- Linear programming
- Clustering
- More machine learning approaches
Major Topics to be Covered (the task perspective)

- Modeling of motifs and \textit{cis}-regulatory modules
- Identification of transcription factor binding sites
- Transcriptome quantification
- Transcriptome assembly
- Regulatory information in epigenomic data
- Genotype analysis and association studies
- Pathways in cellular networks
- Single cell RNA-seq
- Gene regulatory network
Motif Modeling

What sequence motif do these promoter regions have in common?
What configuration of sequence motifs do these promoter regions have in common?
Transcriptome Analysis with RNA-Seq

What genes are expressed and at what levels?
Transcriptome assembly

RNA-Seq Reads → Align reads to genome → Assemble reads into transcripts → Ab initio assemblies → Merge into meta-assembly (i.e., run TACO) → Changepoint detection → Meta-assembly

https://tacorna.github.io/
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

Chr1 20,796,219

EH37E0074243  EH37E0074249

CAMK2N1

rs112660177
Identifying Signaling Pathways

How do proteins coordinate to transmit information?

Yeger-Lotem et al., *Nature Genetics*, 2009
Cell-type

gene regulatory networks

- Cell-type-specific GRNs would be key tools for the study of cellular heterogeneity

- Cell-type-specific GRNs will reveal key regulatory factors and circuits for specific cell types, facilitating mapping between disease-associated variants and affected cell types
Machine Learning for Multiomics

\[ \min_{f} \ell(f) + \Omega(f) \]

Xu. Genome Bio. 2019
Other Topics

• Many topics we aren’t covering
  – RNA structures
  – Metagenomics
  – Metabolomics
  – Graph genomes
  – Mass Spectrometry
  – 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