CS 466
Introduction to Bioinformatics

Instructor: Jian Peng
Teaching Assistant: Wei Qian
Introduction

Instructor:
- Jian Peng
  Office location: 2118 SC
  Office hour: Mon/Wed, 11:00am-11:59am
  Email: jianpeng@illinois.edu
- My own research:
  Computational Biology and Machine Learning

Teaching Assistant:
- Wei Qian, PhD student
  Office location: 1218 SC
  Office hour: Tue/Thu, 1:00pm-2:00pm
  Email: weiqian3@illinois.edu
- Wei’s research:
  Machine Learning and Biological Data Analysis
Prerequisites

• Programming skills (equivalent to CS 225) for doing the mini-project.
• Knowledge of basic probability and statistics for understanding several lectures.
• No biology background is necessary.
Course logistics

• Course website: https://courses.engr.illinois.edu/cs466/sp2018/

• Piazza website: https://piazza.com/illinois/spring2018/cs466/home

• Lecture slides will be released before each class.
• Participation is encouraged.
• Come to class having read the day’s lecture slides and reading assignments, if any.
Course Objectives

Introduction to bioinformatics

• Basic problems in computational biology
• Statistics and machine learning for data analysis
• Algorithms for data processing

Learning to do research

• Course project experience
• Hands-on practice with real datasets
• Propose and perform independent research
Grading

For 3-credit students
- Five problem sets (30%)
- Midterm (25%)
- Final (25%)
- Team-based mini-project and report (20%)

For 4-credit students
- Five problem sets (20%)
- Midterm (25%)
- Final (25%)
- mini-project + individual report (30%)
Assignments

• See the University Policy on Academic Integrity, especially the section on plagiarism.
• Late submission within 3 days (72 hours) is worth 80% credit.
• A student may request an extension of 3 days at most once in the semester.
Course Project

Computational techniques
• Comparing algorithms
• Efficient implementation of algorithms that scale on large datasets
• New probabilistic models for biological data

Biological problems
• Comparative analysis
• Interesting data analysis
• New computational biological problems
Course Project

• Team size
  • One or two (4-credit students)
  • Up to four (3-credit students)
  • Make clear your contribution in the project report

• Implementation
  • Put your code/data on Github
  • Get your hands dirty and work on real-world datasets
Grading

Approximate data from a recent offering:
• Enrollment (who completed course): 43
• 27 A grades (2 A+, 23 A, 2 A-)
• 16 B grades (10 B+, 6 B)

This is not a statement about what the distribution this semester will be.
Questions about the course logistics?
Introduce yourself
Bioinformatics

• Is not about one problem (e.g., designing better computer chips, better compilers, better graphics, better networks, better operating systems, etc.)

• Is about a family of very different problems, all related to biology, all related to each other

• How can computers help solve any of this family of problems?
Bioinformatics and You

• You can learn the tools of bioinformatics
• These tools owe their origin to computer science, information theory, probability theory, statistics, etc.
• You can learn the language of biology, enough to understand what the problems are
• You can apply the tools to these problems and contribute to science
Important Biological Questions?

“Why do humans have so few genes?”

“Can we understand DNA code?”

“Can we understand gene function?”

“How did cooperative behavior evolve?”

“Can we cure cancer?”

……
What does biological data look like?

Sequence data

- Protein/DNA sequence
- Probabilistic models for sequences
- Dynamic programming

Matrix data

- Gene expression
- Dimensionality reduction and feature selection
- PCA and clustering
Biological Data

Graph data

- Molecular interaction networks
- Graph algorithms

Heterogeneous data

- Dimensionality reduction
- Probabilistic models for data integration
- Network-based data integration
Please read “Molecular Biology for Computer Scientists” by Lawrence Hunter
Examples of my research projects
Recent research

Cell Systems, 2016

Cell Systems, 2017

Nature Communications, 2017

Cell Systems, 2018
Protein sequence, structure and function
Network analysis for disease modeling

Human disease network

New disease biology (potential drug targets)

Human disease network

Validation: Q-PCR & Bioscreen
Pharmacogenomics and cancer genomics

Figure from the DREAM challenge website