Welcome to CSE 527
Computational Biology

Lecture 1 - Sep 29, 2015
CSE 527 Computational Biology
Instructor: Su-In Lee
TA: Javad Hosseini
TTh 12:00-1:20 @ SAV 130

Introducing teaching staff

- Professor: Su-In Lee
  - Computer Science & Engineering, Genome Sciences
  - Office hour: Mon 2-3:20 pm
  - Office: Rm 536, Paul Allen Center (CSE building)

- Teaching assistants
  - Javad Hosseini, CSE PhD student
  - Office hour: Wed 2:30-3:50pm @ PAC 220

CS and biology seem to have very different goals and methodologies

Computational Biology

- Image of computer with keyboard and circuit board
- Image of cell with organelles
- Image of DNA strand
Biology is an information science

DNA (3 billion-long string)
AGCTATACGATGACTACAGACAGC
ATACAGACATTTTTAAAAAGCAGACA
AAAATGCTAAAACAGGGTTTACTAG
GACACTTACAACTAGT...

DNA acts as the “brain” of the cell, telling the cell how to properly grow and work.

Each individual has a slightly different version of a DNA sequence

Each species has a unique DNA sequence.

DNA tells the cell how to perform various tasks in a cell

Gene regulation
RNA degradation
Gene interaction map ("social network" of genes)
A cell’s biological state can be described by millions of numbers

**Gene** (~20,000 in human)

DNA

RNA

Protein

### Biological information in the 21st century
- DNA sequence: >1M letters known to differ among individuals.
- RNA expression levels of 20K genes
- Protein levels of 20K genes
- Each DNA letter can be biochemically modified
- How many numbers? Definitely >1M!

Modern biology is about mining very large, complex data

- Interesting questions arise such as
  - How the DNA sequences are different between human and chimpanzee?
  - Which parts of the DNA determine susceptibility to dementia?
  - How epigenetic modifications to DNA influence cell functions?
  - How the “social network” of genes are different between cancer and normal cells?
- CS and Statistics play a key role!

Course structure

**Topics to be discussed**

- **Part 1: Sequence analysis**
  - How the DNA sequences are different between human and chimpanzee?
- **Part 2: Genetics**
  - Which parts of the DNA determine susceptibility to dementia?
- **Part 3: Epigenomics**
  - How epigenetic modifications to DNA influence cell functions?
- **Part 4: Systems biology**
  - How the “social network” of genes are different between cancer and normal cells?

Course structure

**Topics to be discussed**

- **Part 1: Sequence analysis**
- **Part 2: Genetics**
- **Part 3: Epigenomics**
- **Part 4: Systems biology**

**Methods broadly applicable to other applications**

- Dynamic programming, expectation maximization (EM) algorithm, hidden Markov models, Bayesian networks, clustering, classification
Prerequisites

- No background in biology is required
- Familiar with:
  - Basic statistics
  - Probability theory
  - Basic ML methods (e.g., linear regression) (optional)
- Programming skills in MATLAB, R, JAVA, C++, Perl, or Python
- Please contact us if you are unsure of our expectations

LOGISTICS

Course resources

- Course website: [http://cs.washington.edu/527](http://cs.washington.edu/527)
  - Lecture notes, readings, HW assignments
  - Due dates for homework
  - Discussion board, HW dropbox and gradebook
- Mailing list: cse527a_au15@uw.edu

Lecture notes

- Lecture slides will become available in the morning before the class
  - [CSE 527 Computational Biology](#)
Readings

- No required textbook
- Readings will include papers and books listed on course website
  - Biological background
    - The Cell, a molecular approach by Copper
    - Genetics, from genes to genomes by Hartwell and more
    - Principles of Population genetics by Hartl & Clark
  - Computational background
    - Probabilistic graphical models by Daphne Koller & Nir Friedman
    - Andrew Ng’s machine learning lecture note (cs229.stanford.edu)

Homework

- 4 homework assignments
- Due at 11:59pm on the assigned due date
- Allowed 3 free late (calendar) days
- Collaboration allowed, but write-ups and coding must be done individually
  - Please write down team members’ names

Project

- Individual, or teams of two
- Topic
  - Can be connected to your research
  - Open-ended
- Successful project reports will be submitted to ML/compbio conferences or journals

Project deliverables

- Project proposal (1 page) – 10/19
- Midterm report (up to 2 pages) – 11/16
- Presentation (conference style) – final week
- Final report (up to 8 pages) – 12/16
Grading

- 4 HWs (70%)
- Final project (25%)
  - Proposal (5%)
  - Midterm report (5%)
  - Presentation (5%)
  - Final report (10%)
- Attendance/participation (5%)
  - Good questions/answers in class

Course schedule

<table>
<thead>
<tr>
<th>HW assignments</th>
<th>Project assignments</th>
</tr>
</thead>
<tbody>
<tr>
<td>Week 1 HW1 out</td>
<td>Project proposal</td>
</tr>
<tr>
<td>Week 2 HW1 due, HW2 out</td>
<td>About 1 month</td>
</tr>
<tr>
<td>Week 3 HW2 due, HW3 out</td>
<td>Midterm report</td>
</tr>
<tr>
<td>Week 4 HW3 due, HW4 out</td>
<td>About 1 month</td>
</tr>
<tr>
<td>Week 5 HW4 due</td>
<td>Presentation</td>
</tr>
<tr>
<td>Week 6</td>
<td>Final report due 3/21</td>
</tr>
<tr>
<td>Week 7</td>
<td></td>
</tr>
<tr>
<td>Week 8</td>
<td></td>
</tr>
<tr>
<td>Week 9</td>
<td></td>
</tr>
<tr>
<td>Week 10</td>
<td></td>
</tr>
</tbody>
</table>

Please start working on the project as early as possible!