Introduction

CS482/682
Computational Techniques in Biological Sequence Analysis
Outline

• Course logistics
• A few example problems
Course staff

• Instructor: Bin Ma (DC 3345, http://www.cs.uwaterloo.ca/~binma)

• Course webpage: https://cs.uwaterloo.ca/~binma/cs482
Textbooks, notes

• No required textbooks.
• Lecturing ppt available at website before lecturing with best efforts.
  – These are my lecturing outline rather than course notes.
  – Cannot replace attending the lectures.
Why Do People Do Bioinformatics?

• A primary drive is human health.
  – A lot of human diseases are related to genetics.
  – [http://youtube.com/watch?v=-GwdZIqJf8g](http://youtube.com/watch?v=-GwdZIqJf8g)
  – Others related to proteomics, metabolomics, glycomics, lipidomics, ...
  – [https://www.youtube.com/watch?v=GD0gcZoqtcM](https://www.youtube.com/watch?v=GD0gcZoqtcM)
• Human Genome Project
  – 3B$ from 1990-2003 to study human genome.
  – identify all the approximately 20,000-25,000 genes in human DNA,
  – determine the sequences of the 3 billion chemical base pairs that make up human DNA,
• 1000 Genome Project
• 1 million genome project (several of them)
• Human Proteome Organization (HUPO)
• Alphabet’s companies Verily and Calico.
Objectives of This Course

• Know bioinformatics
  – Purpose and method
  – General topics
• Learn classic problems and algorithms in bioinformatics
• Learn wide-applicable computational techniques
  – String algorithms
  – Hidden Markov Model
  – Log likelihood ratio score
  – Statistical validation
  – A bit of machine learning
• Sample the research style in bioinformatics
Grades

- 4 assignments: 40% = 10+10+10+10
- In class midterm: 20%
- Final exam: 40%
Assignments

• All are programming assignments. You submit source code and a half-page document.
  – Based on your own work. Use of library needs to be documented.

• Evaluation is mostly based on correctness and the performance of the program (speed, accuracy, etc.).
A Typical Problem

• Human genome has \( \sim 3G \) base pairs (letters).

> A substring of the genome

```
GCTTACACGTCAACCATCTGTGCCACCACCCCATGTCTCTTAGTGAT
CCCTCATAAGTTCCAACACAAAGTTTTCGAGTACTCAACACCAACA
TTGATGGGCAATGGAAATAGCCTTCGCCATCACCACATTAAGG
GTGATGTGTAGGAAAGCAGACATTGACCTCACCAGAGGGGCAAG
CGAGCTCAGGGTAGGTAGGGGTGGGAGCATATGATCACCATCATA
AGAACTCACCAAGATTCAGACTGGTTC
```

• Only 1% encode genes.
  – where are they?
A Human Gene Finder

Listen to this part: 'AGGCTAATCGCATAACTG'

Wow!
An Analog

• Find the English words.

fbjpsikocxltfestkvdvjiixjsasisxmhbpvpwb
ulfddurluvwrrirtrbshcpeyhbekydaibmwytj
nwvporswuanhvsdgknpkzihjqagrpispixtzqphhk
twbiinformaticsisusefulcrsqibcadosyvoz
vhuzdxabqrjzfagiysqfcmyrkqdytjtjbusysqga
etylklzbinma@uwaterloo.cawpxuprgokixkoiyv
Another Typical Problem

• Find the longest shared substring between human and mouse genomes
  – Each has $3 \times 10^9$ base pairs.
  – Cannot afford $3 \times 10^9 \times 3 \times 10^9$ comparisons.
Longest Common Substring

• Longest Common Substring can be done in linear time!

• What if all similarities (instead of exact matches) are to be found?
  – This leads to the homology search problem.
  – Some of the key techniques are developed by profs in this school.
A Third Typical Problem

<table>
<thead>
<tr>
<th>R002760</th>
<th>AUBL_BOVIN</th>
<th>259.02</th>
<th>39</th>
<th>50</th>
<th>69205.547</th>
<th>Serum albumin OS=Bo taurus GN=ALB</th>
</tr>
</thead>
<tbody>
<tr>
<td>Q29431</td>
<td>TRPE_BOVIN</td>
<td>253.39</td>
<td>57</td>
<td>50</td>
<td>77733.203</td>
<td>Sorotransferrin OS=Bo taurus GN=TF</td>
</tr>
<tr>
<td>B13720</td>
<td>BGAL_ECOLC</td>
<td>242.52</td>
<td>36</td>
<td>5</td>
<td>116416.711</td>
<td>Beta-galactosidase OS=Escherichia coli</td>
</tr>
<tr>
<td>B7U32</td>
<td>BGAL_ECO27</td>
<td>234.04</td>
<td>41</td>
<td>5</td>
<td>116511.883</td>
<td>Beta-galactosidase OS=Escherichia coli</td>
</tr>
<tr>
<td>B1L99</td>
<td>BGAL_ECOGM</td>
<td>229.52</td>
<td>32</td>
<td>0</td>
<td>116541.969</td>
<td>Beta-galactosidase OS=Escherichia coli</td>
</tr>
<tr>
<td>B7N07</td>
<td>BGAL_ECOUL</td>
<td>226.14</td>
<td>32</td>
<td>0</td>
<td>116530.448</td>
<td>Beta-galactosidase OS=Escherichia coli</td>
</tr>
<tr>
<td>P00030</td>
<td>ADH1_BOVIN</td>
<td>163.56</td>
<td>16</td>
<td>12</td>
<td>36899.184</td>
<td>Alcohoh dehydrogenase 1 OS=Equus f ...</td>
</tr>
<tr>
<td>P49468</td>
<td>BGC_CITK3</td>
<td>146.05</td>
<td>10</td>
<td>0</td>
<td>116279.383</td>
<td>Beta-galactosidase OS=Citrobacter kios ...</td>
</tr>
<tr>
<td>P08691</td>
<td>YSC_CHECK</td>
<td>146.46</td>
<td>8</td>
<td>5</td>
<td>16238.638</td>
<td>Lysozyme C OS=Gallus gallus GN=LYZ</td>
</tr>
<tr>
<td>PH5803</td>
<td>EYC_BOVIN</td>
<td>136.21</td>
<td>9</td>
<td>9</td>
<td>11763.332</td>
<td>Cytochrome c OS=Bo taurus GN=CYCS</td>
</tr>
<tr>
<td>R09571</td>
<td>TRPE_PIG</td>
<td>123.60</td>
<td>8</td>
<td>1</td>
<td>76967.983</td>
<td>Sorotransferrin OS=Sus scrofa GN=TF</td>
</tr>
<tr>
<td>R00331</td>
<td>ADH2_BOVIN</td>
<td>121.28</td>
<td>5</td>
<td>1</td>
<td>36731.918</td>
<td>Alcohol dehydrogenase 2 OS=Equus f ...</td>
</tr>
<tr>
<td>A77720</td>
<td>BFC_BOVIN</td>
<td>115.40</td>
<td>9</td>
<td>9</td>
<td>56370.816</td>
<td>Gluteraldehyde OS=Escherichia coli O13</td>
</tr>
</tbody>
</table>

[Image: Picture of a hand holding a test tube with blood.]
Peptide Identification

peptide sequence: LGSSEVEQVQLVVDGVK

tandem mass spectrometry:

MS/MS spectrum
Assignments

• 1. Pairwise sequence alignment.
  – Every Bioinformatics course does this.
• 2. Is it a natural/real peptide?
  – A taste of scoring and machine learning.
• 3. NGS Reads mapping.
  – Index and search in gigabytes of data with a PC.
• 4. Peptide identification from mass spec.
  – Can you write best-selling commercial software?
A brief review of biology

- Modern molecular biology studies a few types of biologically important molecules: DNA, RNA, protein, lipid, glycan
- Bioinformatics has mostly studied DNA, then RNA and protein.
  - Because they are “easier”
  - their primary structures are sequences.
  - Also because the measuring technology has been developed.
Central Dogma of Molecular Biology

- DNA Polymerase (replication: DNA -> DNA)
- RNA Polymerase (transcription: DNA -> RNA)
- Ribosome (translation: RNA -> Protein)
DNA

- The most important information carrying molecule that passes information to children
  - responsible to many genetic diseases.
- The simplest to model in a computer
  - DNA is modeled as a string over \{A,C,G,T\}
RNA

• RNA’s structure is important to RNA’s function. Not a simple string anymore.
Protein

Primary structure is a sequence.
20 frequent amino acids.
Often have post-translational modifications.
Fold into a complex 3D structure.
Read more by yourself

• If you don’t have much biology background, read the following articles (and other related articles) from wikipedia

• Protein, DNA, RNA, gene, genome, genetic code.

• We will briefly review the necessary biology knowledge when needed.
Bioinformatics and Biology

• How bioinformatics is used in biology
  • Bio samples → data → software → discovery

• Bioinformatics research cycle:
  • biological problem → math model → algorithm → software → biology

• Sometimes the problem is too hard
  – NP-hard.

• Sometimes the data is too large.
  – polynomial time is no good any more.
  – Most of time, there are data errors.
  – Garbage in, knowledge out?
Bioinformatics General Topics

Bioinformatics is a very broad area now:

- **Sequence comparison**
- **Phylogeny**
- **Gene prediction** and annotation
- **Motif finding**
- **Protein identification and quantification with mass spectrometry**
- Genome sequencing. **Next generation sequencing.**
- Genetic variations: SNPs, alternative splicing, and diseases.
- Gene expression and biomarker
- Regulatory network
- Protein-protein interaction network
- Protein structure comparison and prediction
- Protein design
- RNA structure, RNA gene prediction, RNAi.
- ......
Summary

• We talked about:
  • course logistics
  • basic biology (wikipedia good resource)
  • course topics

• Next time: sequence alignment