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.
Bioinformatics

The use of computing to help solve biological problems.
Biology Can be Studied at Different Scales

Organisms: living things

Organs and tissues
Cell Level

Genetic information storage
Production of functional proteins from genetic info.

The Animal Cell
Molecular Level

DNA: chain of nucleotide bases

...GCTTACACGTCAACCAT...

Protein: chain of amino acids.

...LVQSGAEVKKP...
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=-GwdZlqJf8g](http://youtube.com/watch?v=-GwdZlqJf8g)
  – New treatment of diseases require in-depth understanding (information) about the body system at molecular level.
  – [https://www.youtube.com/watch?v=Q1CwARpnfe8](https://www.youtube.com/watch?v=Q1CwARpnfe8)
Difference Made by Bioinformatics

• Example: genome sequencing.
• 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.
  – Bioinformatics played an essential role in analyzing the data and assemble the genome.
• Today one can sequence a human’s genome with 1000$ in a couple of weeks. Bioinformatics is the key to utilize the NGS (next generation sequencing) data for genome sequencing.
• As such, today’s cancer treatment starts to become personalized.
New Data and Big Data

• Big data
  – 1M genome project.
  – Single cell sequencing.

• New data
  – Proteomics

• 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 ~ 3G base pairs (letters).

> A substring of the genome
GCTTACACGTCACCATCTGTGCCACCCACCCCATGTCTCTTAGTGAT
CCCTCATAAGTTCCAACAAAGTTTTCGAGTACTCAACACACCCAACA
TTGATGGGCAATGGAAAATAGCCTTCGCCATCACACCATTAAGG
GTGATGTTGAGGAAGGCAGACATTGACCTCACCAGAGAGGGGAGG
CGAGCTCAGGTAGGATGAGGTGGAGCATATGATCACCATCATAC
AGAACTCAACCAAGATTTCCAGACTGGTTC

• Only 1-2% encode proteins (genes).
  – where are they?
A Human Gene Finder

Listen to this part: 'AGGCTAATCGCATAACTG'

Wow!
An Analog

• Find the English words.

fbjpsikocxltfestkvdvjiixjsasisxmhbpvpwb
ulfddurluvwrriritrbsbhcpeyhbekydaibmwfyntj
nwvporabwuahvsdgnpkzihjqagrpspixtzqphhk
tvwbioinformaticsisusefulcrsqiqbcadosyvoz
vhuzdxabqrrjzfagiysqfcmyrkqdytjtjbusysqga
etyllzbinma@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>RO325G</th>
<th>AUSHI_RBOVIN</th>
<th>359.02</th>
<th>268.42</th>
<th>416</th>
<th>57</th>
<th>50</th>
<th>69233.547</th>
<th>Serum albumin OS=A soltana GN=ALB</th>
</tr>
</thead>
<tbody>
<tr>
<td>Q29411</td>
<td>TRYFE_BOVIN</td>
<td>253.39</td>
<td>267.05</td>
<td>168</td>
<td>57</td>
<td>30</td>
<td>77733.203</td>
<td>Sorotransferrin OS=B taurus GN=TP</td>
</tr>
<tr>
<td>B11075</td>
<td>RGapE_RCOIL</td>
<td>242.52</td>
<td>267.05</td>
<td>168</td>
<td>41</td>
<td>5</td>
<td>116511.883</td>
<td>Beta galactosidase OS=Echerichia col</td>
</tr>
<tr>
<td>B7UJD5</td>
<td>RGapE_RCOIL</td>
<td>234.04</td>
<td>267.05</td>
<td>168</td>
<td>36</td>
<td>0</td>
<td>116416.711</td>
<td>Beta galactosidase OS=Echerichia col</td>
</tr>
<tr>
<td>B7UDM9</td>
<td>RGapE_RCOIL</td>
<td>229.52</td>
<td>267.05</td>
<td>168</td>
<td>32</td>
<td>0</td>
<td>116541.969</td>
<td>Beta-galactosidase OS=Echerichia col</td>
</tr>
<tr>
<td>B7UKQ1</td>
<td>RGapE_RCOIL</td>
<td>226.24</td>
<td>267.05</td>
<td>168</td>
<td>32</td>
<td>0</td>
<td>116350.648</td>
<td>Beta galactosidase OS=Echerichia col</td>
</tr>
<tr>
<td>P00310</td>
<td>ADH1_YEAST</td>
<td>168.36</td>
<td>51</td>
<td>16</td>
<td>16</td>
<td>12</td>
<td>36849.184</td>
<td>Alcohol dehydrogenase 1 OS=Saccharo</td>
</tr>
<tr>
<td>A3V90B</td>
<td>RGapE_CITIB</td>
<td>146.05</td>
<td>51</td>
<td>16</td>
<td>16</td>
<td>12</td>
<td>36849.184</td>
<td>Alcohol dehydrogenase 1 OS=Saccharo</td>
</tr>
<tr>
<td>P00688</td>
<td>YSCCHKERRQ</td>
<td>146.46</td>
<td>51</td>
<td>16</td>
<td>16</td>
<td>12</td>
<td>36849.184</td>
<td>Alcohol dehydrogenase 1 OS=Saccharo</td>
</tr>
<tr>
<td>P02891</td>
<td>YSC朐H_BOVIN</td>
<td>136.21</td>
<td>51</td>
<td>16</td>
<td>16</td>
<td>12</td>
<td>36849.184</td>
<td>Alcohol dehydrogenase 1 OS=Saccharo</td>
</tr>
<tr>
<td>P09571</td>
<td>TRYFE_PIG</td>
<td>123.60</td>
<td>51</td>
<td>16</td>
<td>16</td>
<td>12</td>
<td>36849.184</td>
<td>Alcohol dehydrogenase 1 OS=Saccharo</td>
</tr>
<tr>
<td>P00311</td>
<td>ADH2_YEAST</td>
<td>121.28</td>
<td>51</td>
<td>16</td>
<td>16</td>
<td>12</td>
<td>36849.184</td>
<td>Alcohol dehydrogenase 2 OS=Saccharo</td>
</tr>
<tr>
<td>P02890</td>
<td>RGapE_YEAST</td>
<td>115.40</td>
<td>51</td>
<td>16</td>
<td>16</td>
<td>12</td>
<td>36849.184</td>
<td>Alcohol dehydrogenase 2 OS=Saccharo</td>
</tr>
</tbody>
</table>
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.
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 \(\rightarrow\) data \(\rightarrow\) software \(\rightarrow\) discovery

• Bioinformatics research cycle:
  • biological problem \(\rightarrow\) math model \(\rightarrow\) algorithm
    \(\rightarrow\) software \(\rightarrow\) biology

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

• Sometimes the data is too large.
  – polynomial time is no good here ...

  – Most of time, there are data errors.
    – Garbage in, knowledge out?
Bioinformatics General Topics

Bioinformatics is a very broad area now. We sample only the underlined topics in this course.

• **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