**Announcement**

**Article 2 Summaries/Critiques**

For full credit, you needed to include at least one technical detail about the methodology in BLAST that makes it unique and distinguishable from all previous methods (at that time):

- Threshold (T score)
- MSP
- Bitwise searching
- Extending a match
- FSM

The critique needed to be constructive. Saying “the article was difficult to read” is not enough. Instead ...

- “Add more pseudo code”
- “Additional visual explanations of the algorithm would have been helpful”
- “Including a full step-by-step explanation of BLAST on a small example would make the algorithm easier to understand”
Announcement

Article 3 Summaries/Critiques


Your goal is to write a summary and critique that is easily digestible by somebody who has NOT read the article. Still focusing on the WHAT, WHY, and HOW
Announcement

Exam

Take home (over 2 nights), or in-class?

Schedule

<table>
<thead>
<tr>
<th>Week</th>
<th>Date(s)</th>
<th>Lectures</th>
<th>Article Reading</th>
<th>Lab(s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>04 April</td>
<td>Watson &amp; Crick, DNA</td>
<td>Lab 1: Online databases</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>11 April</td>
<td>Lipman et al., BLAST</td>
<td>Lab 2a: Sequence Alignment, online</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>18 April</td>
<td>Lab 2b: Sequence Alignment, coding</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>25 April</td>
<td>Lab 3a: MSA, online</td>
<td>Lab 3b: MSA, scripting</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>02 May</td>
<td>Lab 4a: MSA, online</td>
<td>Lab 4b: MSA, scripting</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>09 May</td>
<td>Lab 5a: visualization, online</td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>16 May</td>
<td>Exam</td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>23 May</td>
<td>Project Proof of Concept</td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>30 May</td>
<td>Project Final Results</td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>06 June</td>
<td>Project Presentations</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Announcement

Exam

Take home (over 2 nights), or in-class?

Course Project

• Group formation : 2-6 May
• Proposal : 10 May (presentation)
• Proof of Concept (Canvas submission)
• Final Results (Canvas submission)
• Presentation (31 May – 03 June)
Project Idea 1: “constrained” sequence alignment

**Q:** With which genes does ACCTCGCTAGCAT align best in the human genome ...
Announcement

Project Idea 1: “constrained” sequence alignment

**Q:** With which genes does
ACCT\text{CGC}\text{TAG}\text{CAT} align best
in the human genome ... such
that \text{CGC} and \text{CAT} align
perfectly?
Announcement

Project Idea 1: “constrained” sequence alignment

Q: With which genes does ACCTCGCTAGCAT align best in the human genome ... such that CGC and CAT align perfectly?

Goal

- Implement an algorithm that achieves this task for DNA or proteins.
- Demonstrate the algorithm’s use.
- Report on its scalability and runtime performance.
From last time ... 

Mutations to the DNA sequences

- Substitution
- Deletion
- Insertion
- Inversion
- Translocation
From last time ...

Mutations to the DNA sequences

- Substitution
- Deletion
- Insertion
- Inversion
- Translocation
From last time ...

Mutations to the DNA sequences

- **Substitution**
- **Deletion**
- **Insertion**
- **Inversion**
- **Translocation**

<table>
<thead>
<tr>
<th></th>
<th>Human1</th>
<th>Pig</th>
<th>Chick</th>
<th>Human2</th>
<th>Mouse</th>
<th>Bacteria</th>
<th>FLY</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sequence</td>
<td>GLKELPMRNLQEIILHGAV</td>
<td>GLRELPMRNLQEIILQGAV</td>
<td>GLRELPMKRLSEILNGGV</td>
<td>GLQELGLKNLTEILNGGV</td>
<td>HLKELGLYNLMNITRGSV</td>
<td>DMQKIGLYSLQNITRGSV</td>
<td>KDLEIGLYNLRNITRGAV</td>
</tr>
</tbody>
</table>

Substitution Mutation
From last time ...

Mutations to the DNA sequences

- **Substitution**
- **Deletion**
- **Insertion**
- **Inversion**
- **Translocation**

Residues that have specificity to certain subfamilies of organisms are a case-by-case basis when inferring their effect. Conserved residues are most likely conserved for a very good reason, so mutations to those locations are probably **functional** mutations.
Q: What is it?
Visualization

Q: What is it?
Visualization

Q: What is it?

- Visualization of raw data
- Visualization of the analysis/interpretation of raw data

The best “visualization” is a mix of raw data and analysis of it
Visualization - Motivation

Napoleon’s Russian campaign, 1812
Visualization - Motivation

Napoleon’s Russian campaign, 1812
Visualization - Motivation

Napoleon’s Russian campaign, 1812

“Decisive” Russian Victory

Ratio of initial troops to survivors

Q: How “bad” was it
Visualization - Motivation

Napoleon’s Russian campaign, 1812
Visualization - Motivation

Napoleon’s Russian campaign, 1812
Visualization - Motivation

Napoleon’s Russian campaign, 1812
Visualization - Motivation

Napoleon’s Russian campaign, 1812

Q: What details does this plot provide?

Q: What details does this plot not provide?
Visualization - Motivation

Q: What details do this plot provide?

Q: What details do this plot not provide?

Q: What is the one big factor most often credited with causing Napoleon's defeat?
Q: How should we “combine” all four pieces of data

- Troop movement
- Troop numbers
- Temperature
- Date(s)
Q: How should we “combine” all four pieces of data

- Troop movement
- Troop numbers
- Temperature
- Date(s)

Until 3D or $nD$ ($n > 3$) tools become available, how do we visualize high-dimensional data using only 2D dimensional mediums (printed paper, standard graphics and “plots”)
Visualization - Motivation

Napoleon’s Russian campaign, 1812

Task: Identify the “data” on this plot (a) Num of troops, (b) Distance, (c) temperature
Visualization – protein and DNA sequences

HIV-1 protease is essential for the life-cycle of HIV, the retrovirus that causes AIDS. HIV protease cleaves virion pre-cursor polyproteins which then assemble to form the HIV virion.
HIV-1 protease is essential for the life-cycle of HIV, the retrovirus that causes AIDS. HIV protease cleaves virion pre-cursor polyproteins which then assemble to form the HIV virion.

Q: Does identification of the amino acids of the HIV-1 protease polypeptide chain inform us about how the protein might work?

If “no,” then what other information is needed?
Visualization – protein and DNA sequences

HIV-1 protease is essential for the life-cycle of HIV, the retrovirus that causes AIDS. HIV protease cleaves virion pre-cursor polyproteins which then assemble to form the HIV virion.

A little better ...

Unless you can visualize 99 amino acids and their atoms in 3D space, even this data set of experimentally-derived atom coordinates do not provide much information about the function of the protease.

<table>
<thead>
<tr>
<th>ATOM</th>
<th>1</th>
<th>N</th>
<th>PRO A 1</th>
<th>52.574</th>
<th>58.851</th>
<th>-7.646</th>
<th>1.00</th>
<th>34.60</th>
</tr>
</thead>
<tbody>
<tr>
<td>ATOM</td>
<td>2</td>
<td>CA</td>
<td>PRO A 1</td>
<td>51.842</td>
<td>59.784</td>
<td>-6.815</td>
<td>1.00</td>
<td>34.88</td>
</tr>
<tr>
<td>ATOM</td>
<td>3</td>
<td>C</td>
<td>PRO A 1</td>
<td>52.146</td>
<td>59.438</td>
<td>-5.356</td>
<td>1.00</td>
<td>36.01</td>
</tr>
<tr>
<td>ATOM</td>
<td>4</td>
<td>O</td>
<td>PRO A 1</td>
<td>53.031</td>
<td>58.612</td>
<td>-5.150</td>
<td>1.00</td>
<td>35.05</td>
</tr>
<tr>
<td>ATOM</td>
<td>5</td>
<td>CB</td>
<td>PRO A 1</td>
<td>59.301</td>
<td>59.581</td>
<td>-7.189</td>
<td>1.00</td>
<td>34.07</td>
</tr>
<tr>
<td>ATOM</td>
<td>6</td>
<td>CG</td>
<td>PRO A 1</td>
<td>50.353</td>
<td>58.166</td>
<td>-7.724</td>
<td>1.00</td>
<td>32.00</td>
</tr>
<tr>
<td>ATOM</td>
<td>7</td>
<td>CD</td>
<td>PRO A 1</td>
<td>51.621</td>
<td>58.242</td>
<td>-8.540</td>
<td>1.00</td>
<td>31.72</td>
</tr>
<tr>
<td>ATOM</td>
<td>8</td>
<td>N</td>
<td>GLN A 2</td>
<td>51.488</td>
<td>60.047</td>
<td>-4.359</td>
<td>1.00</td>
<td>35.47</td>
</tr>
<tr>
<td>ATOM</td>
<td>9</td>
<td>CA</td>
<td>GLN A 2</td>
<td>51.577</td>
<td>59.564</td>
<td>-2.997</td>
<td>1.00</td>
<td>35.15</td>
</tr>
<tr>
<td>ATOM</td>
<td>10</td>
<td>C</td>
<td>GLN A 2</td>
<td>50.109</td>
<td>59.273</td>
<td>-2.094</td>
<td>1.00</td>
<td>33.50</td>
</tr>
<tr>
<td>ATOM</td>
<td>11</td>
<td>O</td>
<td>GLN A 2</td>
<td>49.368</td>
<td>60.231</td>
<td>-2.829</td>
<td>1.00</td>
<td>37.45</td>
</tr>
<tr>
<td>ATOM</td>
<td>12</td>
<td>CB</td>
<td>GLN A 2</td>
<td>52.076</td>
<td>60.630</td>
<td>-2.075</td>
<td>1.00</td>
<td>37.17</td>
</tr>
<tr>
<td>ATOM</td>
<td>13</td>
<td>CG</td>
<td>GLN A 2</td>
<td>52.024</td>
<td>59.911</td>
<td>-0.749</td>
<td>1.00</td>
<td>42.02</td>
</tr>
<tr>
<td>ATOM</td>
<td>14</td>
<td>CD</td>
<td>GLN A 2</td>
<td>52.639</td>
<td>60.592</td>
<td>0.439</td>
<td>1.00</td>
<td>43.50</td>
</tr>
<tr>
<td>ATOM</td>
<td>15</td>
<td>OE1</td>
<td>GLN A 2</td>
<td>53.841</td>
<td>60.484</td>
<td>0.675</td>
<td>1.00</td>
<td>44.41</td>
</tr>
<tr>
<td>ATOM</td>
<td>16</td>
<td>NE2</td>
<td>GLN A 2</td>
<td>51.824</td>
<td>61.272</td>
<td>1.246</td>
<td>1.00</td>
<td>43.72</td>
</tr>
<tr>
<td>ATOM</td>
<td>17</td>
<td>N</td>
<td>ILE A 3</td>
<td>49.555</td>
<td>58.119</td>
<td>-2.306</td>
<td>1.00</td>
<td>30.33</td>
</tr>
<tr>
<td>ATOM</td>
<td>18</td>
<td>CA</td>
<td>ILE A 3</td>
<td>48.107</td>
<td>57.932</td>
<td>-2.171</td>
<td>1.00</td>
<td>27.03</td>
</tr>
<tr>
<td>ATOM</td>
<td>19</td>
<td>C</td>
<td>ILE A 3</td>
<td>47.924</td>
<td>57.639</td>
<td>-0.692</td>
<td>1.00</td>
<td>26.95</td>
</tr>
<tr>
<td>ATOM</td>
<td>20</td>
<td>O</td>
<td>ILE A 3</td>
<td>48.736</td>
<td>56.921</td>
<td>-0.162</td>
<td>1.00</td>
<td>25.73</td>
</tr>
<tr>
<td>ATOM</td>
<td>21</td>
<td>CB</td>
<td>ILE A 3</td>
<td>47.623</td>
<td>56.740</td>
<td>-3.053</td>
<td>1.00</td>
<td>24.30</td>
</tr>
<tr>
<td>ATOM</td>
<td>22</td>
<td>CG1</td>
<td>ILE A 3</td>
<td>47.870</td>
<td>57.072</td>
<td>-4.513</td>
<td>1.00</td>
<td>24.07</td>
</tr>
<tr>
<td>ATOM</td>
<td>23</td>
<td>CG2</td>
<td>ILE A 3</td>
<td>46.116</td>
<td>56.491</td>
<td>-2.898</td>
<td>1.00</td>
<td>22.78</td>
</tr>
<tr>
<td>ATOM</td>
<td>24</td>
<td>CD1</td>
<td>ILE A 3</td>
<td>47.651</td>
<td>55.899</td>
<td>-5.455</td>
<td>1.00</td>
<td>21.74</td>
</tr>
<tr>
<td>ATOM</td>
<td>25</td>
<td>N</td>
<td>THR A 4</td>
<td>46.933</td>
<td>58.229</td>
<td>-0.055</td>
<td>1.00</td>
<td>24.74</td>
</tr>
<tr>
<td>ATOM</td>
<td>26</td>
<td>CA</td>
<td>THR A 4</td>
<td>46.701</td>
<td>58.033</td>
<td>1.344</td>
<td>1.00</td>
<td>25.63</td>
</tr>
<tr>
<td>ATOM</td>
<td>27</td>
<td>C</td>
<td>THR A 4</td>
<td>45.538</td>
<td>57.047</td>
<td>1.421</td>
<td>1.00</td>
<td>27.68</td>
</tr>
<tr>
<td>ATOM</td>
<td>28</td>
<td>O</td>
<td>THR A 4</td>
<td>44.798</td>
<td>56.981</td>
<td>0.423</td>
<td>1.00</td>
<td>29.86</td>
</tr>
<tr>
<td>ATOM</td>
<td>29</td>
<td>CB</td>
<td>THR A 4</td>
<td>46.378</td>
<td>59.394</td>
<td>1.948</td>
<td>1.00</td>
<td>25.77</td>
</tr>
<tr>
<td>ATOM</td>
<td>30</td>
<td>CG1</td>
<td>THR A 4</td>
<td>45.232</td>
<td>59.946</td>
<td>1.314</td>
<td>1.00</td>
<td>26.30</td>
</tr>
<tr>
<td>ATOM</td>
<td>31</td>
<td>CG2</td>
<td>THR A 4</td>
<td>47.406</td>
<td>60.335</td>
<td>1.728</td>
<td>1.00</td>
<td>25.63</td>
</tr>
<tr>
<td>ATOM</td>
<td>32</td>
<td>N</td>
<td>LEU A 5</td>
<td>45.285</td>
<td>56.360</td>
<td>2.548</td>
<td>1.00</td>
<td>26.30</td>
</tr>
</tbody>
</table>
Visualization – protein and DNA sequences
Visualization – protein and DNA sequences

The visual 3D representation of HIV-1 protease permits us to “see” the protein’s structure.

And now we can ask question ... cleft ... flaps ... virion pre-cursor cells.
Q: Does the inclusion of a ligand (polypeptide) into the cleft alter the protein’s function?
Visualization – protein and DNA sequences

Q: Does the inclusion of a ligand (polypeptide) into the cleft alter the protein’s function?

(this week’s first lab)
Visualization – protein and DNA sequences

Human1  GLKELLHGAV
Pig      GLRELLQGAV

What is one approach of “visually” representing alignment (that you’ve already seen in lab)?
Q: What other commonly used visualization techniques for inferring correlations are available?

What is one approach of “visually” representing alignment (that you’ve already seen in lab)?

**Human1**

GLKELLHGAV

**Pig**

GLRELLQGAV
Visualization – protein and DNA sequences

Step 1: write sequences along top row and left-most column

Human1  GLKELLHGAV
Pig      GLRELLQGAV
Visualization – protein and DNA sequences

Human1: GLKELLHGA
Pig: GLRELLQGA

Step 1: write sequences along top row and left-most column
Visualization – protein and DNA sequences

Human1: GLKELLHGAV  
Pig:    GLRELLQGAV

Step 1: write sequences along top row and left-most column

Step 2: Identify rows/columns that are a match along the diagonal
Visualization – protein and DNA sequences

Human1  GLKELLHGAV
Pig      GLRELLQGAV

Step 1: write sequences along top row and left-most column

Step 2: Identify rows/columns that are a match along the diagonal
Visualization – protein and DNA sequences

Human1  GLKELLHGAV
Pig      GLRELLQGAV

Step 1: write sequences along top row and left-most column

Step 2: Identify rows/columns that are a match along the diagonal

Step 3: Identify all pair-wise letters that match
Visualization – protein and DNA sequences

Step 1: write sequences along top row and left-most column

Step 2: Identify rows/columns that are a match along the diagonal

Step 3: Identify all pair-wise letters that match
Visualization – protein and DNA sequences

Human1: GLKELLHGAV
Pig: GLRELLQGAV

Step 1: Write sequences along top row and left-most column

Step 2: Identify rows/columns that are a match along the diagonal

Step 3: Identify all pair-wise letters that match
Visualization – protein and DNA sequences

Devise of a dot-matrix(ish) approach for visually inferring the alignment of 3 sequences

In class worksheet exercise

<table>
<thead>
<tr>
<th>Sequence 1</th>
<th>Sequence 2</th>
<th>Sequence 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>SSVT</td>
<td>GSDT</td>
<td>WSWTW</td>
</tr>
</tbody>
</table>

For the below 3 sequences, use a dot-matrix approach to identify similar (based on amino acid) regions. Devise of a “scoring” scheme to identify high match (3 matches) from Medium match (2 matches) from Low match (no match).

<table>
<thead>
<tr>
<th>Sequence 1</th>
<th>SSVT</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sequence 2</td>
<td>GSDT</td>
</tr>
<tr>
<td>Sequence 3</td>
<td>WSWTW</td>
</tr>
</tbody>
</table>
Visualization – protein and DNA sequences

Other variants of matrix-ish approaches
Visualization using R, gnuplot, matlab, etc.