Homework 0

CS1820

February 19, 2017

Out: February 9, 2017
Due: 6:00 PM February 17, 2017

Handin Instructions: See this piazza post.

1 Alignment (30 points)

For all problems, show the dynamic programming table used to construct the alignments you found. You don’t have to show backtracking arrows, but you do need to tell us what the final alignments are. Also, if you find more than one alignment with the optimal score, write down all of them. Given the following scoring scheme:

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>C</th>
<th>G</th>
<th>T</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>1</td>
<td>-1</td>
<td>-1</td>
<td>-1</td>
</tr>
<tr>
<td>C</td>
<td>-1</td>
<td>1</td>
<td>-1</td>
<td>-1</td>
</tr>
<tr>
<td>G</td>
<td>-1</td>
<td>-1</td>
<td>1</td>
<td>-1</td>
</tr>
<tr>
<td>T</td>
<td>-1</td>
<td>-1</td>
<td>-1</td>
<td>1</td>
</tr>
</tbody>
</table>

1. Find the best global alignments for the following sequences using a \((-2)\) gap penalty:

\[
\begin{align*}
GAT \\
GCT
\end{align*}
\]

2. Find the best local alignments for the two sequences from part 1 of this problem using a \((-1)\) gap penalty.

3. Find the best global alignment using a \((-2)\) gap penalty for the following two sequences:

\[
\begin{align*}
ATCGTC \\
GATCGC
\end{align*}
\]

4. Now, using a different scoring scheme:

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>C</th>
<th>G</th>
<th>T</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>C</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>G</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>T</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>

Find the best local alignment for the following two sequences using a \((-2)\) gap penalty:

\[
\begin{align*}
ATCGTC \\
GATCGG
\end{align*}
\]
2 BLAST: Basic Local Alignment Search Tool (8 points)

Please follow the directions carefully so that you get the intended output.

- Click on “Nucleotide Blast” in the “Basic Blast” section.
- Copy the following DNA sequence and paste it into the provided window (Enter Query Sequence).
  AAACCTCCGATGCCTATCAGTCCGCACGTATCCGGCAGCCCTGTCACCGGTCGGGCT TGGCGCTGCTGAGGACACCGCTGAAACCGAGGAGACGGCAAGGACATCGCCGG AGATCCGCGCCTCGACAACGAGAAACCCCTGCTAGCAGAGCCCGCTGAGAACAC CGCAGCGAGATTCCAGCGTGCGGCCAAAATCGGGCTTTTGAGCAGAGTGCTGTGCTG GTGTCTCTCTCTCACTCTGTCTCTTGTTGTTGTTGTTCCGGACTGCGCTCGGTTCTGAGC AGAGGCTACGGCAAGAAGACATCGGAAGAAGCTGACACCTCTCGCTCAAA GCAGTTCTACCTATCTGCGCAGAGAAGACCTTAGGGCCCGCAGACGATACG AAGCCAAGATAACCGCAATTCGAGAGATTTAAAGAAGCTTTACTCCAAATTAC AATCCCGACATTACTTTAAAAAGAGGAGGAGAACACGAGGAGGAGACGGCTCAT GACACAGAGATGCAAAGACAGCTGACACCTCGCTGCCATCTGTGATATGAACCC ACTGGCGCAGGTTAAAGCTGACACCTCGCTGACAGAGGGCTGGGATGAGGACGGTCAC CATTTGAGAATCA

- Use the default settings under Choose Search Set:
  Database: Other, nr (non-redundant nucleotide databases)
  Organism: leave it blank
  Exclude: leave it blank
  Entrez query: leave it blank

- Use the default setting under Program Selection:
  Highly similar sequences (megablast)

- Hit the blue BLAST button

- Scroll down past the graphic summary and look at the ‘Descriptions’ section

- Now, answer the following questions with one or two word answers:
  1. What protein does this gene encode?
  2. What organism does it come from?
3 PDB: Protein Data Bank (12 points)

What is the most abundant protein in the human body? You might think it would be some important enzyme involved in a metabolic pathway, or a regulatory protein involved in daily processes like sleeping or moving, but the most abundant protein in humans is actually a structural one - collagen. Collagen makes up a lot of the connective tissues in our body, including the skin. There are several types of collagen, and we will focus on collagen type I, which has quite an interesting structure: it is made up of two α1 chains and one α2 chain organized in a triple helix. In this exercise, we will practice using PDB (Protein Data Bank) by exploring the α1 chain of type I collagen.

1. Determine through web search what gene is responsible for encoding the collagen type I α1 chain. What is this gene? (Provide a 6 letter answer.)

2. Go to the protein data bank website (http://www.rcsb.org/pdb/home/home.do), and type in the gene name you found from part 1 of this problem into the search box. You'll come up with several hits, but we'll use another type of filter to narrow down what we want to look at. Click on the Refine Search button, and you'll be brought to a new page to enter advanced search queries. Add another search criteria, this time making the query type X-ray Resolution which is under the Methods header. Put the range as between 0.5 and 2.0. You should now come up with 2 hits; take a look at the resolution of each. What is the 4 letter PDB ID (in bold used to label the hit) of the one with the smaller resolution, and what is the resolution of that protein? (Provide a 4 letter answer, and a number along with units.)

3. Click on the 4 letter PDB ID of the hit with the smaller resolution. You'll be brought to an information page. Here, we see that we have a ton of information, most of which doesn't quite tell us about the α1 chain we are interested in. Scroll through the page and find the Macromolecules section. In this section, read through some of the details, but what we want to do is find the link to COL1A1 Gene View. After clicking on this link, we'll finally be on a page that has information solely about the α1 chain we were interested in. What chromosome is this protein encoded on? (Provide a number answer.)

4. Click on the Protein Feature View button. Now, we can finally take a look at the 3D structure of collagen type I and observe the two α1 chains present. Play around with the structure and the visualization tool. An interesting tool available on this page is being able to see similar proteins that are present in other organisms. Look for the dropdown menu labeled by This protein in other organisms (by gene name). The dropdown will currently be set to PO2452 - Homo Sapiens. This number is the UniProt ID of the COL1A1 gene/protein product in humans. UniProt is another protein information database. What is the only other similar protein for which there is a PDB entry? What species is this protein found in? (Provide a UniProt ID and species name.)

You will have noticed that we had to go through multiple filters and links to get to some of the specific information we wanted. This is common with PDB, as it will return all hits that are related to whatever you search. PDB also allows you to download files that contain information about protein structures.

---

4 Multiple Sequence Alignment (10 points)

UPDATES: For problem 4, we are concerned with only globally aligning the $k$ sequences and in 4.2, we are looking for both a naive approach to the problem and then further optimizations that can be made; furthermore, you can assume all $k$ sequences are of length $n$.

In class, we were only concerned with pairwise alignment. The question of how to best align more than two sequences or texts is actually still an open problem. Please be brief when answering the following two questions.

1. How might you go about aligning 3 sequences using an approach very similar to how we did pairwise alignment? Think carefully about how our dynamic programming table might change, and how we’d fill it in. The answer to this doesn’t have to be extensive, but should describe your ideas from a high level, including: what your DP table looks like and how it gets filled in, where the initializations occur, and what the recurrence may look like.

2. What about aligning more than three sequences, let’s say $k$ sequences? Describe how our pairwise method can be adapted to account for $k$ sequences. Be brief here, and don’t worry about capturing all the intricacies; just explain at a high level. Can you think of any sort of optimizations or shortcuts that can be used when comparing $k$ sequences that would allow us to do the alignment faster?