ECS 129: Structural Bioinformatics
Midterm 2
March 2, 2020

Notes:
1) The midterm is open book, open notes.
2) You have 45 minutes, no more: I will strictly enforce this.
3) The midterm is divided into 2 parts, and graded over 90 points.
4) You can answer directly on these sheets (preferred), or on loose paper.
5) Please write your name at the top right of each page you turn in!
6) Please, check your work! If possible, show your work when multiple steps are involved.

Part I (6 questions, each 10 points; total 60 points)
(Most of these questions are multiple choices; in each case, find the most plausible answer)

1) How many possible alignments of length \(M\), with no gaps, can you form when you compare two sequences of length \(N\) and \(M\), with \(N = 2M\)? (gaps at the beginning or end of a sequence do not count)
   A) 2
   B) \(M\)
   C) \(M+1\)
   D) \(M+2\)
   E) \(2M\)

2) The Ramachandran plot:
   A) Compares the conformation of the side-chains of a protein.
   B) Shows the accessibility of all amino acids in a protein
   C) Shows the relationship between the torsion angles \(\phi\) and \(\psi\), for each amino acid in the protein
   D) Shows the torsion angle for the peptide bond, for each amino acid in the protein
   E) Shows the number of hydrogen bonds that stabilize a protein

3) The figure below shows two base pairs of a hybrid DNA-RNA complex. identify it (note that \(dX\) indicates a deoxyribonucleotide, as contained in a DNA molecule, while \(rX\) refers to a ribonucleotide, as found in an RNA molecule). **Hint: the purines are on the DNA strand.**
   A) \(r(5'\)-UC\(-3')\) – \(d(5'\)-GA\(-3')\)
   B) \(r(5'\)-CU\(-3')\) – \(d(5'\)-GA\(-3')\)
   C) \(r(5'\)-TC\(-3')\) – \(d(5'\)-GA\(-3')\)
   D) \(r(5'\)-GA\(-3')\) – \(d(5'\)-UC\(-3')\)
   E) \(r(5'\)-GA\(-3')\) – \(d(5'\)-CU\(-3')\)
4) We want to find the best alignment(s) between the protein sequences AYAVL and YAVL. The scoring scheme $S$ is defined as follows: $S(i,i) = 10$, and $S(i,j) = -2$ otherwise. There is a length-dependent gap penalty of $-2 - 2N$, where $N$ is the length of the gap (gaps at the beginning are considered). The (partial) dynamic programming matrix is shown below. What is the score in the cell identified with an interrogation mark (?)

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>Y</th>
<th>A</th>
<th>V</th>
<th>L</th>
</tr>
</thead>
<tbody>
<tr>
<td>Y</td>
<td>-2</td>
<td>6</td>
<td>-8</td>
<td>-10</td>
<td>-12</td>
</tr>
<tr>
<td>A</td>
<td>6</td>
<td>-4</td>
<td>16</td>
<td>0</td>
<td>-2</td>
</tr>
<tr>
<td>V</td>
<td>-8</td>
<td>4</td>
<td>0</td>
<td>26</td>
<td>?</td>
</tr>
<tr>
<td>L</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

A) 10  
B) 5   
C) 3   
D) 9   
E) -5

5) We want to find the best alignment(s) between the 2 DNA sequences AATGCT and AGTCT. The scoring scheme $S$ is defined as follows: $S(i,i) = 10$, $S(i,j) = 5$ if $i$ and $j$ are both purines, or both pyrimidines, and $S(i,j) = 0$ if $i$ is a purine and $j$ is a pyrimidine, or if $i$ is a pyrimidine and $j$ is a purine. The is a constant gap penalty of 5 (gaps at the beginning are considered, see below). The score $S_{best}$ and the number $N$ of optimal alignments are (show your final dynamic programming matrix for full credit):

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>A</th>
<th>T</th>
<th>G</th>
<th>C</th>
<th>T</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>10</td>
<td>5</td>
<td>-5</td>
<td>0</td>
<td>-5</td>
<td>-5</td>
</tr>
<tr>
<td>G</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>T</td>
<td>-5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>C</td>
<td>-5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>T</td>
<td>-5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

A) $S_{best} = 40$, $N = 2$  
B) $S_{best} = 35$, $N = 2$  
C) $S_{best} = 35$, $N = 1$  
D) $S_{best} = 40$, $N = 1$  
E) $S_{best} = 30$, $N = 1$

6) We want to find the best alignment(s) between the protein sequences WYWAC and WAWC. The scoring scheme $S$ is defined as follows: $S(i,i) = P$, and $S(i,j) = M$ otherwise. There is a constant gap penalty of $G$ (gaps at the beginning are considered). The dynamic programming matrix is shown below. What were the values of $P$, $M$, and $G$?

<table>
<thead>
<tr>
<th></th>
<th>W</th>
<th>Y</th>
<th>W</th>
<th>A</th>
<th>C</th>
</tr>
</thead>
<tbody>
<tr>
<td>W</td>
<td>5</td>
<td>-4</td>
<td>3</td>
<td>-4</td>
<td>-4</td>
</tr>
<tr>
<td>A</td>
<td>-4</td>
<td>3</td>
<td>1</td>
<td>8</td>
<td>1</td>
</tr>
<tr>
<td>W</td>
<td>3</td>
<td>1</td>
<td>8</td>
<td>-1</td>
<td>6</td>
</tr>
<tr>
<td>C</td>
<td>-4</td>
<td>1</td>
<td>-1</td>
<td>6</td>
<td>11</td>
</tr>
</tbody>
</table>

A) $P=5$, $M=0$, $G=0$  
B) $P=5$, $M=-1$, $G=-1$  
C) $P=5$, $M=-2$, $G=-2$  
D) $P=5$, $M=-3$, $G=-1$  
E) $P=5$, $M=-1$, $G=-3$
Part II (three questions, 10 points each: total 30)

Concanavalin A (ConA) is a lectin (carbohydrate-binding protein) originally extracted from the jack bean, *Canavalia ensiformis*. It binds specifically to certain sugars, glycoproteins, and glycolipids. The structure of concanavalin has been determined by X-ray crystallography, and is stored in the PDB. You are interested to know how similar this lectin is from the other lectins that are known, in particular to the lectin from peanut, whose structure is also known. First, you run BLAST, starting from the sequence of ConA. BLAST does find a match with the peanut lectin:

```
>pdb|2PEL|A  Chain A, Peanut Lectin
pdb|2PEL|B  Chain B, Peanut Lectin
pdb|2PEL|C  Chain C, Peanut Lectin
>51 more sequence titles
Length=236
```

Score = 91.3 bits (225), Expect = 3e-22, Method: Compositional matrix adjust.
Identities = 53/116 (46%), Positives = 67/116 (58%), Gaps = 1/116 (1%)

```
Query 1  ADTVAVELDTPYNDGAPSHPHIDIEKSVRSKK?AKWNMQEGKVGPAHILYNSVDEKR  60
            A V VE DTY N++  DP +GID+ SY S ST WN G V +I+Y+S K
Sbjct 114 ACHFVVEFDTYSNSEYNDPP7GDVGDVHSVDVS9TKFPVWSVSQAVKVVTIVDSSSTKT  173
```

Score = 76.3 bits (186), Expect = 1e-16, Method: Compositional matrix adjust.
Identities = 44/106 (42%), Positives = 64/106 (60%), Gaps = 7/106 (7%)

```
Query 124 DALIFHMQFSDKQDKLILGQDGATTGTDGNHELHRVSSNGSPEGSSVGRALFTAPVFHNE  183
            ++ + FN FD+ ++QDD ++GHI+T ++ +SGC L+ PV EX
Sbjct 2  ETYSFVFNSFSSEGPAINFQGDPYVSNGHNQDTNLN------KNVNGRVLVYAMPVRINS  56
```

```
Query 184 SGAVTV-EPATNAFLK-PSKHAPDGIAEFFISNIDSS1PGSTG  227
            S+ V SF ++F ++K D PAGI FP1++ D+ IP+GS G
Sbjct 57  SAGQVNASFLISFSPFMKO1KDTPDAGIIFTFIAFDTQ1IPAGSG  102
```

a) BLAST found two alignments between subsets of the sequences of ConA and the peanut lectin. Are these two alignments significant? Justify your answer
b) Based on these results from BLAST, draw schematically the dotplot between ConA and the peanut lectin. Only show the major correspondences between the two sequences.

c) The two local alignments found by BLAST are 116 residues long and 106 residues long, respectively. Based on the schematic dotplot you have drawn (from question b), explain why BLAST could not have found a single alignment of length at least 222.
Appendix: Nucleotides