ECS 129: Structural Bioinformatics
March 22, 2014

Notes:
1) The final exam is open book, open notes.
2) The final is divided into 2 parts, and graded over 100 point
3) You can answer directly on these sheets (preferred), or on loose paper.
4) Please write your name at the top right of each page you turn in!
5) Please, check your work! If possible, show your work when multiple steps are involved.

Part I (15 questions, each 4 points; total 60 points)
(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 > M?
   A) 1
   B) N-M
   C) N-M+1
   D) M
   E) N

   There are N-M+1 positions in sequence 1 for the first letter of the sequence of length M that leads to an alignment of length M.

2) You have been studying a large family of protein sequences that are highly homologous to each other. You would like to generate a substitution matrix that is specific to this family; this substitution matrix should prove useful when searching for other members of this family in large protein databases. To build this matrix, you generate first a multiple sequence alignment (MSA) of all the sequences in the family. From this MSA it is possible to build a table of substitution, based on direct count. Here is the table you observed, limited to the aromatic amino acids:

   |   | H | Y | W |
---|---|---|---|
F | 5 | 2 | 100| 4 |
H | 2 | 10| 40 | 3 |
Y | 100| 40| 20 | 5 |
W | 4 | 3 | 5 | 1 |

If this table is called T, T(i,j) represents the number of times that amino acid i is replaced with amino acid j (for example there are 100 mutations F->Y). You have doubts however and you think that you have made some mistakes. What is the most likely hint that makes you think that this table is wrong?

   A) W is the most abundant amino acid type in proteins; however your table has only a small number of them
   B) The mutation F->Y never occurs in proteins
   C) The values on the diagonal are too small compared to the values off diagonal
D) The sums of the numbers on each line should be all equal
E) All of the above

3) The figure below shows a small peptide of six amino acids; give its sequence: (hint: there is one charged amino acid at physiological pH – from pH 5.5 to pH 8.0; hydrogens are not shown)

4) Cytochrome P450 enzymes form a super-family of haem-containing oxygenases that are found in all kingdoms of life. These proteins have very similar structures but show extraordinary diversity in their reaction chemistry. Let us consider these three examples: (A) the human CYP46A1, an enzyme that controls cholesterol turnover in the brain, (B), a human prostacyclin synthase (prostacyclin is a small lipid that inhibits platelet aggregation), and (C), Xpla, a cytochrome P450 from rhodococcus (aerobic, gram-positive bacterium) that has been found to break down explosive pollutants. (A), (B) and (C) are homologous proteins; what else can you say?

A) (A), (B) and (C) are orthologous
B) (A), (B) and (C) are paralogous
C) (A), (B) and (C) are analogous
D) (A) and (B) are paralogous, while (A) and (C) are orthologous
E) (A) and (B) are orthologous, while (A) and (C) are paralogous

5) Given two DNA sequences that are each other’s inverse (for example GATCAT and TACTAG), what does their dotplot look like?

6) Peptide Nucleic Acids, or PNAs, are synthetic oligomers with a protein backbone on which bases (purines and pyrimidines) are linked every second N. Unlike DNA, PNAs do not contain
sugars or phosphate groups. PNAs are represented as proteins, from Nter to Cter. Find the “sequence” from Nter to Cter of the PNA shown below:

![Image of PNA structure]

A) Nter-TACGTA-Cter  
B) Nter-CGTACG-Cter  
C) Nter-CATGCA-Cter  
D) Nter-TGCATG-Cter  

7) How many DNA coding sequences (where a coding sequence includes the START and STOP codon, but no introns) could lead to the following protein sequence:
Met- Lys-Leu-Trp-Ser-Phe-Trp-Val, assuming the standard genetic code?
A) 1  
B) 576  
C) 1152  
D) 1728  
E) 4096  

Count number of codons for each amino acid (3 for stop codons):
1 (Met) x 2 (Lys) x 6 (Leu) x 1 (Trp) x 6 (Ser) x 2 (Phe) x 1 (Trp) x 4 (Val) x 3 (STOP)= 1728

8) You have designed E-Coli such that it can react to light. In the presence of light it generates a white dot, while in the absence of light it generates a black dot. You want to use a bio-film of covered with E-Coli as a synthetic camera. Assuming that the bacteria cover uniformly (with no overlap) your bio-film, and that each bacterium is circular with a radius of 0.5 µm, and assuming you want to generate an image with 400 Megapixels (1 Megapixel = 10^6 pixels) what would be a possible size for your bio-film?

   a) 1 cm x 1 cm  
   b) 1 cm x 2 cm  
   c) 1 cm x 3 cm  
   d) 1 cm x 4 cm

Number of bacteria = Number of pixels = 4x10^8

Surface: Number of bacteria x 1 µm x 1 µm (the bacteria can’t overlap) = 4x10^8 x 10^-6 x 10^-6 m^2 = 4 cm^2
9) We want to find the best alignment(s) between the 2 DNA sequences TATATGCA and ATATC. 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 $S(i,j) = 0$ otherwise. There is a constant gap penalty of 1. The score $S_{best}$ and the number $N$ of optimal alignments are (show your final dynamic programming matrix and alignment(s) for full credit).

<table>
<thead>
<tr>
<th></th>
<th>T</th>
<th>A</th>
<th>T</th>
<th>A</th>
<th>T</th>
<th>G</th>
<th>C</th>
<th>A</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>0</td>
<td>9</td>
<td>-1</td>
<td>9</td>
<td>-1</td>
<td>4</td>
<td>-1</td>
<td>9</td>
</tr>
<tr>
<td>T</td>
<td>9</td>
<td>0</td>
<td>19</td>
<td>8</td>
<td>19</td>
<td>8</td>
<td>13</td>
<td>8</td>
</tr>
<tr>
<td>A</td>
<td>-1</td>
<td>19</td>
<td>8</td>
<td>29</td>
<td>18</td>
<td>24</td>
<td>18</td>
<td>28</td>
</tr>
<tr>
<td>T</td>
<td>9</td>
<td>8</td>
<td>29</td>
<td>18</td>
<td>39</td>
<td>28</td>
<td>33</td>
<td>28</td>
</tr>
<tr>
<td>C</td>
<td>4</td>
<td>9</td>
<td>23</td>
<td>29</td>
<td>33</td>
<td>39</td>
<td>48</td>
<td>38</td>
</tr>
</tbody>
</table>

A) $S_{best} = 50$, $N = 1$
B) $S_{best} = 50$, $N = 2$
C) $S_{best} = 48$, $N = 1$
D) $S_{best} = 48$, $N = 2$

10) The dotplot shown below compares the DNA sequence of the actin muscle gene from Pisaster ochraceus (horizontal) with the RNA corresponding to the same gene (vertical). The six regions of high similarity that shows as black lines correspond to:

A) Introns
B) Repeats
C) Inverted repeats
D) Exons
E) All of the above

Conserved regions between RNA and DNA corresponds to coding regions, hence exons.
11) We want to find the best alignment(s) between the protein sequences WWYCTY and WCYTY. The scoring scheme $S$ is defined as follows: $S(i,i) = 10$, $S(i,j) = 5$ if $i$ and $j$ are both aromatic amino acids, and $S(i,j) = 0$ otherwise. There is a constant gap penalty of 5 (gaps at the beginning are not considered, see below). The score $S_{best}$ and the number $N$ of optimal alignments are (show your final dynamic programming matrix and optimal alignment(s) for full credit):

<table>
<thead>
<tr>
<th></th>
<th>W</th>
<th>W</th>
<th>Y</th>
<th>C</th>
<th>T</th>
<th>Y</th>
</tr>
</thead>
<tbody>
<tr>
<td>W</td>
<td></td>
<td>10</td>
<td>5</td>
<td>0</td>
<td>-5</td>
<td>-5</td>
</tr>
<tr>
<td>C</td>
<td>-5</td>
<td></td>
<td>10</td>
<td>5</td>
<td>15</td>
<td>5</td>
</tr>
<tr>
<td>Y</td>
<td>0</td>
<td>10</td>
<td>20</td>
<td>5</td>
<td>15</td>
<td>20</td>
</tr>
<tr>
<td>T</td>
<td>-5</td>
<td>5</td>
<td>10</td>
<td>20</td>
<td></td>
<td>25</td>
</tr>
<tr>
<td>Y</td>
<td>0</td>
<td>10</td>
<td>15</td>
<td>15</td>
<td>20</td>
<td>35</td>
</tr>
</tbody>
</table>

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

*There is only one optimal alignment (traceback shown in bold):*

```
WWYCTY
WCYTY
```

12) Assume that the whole DNA corresponding to the human genome can be represented as a ribbon of length $L=1$ meter and width $W = 2$nm. Assume also that this DNA contains $3 \times 10^9$ bases. If we could create a material that would allow the same storage as this DNA, how many mm$^2$ (millimeter squared) of such material would you need to store the whole Encyclopedia Britannica, assuming that it contains 30 billion letters (i.e. $3 \times 10^{10}$ letters).

A) 2
B) 0.02
C) 200,000
D) 0.2
E) None of the above

Density of letter = (Number of letter) / (area of DNA ribbon) = $3 \times 10^9 / (2 \times 10^9) = 1.5$ letter / nm$^2$

Number of letters in encyclopedia = $3 \times 10^{10}$
Surface needed : $3/1.5 \times 10^{10} \text{ nm}^2 = 2 \times 10^{10} \text{ nm}^2 = 0.02 \text{ mm}^2$

13) Given that a double stranded DNA molecule contains 21% of Adenosine, find the corresponding percentage of Guanosine it contains

A) 21%
B) 42%
C) 29%
D) 25%
E) Not enough information available

The DNA is double stranded: there are as many As as Ts, and as many Gs and Cs. Since Adenosine represents 21% of the content, T represents another 21%; therefore G+C represents 58%, and since G and C are in equal amount, Guanosine represents 29% of the content.

14) The codon for Tyrosine is UAU. How many different amino acids (not including tyrosine) could possibly result from substitutions of the first base, the second base, or both (the third base will always be U)?

A) 13  
B) 14  
C) 15  
D) 7  
E) None of the above

There are 6 options for single-base substitutions, depending on which base is mutated:
- First base. The mutated codons are AAU, GAU, and CAU, which code for Asn, Asp, and His, respectively.
- Second base. The mutated codons are UUU, UGU, and UCU, which code for Phe, Cys, and Ser, respectively.

There are 9 additional double substitutions: AUU, AGU, ACU, GUU, GGU, GCU, CUU, CGU, CCU that code for Ile, Ser, Thr, Val, Gly, Ala, Leu, Arg, and Pro, respectively.

There are therefore 15 possible amino acids, but Ser occurs twice: therefore 14 possible amino acids.

15) Let us consider a population of 1 million bacteria. We suppose the genome of one bacterium is 3x10^6 base pair long and that we have observed a total number of 9x10^7 point mutations over 100,000 generations. What is your estimate of the intrinsic mutation rate per base pair?

A) 3 x 10^{-9}  
B) 3 x 10^{-10}  
C) 3 x 10^{-8}  
D) 9 x 10^{-7}  
E) None of the above

\[ \text{N}_{\text{mutation}} = (\text{Population size}) \times (\text{Number of base pair}) \times (# \text{ mutation / base pair}) \times (# \text{ of generation}) \]
\[ 9 \times 10^7 = 1 \times 10^6 \times 3 \times 10^6 \times p \times 100000 \]
therefore \[ p = 3 \times 10^{-10} \]
Part II (2 problems, each 20 points; total 40 points)

Problem 1:

   a) BLAST found two alignments between subsets of the sequences of ConA and the peanut lectin. Are these two alignments significant? Justify your answer (2 points)

   The two local alignments have E-values of $3\times10^{-22}$ and $1\times10^{-16}$, respectively: as such, they are highly significant (the corresponding P-values, i.e. the probabilities that these alignments are random, are $3\times10^{-22}$ and $1\times10^{-16}$, respectively).

   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 (3 points)

   ![Dotplot between ConA and peanut lectin]

   c) The two local alignments found by BLAST are 116 residues long and 106 residues long, respectively. Based on the specificity of these two alignments and the schematic dotplot you have drawn (from question b), explain why BLAST could not have found a single alignment of length at least 222 (3 points).

   The two alignments are not sequential along the ConA sequence from Nter to Cter: in fact, ConA contains two domains that have been swapped in peanut lectin (a circular permutation). As BLAST (just like most sequence comparison techniques) works with the sequence given from Nter to Cter, it is not designed to detect domain swap.

   d) From these results, do you expect the structures of ConA and peanut lectin to be similar? Justify your answer (2 points).

   As the two alignments are highly significant, it is expected that the two domains of ConA (from residue 1 to 115 and from residue 124 to 227, respectively) be structurally conserved in peanut lectin. We do not know however if the 3D arrangement of these two domains (i.e. how these two domains are packed in the full protein) is conserved.
Problem 2:

- Question 1:

1) The following eukaryotic DNA sequence was given to you:

5’-CCCTTAATGCGTATCGCTCACGAGATGTTGGGCGGCTAA-3’

You are told that this sequence, or its complementary, codes for one gene.

Find the longest “gene” corresponding to this DNA sequence; remember that there are 6 possibilities, i.e. 3 possible reading frames for one strand and 3 possible reading frames for its complementary.

Transcribe this gene into an RNA sequence and then translate it into a protein sequence.

We don’t know if the sequence given corresponds to the coding strand, so we need to check both this sequence S, and its complementary C:

5’-TTAGCGCCCAACATCTCGTGAGCGATACGCATTAAGGG-3’

The complementary strand C does not contain any ATG (Start codon)

The initial sequence S contains two ATG in phase, and one TAA (stop codon), in phase with both ATG. Consequently, the longest ORF goes from the first ATG to TAA:

5’ ATG CGT ATC GCT CAC GAG ATG TTG GGC GGC TAA-3’

The corresponding RNA sequence is:

5’ AUG CGU AUC GCU CAC GAG AUG UUG GGC GGC UAA-3’

The protein sequence is obtained directly using the genetic code:

Nter – Met Arg Ile Ala His Glu Met Leu Gly Gly – Cter

Question 2:

Predict the secondary structure of this “protein” using the Chou and Fassman method, with the propensities given in Appendix D.

To predict the secondary structure of this peptide, we use the Chou and Fassman propensities:

<table>
<thead>
<tr>
<th></th>
<th>M</th>
<th>R</th>
<th>I</th>
<th>A</th>
<th>H</th>
<th>E</th>
<th>M</th>
<th>L</th>
<th>G</th>
<th>G</th>
</tr>
</thead>
<tbody>
<tr>
<td>P(a)</td>
<td>1.47</td>
<td>0.96</td>
<td>0.97</td>
<td>1.29</td>
<td>1.22</td>
<td>1.44</td>
<td>1.47</td>
<td>1.30</td>
<td>0.56</td>
<td>0.56</td>
</tr>
<tr>
<td>P(b)</td>
<td>0.97</td>
<td>0.99</td>
<td>1.45</td>
<td>0.90</td>
<td>1.08</td>
<td>0.75</td>
<td>0.97</td>
<td>1.02</td>
<td>0.92</td>
<td>0.92</td>
</tr>
</tbody>
</table>
Helix:

- nucleation sequence: IAHEML
- extension: add R and M on Nter side, and first G on Cter side
- Computer average over 9 first residues: 1.18 > 1.0

9 first residues predicted to be helical.

Strand:

- no nucleation site
The prediction is therefore: HHHHHHHHHO

Note: if you use the nucleation sequence: HEMLGG to predict the helical content, you find that the whole peptide is helical! This is a problem of the Chou and Fassman scheme…I counted both answers as correct.
Appendix A: Amino Acids

Hydrophobic Amino Acids

GLY (G)
ALA (A)
Val (V)
Leu (L)
Ile (I)
Pro (P)
Phe (F)
Met (M)

Polar Amino Acid

Ser (S)
Thr (T)
Tyr (Y)
His (H)
Trp (W)
Asn (N)
Gln (Q)
Polar Amino Acids

Appendix B: Nucleotides

Thymine (T)

Adenine (A)

Uracyl (U)

Cytosine (C)

Guanine (G)
Appendix C: Genetic Code

<table>
<thead>
<tr>
<th></th>
<th>U</th>
<th>C</th>
<th>A</th>
<th>G</th>
</tr>
</thead>
<tbody>
<tr>
<td>U</td>
<td>Phe</td>
<td>Ser</td>
<td>Tyr</td>
<td>Cys</td>
</tr>
<tr>
<td></td>
<td>Phe</td>
<td>Ser</td>
<td>Tyr</td>
<td>Cys</td>
</tr>
<tr>
<td></td>
<td>Leu</td>
<td>Ser</td>
<td>STOP</td>
<td>STOP</td>
</tr>
<tr>
<td></td>
<td>Leu</td>
<td>Ser</td>
<td>STOP</td>
<td>Trp</td>
</tr>
<tr>
<td></td>
<td>Leu</td>
<td>Pro</td>
<td>His</td>
<td>Arg</td>
</tr>
<tr>
<td></td>
<td>Leu</td>
<td>Pro</td>
<td>His</td>
<td>Arg</td>
</tr>
<tr>
<td></td>
<td>Leu</td>
<td>Pro</td>
<td>Gln</td>
<td>Arg</td>
</tr>
<tr>
<td></td>
<td>Leu</td>
<td>Pro</td>
<td>Gln</td>
<td>Arg</td>
</tr>
</tbody>
</table>

Appendix D: Chou and Fassman Propensities

<table>
<thead>
<tr>
<th>Amino Acid</th>
<th>Helix</th>
<th>Strand</th>
<th>Turn</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ala</td>
<td>1.29</td>
<td>0.90</td>
<td>0.78</td>
</tr>
<tr>
<td>Cys</td>
<td>1.11</td>
<td>0.74</td>
<td>0.80</td>
</tr>
<tr>
<td>Leu</td>
<td>1.30</td>
<td>1.02</td>
<td>0.59</td>
</tr>
<tr>
<td>Met</td>
<td>1.47</td>
<td>0.97</td>
<td>0.39</td>
</tr>
<tr>
<td>Glu</td>
<td>1.44</td>
<td>0.75</td>
<td>1.00</td>
</tr>
<tr>
<td>Gln</td>
<td>1.27</td>
<td>0.80</td>
<td>0.97</td>
</tr>
<tr>
<td>His</td>
<td>1.22</td>
<td>1.08</td>
<td>0.69</td>
</tr>
<tr>
<td>Lys</td>
<td>1.23</td>
<td>0.77</td>
<td>0.96</td>
</tr>
<tr>
<td>Val</td>
<td>0.91</td>
<td>1.49</td>
<td>0.47</td>
</tr>
<tr>
<td>Ile</td>
<td>0.97</td>
<td>1.45</td>
<td>0.51</td>
</tr>
<tr>
<td>Phe</td>
<td>1.07</td>
<td>1.32</td>
<td>0.58</td>
</tr>
<tr>
<td>Tyr</td>
<td>0.72</td>
<td>1.25</td>
<td>1.05</td>
</tr>
<tr>
<td>Trp</td>
<td>0.99</td>
<td>1.14</td>
<td>0.75</td>
</tr>
<tr>
<td>Thr</td>
<td>0.82</td>
<td>1.21</td>
<td>1.03</td>
</tr>
<tr>
<td>Gly</td>
<td>0.56</td>
<td>0.92</td>
<td>1.64</td>
</tr>
<tr>
<td>Ser</td>
<td>0.82</td>
<td>0.95</td>
<td>1.33</td>
</tr>
<tr>
<td>Asp</td>
<td>1.04</td>
<td>0.72</td>
<td>1.41</td>
</tr>
<tr>
<td>Asn</td>
<td>0.90</td>
<td>0.76</td>
<td>1.23</td>
</tr>
<tr>
<td>Pro</td>
<td>0.52</td>
<td>0.64</td>
<td>1.91</td>
</tr>
<tr>
<td>Arg</td>
<td>0.96</td>
<td>0.99</td>
<td>0.88</td>
</tr>
</tbody>
</table>