ECS 129: Structural Bioinformatics
March 15, 2016

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 least on the front page!
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) Two homologous genes:
   A) Would be expected to have very similar sequences in related organisms
   B) Would be expected to be more similar in distantly related organisms than in organisms that are closely related
   C) May have become similar to each other by random mutations
   D) Cannot be found on the same genome
   E) All of these

   *Homologous means the two sequences are related, often very similar.*

2) In the dynamic programming matrix below, what is the score in the cell identified with an interrogation mark (?). Assume that the score for a perfect match is set to 10, the score of a mismatch is set to 0, and gap penalties are ignored.

   |   | A | T | W | C | Y | T |
---|---|---|---|---|---|---|
| A | 10 | 0 | 0 | 0 | 0 | 0 |
| T | 0  | 20| 10| 10| 10| 20|
| C | 0  | 10| 20| **30**| 20| 20|

   A) 20
   B) 10
   C) 30
   D) 40
   E) 0

3) The figure below shows a non-standard nucleotide base pair; 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).

   A) dG-dC
   B) rG-rC
   C) dG-rC
   D) rG-dC
   E) rC-dG
The nucleotide on the left is a ribonucleotide (as the C2’ carries a O); the nucleotide on the right is a deoxy-ribonucleotide (no O on C2’); the bases are G on the left, C on the right.

4) 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)

5) Given the DNA sequence S = 5’-GAATTC-3’, how does the dotplot between S and its complementary, cS, look like?

S = 5’-GAATTC-3’ and cS = 5’-GAATTC-3’ (remember that if nothing is said, sequences are always assumed to be 5’ to 3’); the two sequences are the same and therefore the corresponding dot plot is A.

6) The figure below shows a small fragment of a protein. From this figure, is it possible to define which extremity is the N-terminal, and which extremity is the C-terminal?

A) Yes: 1 is Nter, 2 is Cter
B) No: there is not enough information
C) Yes, 1 is Cter and 2 is Nter
D) No: Nter and Cter are only defined for nucleic acids
E) No: we would need to know the sequence of this protein fragment
Based on the arrows representing the strands, Nter is at 2, and 1 is Cter.

7) The so-called Rosetta stone for predicting protein-protein interactions is:
   A) Gene fusion
   B) Gene co-expression
   C) Presence of the name of the two proteins concerned in the same scientific paper
   D) A very old stone recently found in Gizeh, Egypt, next to the Sphinx, that describes the code
      for protein-protein interactions in three scripts: hieroglyphic, demotic and Greek
   E) A free software with high success rate

   *Gene fusion is really the so-called Rosetta stone for protein-protein interactions.*

8) Which combination of program / substitution matrix will most likely give you the best
    alignment between two sequences that are highly similar?
   A) BLAST / Blosum45
   B) Dynamic programming / Blosum45
   C) BLAST / Blosum90
   D) Dynamic programming / Blosum90
   E) BLAST / Blosum10

   *BLAST is only heuristic; it is best to use dynamic programming to get a good alignment. As the
   two sequences are highly similar, it is best to use a BLOSUM matrix computed from sequences
   that are very similar, such as BLOSUM90, computed with sequences that have at least 90% sequence identity.*

9) How many possible alignments, with no internal gaps, can you form when you compare a
    sequence of length 4 with a sequence of length 8? (Note that an alignment must have at least one
    letter match between the 2 sequences)
   A) 4
   B) 8
   C) 9
   D) 10
   E) 11

   *The last letter of the sequence of length 4 can face any of the 8 letters of the second sequence,
   but can also be hanging after the last letter, up to 3 letters away; therefore the total number is 8
   + 3 = 11*

10) Only one of these techniques directly studies the behavior of a molecule as a function of
    time:

   A) Molecule dynamics
   B) Monte Carlo sampling techniques
   C) Molecular mechanics
   D) Energy minimization
   E) Simulated annealing techniques
Monte Carlo explores in conformational space; molecular mechanics == energy minimization, no time involved; simulated annealing is just a better sampling technique.

11) We want to find the best alignment(s) between the DNA sequences AGTATCT and AGATGC. The scoring scheme S is defined as follows: S(i,j) = 1 if i = j, and S(i,j) = 0 otherwise. There is a constant gap penalty of -1 (penalty for the first position counts; see table below). The score S_{best} and the number N of optimal alignments are (show your final dynamic programming matrix and the best possible alignment(s) for full credit):

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>G</th>
<th>T</th>
<th>A</th>
<th>T</th>
<th>C</th>
<th>T</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>1</td>
<td>-1</td>
<td>-1</td>
<td>0</td>
<td>-1</td>
<td>-1</td>
<td>-1</td>
</tr>
<tr>
<td>G</td>
<td>-1</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>A</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>T</td>
<td>-1</td>
<td>0</td>
<td>2</td>
<td>2</td>
<td>3</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>G</td>
<td>-1</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>C</td>
<td>-1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>3</td>
</tr>
</tbody>
</table>

A) S_{best} = 3, N = 2
B) S_{best} = 3, N = 1
C) S_{best} = 4, N = 1
D) S_{best} = 3, N = 3
E) S_{best} = 4; N = 3

12) A protein sequence contains one ASP residue. You want to create a new protein sequence, with this ASP being replaced with a TYR. To do this, you first generate the cDNA corresponding to the original protein (with your own choice for the codons you use), then mutate this cDNA to get the sequence corresponding to the new protein. What is the minimum number of mutations needed?

A) 1
B) 2
C) 3
D) 0
E) None of the above

*ASP can be represented with the codon GAU; mutating G with U, you get the codon UAU which codes for TYR.*

13) The Ramachandran plot of the protein structure 1axc in the PDB databank is given on the right. Which of the model of protein structures given below is most likely the corresponding structure:
The Ramachandran plot shows as many residues in helical structures than in strand conformations. Only structure C corresponds.

14) A single stranded DNA contains 15% Adenine, as many Guanines as Cytosines, and 40% of purines. What is the amount (in percent) of Thymine:

A) 25%
B) 15%
C) 35%
D) 40%
E) Not enough information available

There are 40% of purines and 15% Adenine, therefore there are 25% of Guanine. Since there are as many Guanine as Cytosine, there are 25% Cytosine. Finally, there are 35% of Thymine.

15) The protein sequence alignment shown below has a total score of 28. Knowing that the score for an exact match is 5 and the score for a mismatch is -4, what is the score used for the (constant, i.e. independent of length) gap penalty:

GCTGGAAG--GCA--T
GC----AGAGCCT
A) -1
B) -2
C) -3
D) -4
E) Undefined (any value would give the same total score)

Total score = 28 = 5*8 +3*x -> x = (28-40)/3 = -4 (it was important to notice that the cost of a gap is independent of its length, as said explicitly in the text of the question; there are therefore 3 gaps to consider, it does not matter that there are 6 residues total in those gaps).
16) Docking is the process of predicting the conformation of the complex formed by a receptor and a ligand. Which of these four statements about docking is most likely to be true?

A) Rigid, bound docking is the most difficult situation for predicting the conformation of the complex
B) We only need the conformation of the receptor to perform docking
C) The lock-and-key concept relates to rigid docking
D) Docking can be solved with a simple energy minimization.

Lock-and-key assumes that the two partners are rigid… which is the underlying assumption for rigid docking.

17) Dynamic programming, popular for sequence alignment, can also be used for spell checking. Assuming that a match is worth 10, a mismatch is worth 5, and a gap “costs” -5, which of these four words is closest to the word “graffe” typed by a user? Write the score of the optimal alignment next to each word (gaps at the start or at the end do not count).

A) gaff best score: 40-5 = 35
B) graft best score: 40+5 = 45
C) grail best score: 30+5+5 = 40
D) giraffe best score: 60-5 = 55

18) Let us consider the Luria and Delbruck experiment. The distribution of the number of mutations that occur during the growth of parallel cultures has a Poisson distribution. If there are no mutants, there were no mutations, and so the mean number of mutations \( m \) that occurs during the growth of a culture can be calculated from \( p_0 \), the proportion of cultures with no mutants: \( m = -\log(p_0) \). Let us consider a bacterium B that is sensitive to a bacteriophage T, unless it carries a mutation M. 50 cultures of the bacterium, each with approximately \( 3 \times 10^7 \) bacteria, are subjected to the bacteriophage; 40 of those cultures show no resistance, i.e. none of their bacteria carried the mutation. Estimate the mutation rate \( \mu \) per bacterium B:

A) \( 7.4 \times 10^{-9} \)
B) \( 2.9 \times 10^6 \)
C) 0.097
D) \( 1 \times 10^{-9} \)
E) Not enough information available

\( m, \) number of mutations per culture = \(-\log(40/50) = 0.22 \)

\( \mu = m/(3\times10^7) = 7.4 \times 10^{-9} \)

19) You want to design a small peptide that can interact with the TATA box of a specific gene (the TATA box is a small DNA sequence upstream from the gene that serves as transcription initiator). Your constraints are: the peptide should contain a strand (at least predicted to be mostly in extended conformation, based on Chou and Fassman, see appendix D), and it should contain 12 residues. Which of the following peptide would be a good candidate?

A) MPGCLPQALGLP
B) MPGLEWQLPGLP
C) MLGYTWTTVSVT
D) MVTTVWYVTGT

A and B are unlikely due to the presence of many prolines; D is only 11 residue long.

20) The cDNA corresponding to a small peptide is ATGTATGATCAATGCAGCGGGCCTTATAG. The corresponding amino acid sequence is Met-Tyr-Asp-Glu-Cys-Ser-Gly-Pro-Leu-Stop. A mutation occurs at the DNA level, with the C at position 15 being substituted with T. What effect do you think this mutation might have on the expression of this gene?

A) It introduces a stop codon and the peptide will be shorter
B) The Cys in position 5 of the protein sequence will be replaced with Trp
C) The Start and Stop codons won’t be in phase anymore and the gene won’t be expressed
D) This is a silent mutation as it will have no impact on the protein sequence

The codon TGC is mutated to TGT... both code for Cysteine; the mutation is silent.

Part II (2 problems; total 40 points)

Problem 1 (4 questions, each 8 points)

1) The following eukaryotic DNA sequence was given to you:
   5'-TAATGACCTTAGAAGAGGTCTCGCGAAACAC-3'

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

Find the longest “gene”, or open reading frame (ORF) 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 ORF into an RNA 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’-CCTTAGTGGTTTCGCGAGACCCTCTTTCTAAGGCCATTATA-3’

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

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

   5’ ATG GCC TTA GAA GAG GGT CTC GCG AAA CAC TAA-3’

The corresponding RNA sequence is:
2) As this is a eukaryotic sequence, it may contain an intron. For simplicity, we will assume that introns always start with GU and end with CA. Identify all possible introns, and explain why their removal would result in the loss of the gene.

There is one GU and one CA in the RNA sequence:

5’ AUG GCC UUA GAA GAG GGU CUC GCG AAA CAC UAA-3’

If we remove the corresponding intron GU CUC GCG AAA CAC, we would get the new RNA sequence:

5’ AUG GCC UUA GAA GAG GC UAA-3’

in which the start and stop codon would not be in phase anymore; the gene would be lost.

3) Based on question 2 just above, we know that the RNA is not spliced. Find the sequence of the “protein” it encodes.

The mRNA is:

5’ AUG GCC UUA GAA GAG GGU CUC GCG AAA CAC UAA-3’

The protein sequence is obtained directly using the genetic code:

Nter – Met Ala Leu Glu Glu Gly Leu Ala Lys His – Cter

Or, in one-letter code:

Nter- MALEEGLAKH-Cter

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

We start by writing the propensities:

<table>
<thead>
<tr>
<th></th>
<th>M</th>
<th>A</th>
<th>L</th>
<th>E</th>
<th>E</th>
<th>G</th>
<th>L</th>
<th>A</th>
<th>K</th>
<th>H</th>
</tr>
</thead>
<tbody>
<tr>
<td>P(helix)</td>
<td>1.47</td>
<td>1.29</td>
<td>1.30</td>
<td>1.44</td>
<td>1.44</td>
<td>0.56</td>
<td>1.30</td>
<td>1.29</td>
<td>1.23</td>
<td>1.22</td>
</tr>
<tr>
<td>P(strand)</td>
<td>0.97</td>
<td>0.9</td>
<td>1.02</td>
<td>0.75</td>
<td>0.75</td>
<td>0.92</td>
<td>1.02</td>
<td>0.9</td>
<td>0.77</td>
<td>1.08</td>
</tr>
</tbody>
</table>
There are no initiation sites for strands. However, there are multiple possible initiation sites for helices. We can pick the Nter of the sequence: MALEEG. We can prolong:

- EEGL: sum $P(\alpha) = 1.44 + 1.44 + 1.30 = 4.74 > 4$
- EGLA: sum $P(\alpha) = 1.44 + 0.56 + 1.30 + 1.29 = 4.59 > 4$
- GLAK: sum $P(\alpha) = 0.56 + 1.30 + 1.29 + 1.23 = 4.38 > 4$
- LAKH: sum $P(\alpha) = 1.30 + 1.29 + 1.23 + 1.22 = 5.04 > 4$

Finally, we compute the average $P(\alpha)$ over the whole peptide:

Sum = $1.47 + 1.29 + 1.30 + 1.44 + 0.56 + 1.30 + 1.29 + 1.23 + 1.22 = 12.54$

Average = $12.54 / 10 = 1.254 > 1$

The whole peptide is predicted to be helical.

**Problem 2 (8 points)**

You have isolated an important gene that regulates the size of a newly found frog from the island of Borneo. You have also been able to find the sequence of the protein encoded by this gene. You suspect that sequences similar to this sequence can be found in other organisms, but with circular permutation:

In a circular permutation, N amino acids (N can take any value between 1 and M-1, where M is the total length of the protein) at the end of the original sequence will appear at the beginning of the permuted sequence (i.e. before the remaining M-N amino acids).

Propose an efficient strategy for detecting all possible permuted sequences of your frog sequence in a large database of protein sequences.

*The most efficient strategy is to generate a pseudo sequence in which the sequence of your frog protein is repeated twice. This pseudo sequence will look like (following the drawing of the question):*
If you search a protein sequence database with this sequence, you will detect all possible permutations!
Appendix A: Amino Acids

Hydrophobic Amino Acids

- Gly (G)
- Ala (A)
- Val (V)
- Ile (I)
- Leu (L)
- Met (M)

Polar Amino Acid

- Ser (S)
- Thr (T)
- Tyr (Y)
- His (H)
- Trp (W)
- Asn (N)
- Gln (Q)
Appendix B: Nucleotides

Thymine (T)  Adenine (A)  Guanine (G)  Uracyl (U)
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>C</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>
<tr>
<td>A</td>
<td>Ile</td>
<td>Thr</td>
<td>Asn</td>
<td>Ser</td>
</tr>
<tr>
<td></td>
<td>Ile</td>
<td>Thr</td>
<td>Asn</td>
<td>Ser</td>
</tr>
<tr>
<td></td>
<td>Ile</td>
<td>Thr</td>
<td>Lys</td>
<td>Arg</td>
</tr>
<tr>
<td></td>
<td>Met/Start</td>
<td>Thr</td>
<td>Lys</td>
<td>Arg</td>
</tr>
<tr>
<td>G</td>
<td>Val</td>
<td>Ala</td>
<td>Asp</td>
<td>Gly</td>
</tr>
<tr>
<td></td>
<td>Val</td>
<td>Ala</td>
<td>Asp</td>
<td>Gly</td>
</tr>
<tr>
<td></td>
<td>Val</td>
<td>Ala</td>
<td>Glu</td>
<td>Gly</td>
</tr>
<tr>
<td></td>
<td>Val</td>
<td>Ala</td>
<td>Glu</td>
<td>Gly</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>