B363: Bioinformatics algorithms

HW2 (Due: **Oct. 10 Monday 5pm**)

http://darwin.informatics.indiana.edu/col/courses/B363-16

(You do not need to write computer programs for the following questions.)

1. (10 pts) Reconstruct the DNA sequence spelled by the following Eulerian path of (2, 1)-mers: (AG|AG) → (GC|GC) → (CA|CT) → (AG|TG) → (GC|GC) → (CT|CT) → (TG|TG) → (GC|GC) → (CT|CA). Note: (2, 1)-mer is a pair 2-mer with fixed distance of 1.

2. (10 pts) To find a Eulerian cycle in a balanced, strongly connected graph, Dr. Smart suggested to start from a vertex with high indegree (and high outdegree). What is his rationale? Assuming you start from the vertex with the highest indegree in the graph, is it always true that the first cycle you get is a Eulerian cycle? Justify your answer.

3. (10 pts) A *bubble* in a *de Bruijn* graph consists of two *parallel* paths connecting from the same vertex A to the same vertex B. 1) Devise an algorithm for detecting bubbles in a given de Bruijn Graph. 2) After a bubble is detected, you must decide which of the two paths in the bubble to remove. How should you make this decision? Explain your answer.

4. (10 pts) Mr. Study devises the following algorithm to solve the spectrum convolution problem (page 203 of the textbook). Is the algorithm correct? What is the run time of the algorithm? Can you revise the algorithm to make its run time only dependent on the size of *spectrum* and independent on *MaxMass*?

```
Input: A collection of integers Spectrum
NovelSpectrumConvolution(Spectrum)
MaxMass ← maximum integer in Spectrum
Initialize an array Count of size MaxMass with all values = 0
Sort Spectrum in increasing order
for each M1 in Spectrum
      for each M2 > M1 in Spectrum
            Diff ← M2 – M1
            Count[Diff] ← Count[Diff] + 1
ConvolutionSpectrum ← empty set
for i ← 1 to MaxMass
      if Count[Diff] > 0
            add (Diff, Count[Diff]) into ConvolutionSpectrum   //(mass, multiplicity)
Sort ConvolutionSpectrum in decreasing order of the multiplicity
Output ConvolutionSpectrum
```

5. (10 pts) Mr. Fuzzy claims that the theoretical spectrum of a cyclopeptide is a superset of the theoretical spectrum of any corresponding linear peptide (that

resulting from the break of the circular peptide at a arbitrary location), and thus in order to reconstruct a cyclopeptide from a given mass spectrum, one can always first reconstruct a linear peptide. Is Mr. Fuzzy correct? If yes, explain your answer; otherwise, give a counterexample.

6. (10 pts) Devise a dynamic programming algorithm to solve the *Counting peptides with Given Mass Problem* (page 193): given an integer, count the number of linear peptides having the integer mass m.

7. (10 pts) In a noisy mass spectrum of a cyclopeptide, the maximum value in the spectrum may not be the parent mass of the cyclopeptide. Devise an algorithm to compute the parent mass of a cyclopeptide from a given noisy mass spectrum.

8. (10 pts) Construct the alignment graph of ACGTTAA and AGTTTA (using the score = 3 for matches, and = -2 for mismatches and gaps). Show the optimal alignment, and its corresponding path in the alignment graph.

9. (10 pts) Modify the dynamic programming algorithm for the global alignment of two DNA sequences to solve the Overlap Alignment Problem (page 264): given two strings *u* and *v*, and a scoring matrix *Score*, find a highsest-scoring *overlap alignment*, i.e. a global alignment of a suffix of *u* and a prefix of *v*.

10. (10 pts) Does an optimal multiple alignment always induce optimal pairwise alignments? If yes, explain your answer; otherwise, give a counterexample.

11. (Bonus question, extra 10 pts) The challenge problem on page 208 of the textbook.