Bioinformatics: Homework 2
Jane Lockshin
October 04, 2018

Problem 2 (10 points) Answer the following questions in a few sentences:
1. In protein alignments we don't consider just match/mismatches, but instead consider similarities between amino acids. How are these similarities represented? What does "similarity" between amino acids even mean?
Similarities in protein alignments are usually represented by visualizations that highlight "conserved" areas among sequences (denoted by a "*"), along with areas that are "conserved mutations" (":"), "non-conserved mutations" (" "), and "semi-conserved mutations" ("."). These representations help the viewer judge the similarities of amino acid substitution.

"Similarities" means the degree of resemblance between amino acids so we can infer information about patterns in a particular region in an observed sequence. This is important so we can gather information about when and where a sequence has been conserved or contains a mutation.

2. Explain the advantage of BLAST's seeding method over a standard Needleman-Wunsch local alignment algorithm.

First of all, BLAST is much faster than a standard local alignment algorithm, because instead of going through the entire data that the query sequence is aligning to, BLAST only uses the dynamic programming algorithms in areas that look like they are most likely to match with the query sequence. This results in a system that is 10-50 times faster than the standard Smith-Waterman approach for local alighments. In addition, BLAST only returns local alignments that are greater than or equal to a scoring metric, returning only the most accurate results.

Problem 3 (10 points) Suppose that the true genome sequence of an organism is AGTCGATCGTG and in a sequencing experiment, one obtains exactly three reads: AGTCG, CGATC, and TCGTG.

1. Suppose you use the greedy algorithm for assembling these three reads into a super string. Will this approach recover the true genome sequence? Explain your reasoning.
Yes, the greedy algorithm will recover the true genome sequence. That is because the greedy algorithm will look at the reads and overlap the similar bases of each read. For example, the first two reads will be compared, and then combined into AGTCGATC, and then that will be compared to the third read, giving the result: AGTCGATCGTG.

2. Suppose instead that you use the set of k-mers found in the reads as the spectrum (the de Bruijn approach), with k = 4. Will this approach recover the true genome sequence? Explain your reasoning.

No, I don’t believe that this approach will recover the true genome sequence, because a k of 4 is too large for reads that only have 5 bases. While constructing the graph, with k = 4, you get: AGTC, CGAT, and TCGT. Then when you find the prefixes/suffixes and construct the graph, it will only contain four edges, and an Euler cycle cannot be found between those four.