Computational Biology Problem Set 3 Write-Up Questions

Kerim Celik

01/25/2018

A. Test-Cases
    a. Strategy
        i. To ensure that my code was working correctly, I first ran a few test cases for alignments with obvious answers so I could manually check that scores were correct. I then tried a few combinations to multiple correct optimal alignments, to make sure that didn't break anything either. Finally, I made sure my code could handle empty strings in one or both of the sequence inputs. Having tested all the edge cases I could think of and having verified the accuracy of the algorithm in a variety of cases, I concluded testing.
    b. Local Alignment Test Cases
        i. Two identical strings
        ii. Two strings with no characters in common
        iii. Two identical regions on one string that have equal local scores when aligned with the second string
        iv. One empty string
        v. Both empty strings
    c. Global Alignment Test Cases
        i. Two identical strings
        ii. Two strings with no characters in common
        iii. Two identical strings, but one string has a middle section of characters not in the other
        iv. One shorter string that can optimally align to multiple locations on a longer string
        v. One empty string
        vi. Two empty strings
B. BRCA1 Alignment
    a. I used the following score values:
        i. Match: +5
        ii. Mismatch: -5
        iii. Gap-open: -50
        iv. Gap-extend: -2
    b. We know that exons are parts of the whole gene with the intron regions spliced out. Therefore we should expect that the exons will have long gaps between each other when globally aligned to the whole gene or a fragment of the whole gene. Since the gaps are long, we should not penalize gap extension much at all. However, we really want to discourage gap opening, as we do not want our

scoring matrix to allow gaps to open in the middle of exon sequences. Therefore we assign a low gap extension penalty, but a very high gap open penalty. There may still be some variation between the exon sequences we have obtained and the partial gene, so mismatches will be penalized as normal, but nowhere near as heavily as opening new gaps.

c. Overall global alignment with affine gap score: 2800

C. TP53 Alignment

a. I used the following score values:
    i. Match: +2
    ii. Mismatch: -1
    iii. Gap-open: -4
    iv. Gap-extend: -1

b. Resulting pairwise alignment scores:
    i. Human to cat, local: 336
    ii. Human to cat, global: 111
    iii. Human to rat, local: 259
    iv. Human to rat, global: 47
    v. Cat to rat, local: 241
    vi. Cat to rat, global: 24

c. Scoring justification
    i. These three sequences are different versions of the same gene, taken from three closely related species (all mammals). Therefore we should expect a high level of conservation between the three versions. We don't want to penalize gaps too much, but want to discourage them. We do this to abide by the rule that fewer indels better explain genetic variation than more indels, as indels require some sort of mutation event to appear. We want to incentivize matching where possible to find the similarities across gene versions, so we give matches a higher score relative to most penalties. Only gap opening changes the overall score more than matching, as otherwise creating new gaps would be easily paid for with only a couple of matches, allowing for too many gaps.

d. Most Similar Sequences
    i. The cat and human sequences are the most similar. Using these scoring parameters, both cat to human alignments outscore the other pairwise alignments of similar type by a significant margin. The high global alignment score is especially convincing because it measure a high level of similarity over the entirety of the gene fragment. In contrast, local alignment could just highlight that a section of two gene versions is very similar. Since the human to cat global alignment is much higher than the other global alignments in addition to the higher local alignment score, we can say with confidence that the human and cat sequences are most similar.

e. Visualization

i. The alignments are displayed in Excel files in the Problem Set directory. They are titled according to the two species' sequences aligned along with whether the alignment is local or global. The alignment is shown using red cells. For best viewing, zoom all the way out on the file. Unfortunately, even at widest zoom, a full picture cannot be observed (sorry).