Exploring Bioinformatics: A Project-Based Approach

Key for Web Exploration, Chapter 3

1. How many matching nucleotides are there between your two sequences? What is the alignment score?

   1319 nucleotides match, giving an alignment score of 4934.5.

2. How many gaps were needed to align these sequences? Is there any particular pattern to where or how the gaps occur?

   131 gaps were needed to align the sequence; most are at the beginning or end of the shorter sequence, but there are gaps scattered throughout.

3. Can you suggest where the coding sequence might occur within this segment? What is your evidence?

   Although some gaps are needed at the beginning and end of the shorter sequence to position the sequences correctly, there are many matching nucleotides between these points, so this un-gapped central region seems likely to be the coding sequence, a hypothesis strengthened by the fact that the first aligned nucleotides are ATG, potentially the start codon.

4. What is the logic behind the affine gap penalty, which imposes a large penalty for opening a new gap but a much smaller penalty for extending the size of an existing gap?

   A linear gap penalty imposes the same negative score on each and every gap. An Affine gap penalty recognizes that one indel event that removes or adds several nucleotides actually represents less disruption than many one-base insertions or deletions and should not be penalized as much. So, there's a strong penalty for the existence of a gap, but allowing the gap to extend further doesn't add much additional penalty.

5. When you align the two HA sequences using a higher gap opening penalty, does the percent identity change significantly? How about the number of gaps and their placement or size?

   Now, with the greater penalty on gaps, the optimal alignment has only 81 gaps and all but three are at the beginning and end to position the sequence. Only a single three-nucleotide gap occurs within the coding sequence.

6. Your alignments with higher and lower gap opening penalties are both optimal alignments (the best alignments given the parameters), and they give quite similar scores. Which alignment do you think is "better," biologically, and what is your justification? (Hint: what striking observation did you make when looking at the gaps in the second alignment?)

   Sometimes there's no absolutely clear way to decide which of two alignments is "better," but in this case it is very noticeable that only a single three-base gap occurs within the coding sequence. Deleting three nucleotides in one place could mean that one codon is deleted, and this would be far less disruptive to a protein sequence (which would just be missing one amino-acid) than multiple one- and two-nucleotide indels.

7. Discuss how closely the HA segments of the two modern viruses are related to each other and how closely they resemble the 1918 virus. Can you draw any conclusions from your data about the origin of HA in the 2009 pandemic virus?

   The alignment scores are a good way to compare different alignments. When the 2009 virus was aligned to a the modern, circulating Brisbane H1N1 virus, an alignment score of 4768 resulted (when the higher gap opening penalty was used). Aligning each to the 1918 virus with the same parameters gives a score of 5707 for California and 6090 for Brisbane. The relatively good alignments suggest that the 1918 virus could have been the source for both modern HA genes, but that the non-pandemic Brisbane virus is closer to this potential parent sequence today.

8. If you were to use a different segment from the same viruses for your sequence comparisons, you might come up with different answers. How is this possible?

   Because two viruses can infect one host cell, segments can be mixed and matched during the assembly of progeny viruses (“antigenic shift”). So, a single virus can actually be a comite of segments with different origins.

9. How good are the score and the percentage of sequence identity for this comparison? Why don't these statistics tell the full story in this case?

   The percent identity is only about 26%, with 71% of the sequence in gaps, giving a score of 1077, which doesn’t seem very good. However, looking at the alignment reveals that one sequence is much shorter than the other, requiring a large number of terminal gaps.

10. Suppose we only looked at the portion of the 2009 segment that actually aligned with the M2 coding region of the Brisbane strain. How would this change the percent identity? Is this degree of similarity as high as you would expect for these related viruses?

    The portion that aligns has a much higher percent identity than the overall percentage, representing a degree of similarity much greater than the initial score suggests—much more like the expected degree of relatedness.

11. How does this alignment differ from the previous one? Is the percent identity, either for the whole alignment or just for the regions that actually match, significantly better than before?

    Now, there is a small aligned region at the beginning of the sequence, a long gap, and then another aligned region; both of the aligned regions have very high similarity.

12. There is an obvious difference in how the subsequences of the M2 coding region align with the 2009 segment 7 sequence in the local alignment. Can you suggest a hypothesis for why the sequences align this way? (Hint: remember that the M2 sequence is the protein-coding sequence.) Based on your hypothesis, is the local

alignment superior to the global alignment in terms of its ability to help us understand the viruses biologically?

The excellent match to the M2 coding sequence seen in two separated sections of the whole segment 7 sequence strongly suggests that the M2 gene has two exons separated by an intron. The ability to observe this makes the local alignment the better one, biologically, in this particular instance.

13. What are the scores and sequence identities for a comparison of the two avian viruses? Are the differences between the human isolate and the avian isolates greater than the differences among avian isolates?

The two avian H5N1 sequences are 96% identical to each other, with a score of 8205. Avian sequence #1 is 97.5% identical to the human isolate, with a score of 8575, and avian sequence #2 is 95.2% identical to the human isolate, with a score of 8286. It appears that the differences between the avian and human strains are comparable to the differences between the two avian strains.

14. Based on your results (which of course are limited—it would be necessary to do many more comparisons in reality), do you think there is evidence that human adaptation is occurring in H5N1 viruses that might merit concern about human-to-human transmission in the near future?

At least in this limited data set, there is no evidence that the H5N1 HA is changing especially rapidly. It seems likely that this human isolate is just an avian strain that some human in close contact with birds was able to catch, and not a developing human H5N1.