BIO/CMPSC 300 Introduction to Bioinformatics
Spring 2016
Kristen Webb and Janyl Jumadinova
http://cs.allegheny.edu/sites/jjumadinova/300

Lab Week 6 – Investigation of Influenza Virus Strains

Given: Friday February 26, 2016
Due: Friday March 4 by 2:30pm

Objectives:

- **To understand the value of aligning genes and recognize the practical applications of this technique.**
- **To gain familiarity with the use of Web-based alignment tools to explore sequence similarity and understand how to modify their parameters.**
- **To know how the Needleman-Wunsch algorithm optimally aligns any two sequences.**
- **Understand how the Needleman-Wunsch algorithm can be modified to yield other alignments.**

Reading Assignment:

Chapter 3 in Exploring Bioinformatics textbook.

Required Deliverables (submitted through your Bitbucket repository):

1. **An electronic version of the report containing the answers to the lab comprehension questions in red.**
2. **A properly formatted and commented Python program that implements other types of alignments and a snapshot of an output that your program produces.**

General Guidelines for Labs

Work on the Alden Hall computers. If you want to work on a different machine, be sure to transfer your programs to the Alden machines and re-run them before submitting.
Keep all of your files! Don't delete your programs and reports after you hand them in---you might need them again later.
Back up your files regularly. Use a flash drive or Google Drive or whatever your favorite backup method is.
Review the Honor Code policy on the syllabus. Remember that you may discuss experiments and programs with others, but copying answers or programs is a violation of the Honor Code

Part I: Pairwise Global Alignment with the Needleman-Wunsch Algorithm

The genomes of the influenza viruses are divided into eight segments, each representing essentially the coding information for a single protein (Figure 1). Segment 4 contains the gene for hemagglutinin (HA), the viral surface protein essential for the initial interaction between the virus and its host cell. HA is one key determinant of which host(s) a particular virus can infect, because the virus cannot replicate or cause disease without being able to first bind to a host cell. The HAs of one of the major seasonal human viruses circulating before 2009, the 2009 H_1N_1 pandemic virus, and the 1918 human pandemic virus are all classified as the H_1 type, whereas recent outbreaks of severe avian flu are caused by a virus with HA classified as H_5. These classifications are based on binding of antibodies of known specificity, but sequence alignment provides much more detailed information about similarities and differences and where changes have occurred.

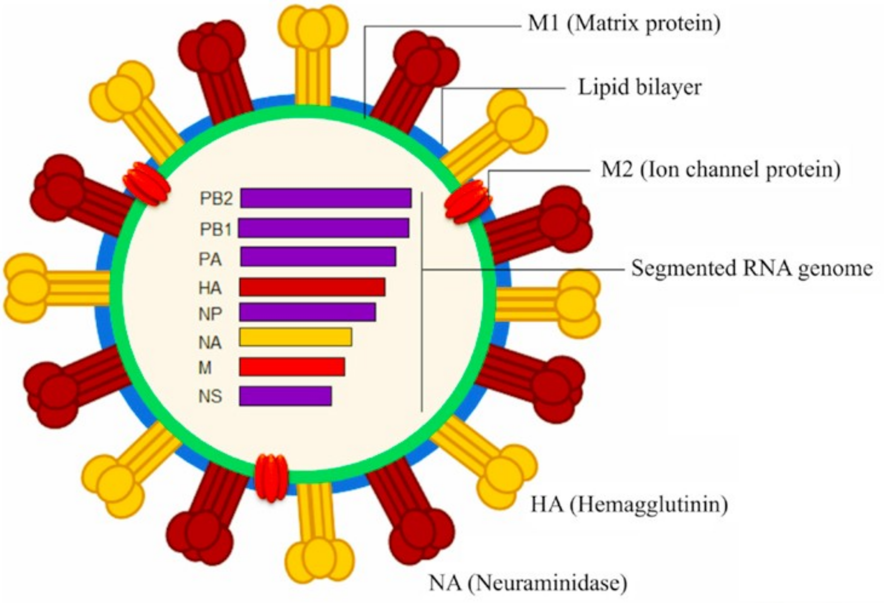

Figure 1. Schematic diagrams of influenza A virus - The segmented genome of influenza A virus encodes three envelope proteins (hemagglutinin, neuraminidase, and ion channel M2 protein), and internal nucleoprotein (NP), polymerases (PA, PB1, and PB2), matrix protein 1 (M1), and non-structural proteins (NS). The lipid bilayer is derived from host cell membrane.

We can use the Needleman-Wunsch algorithm to compare influenza virus HA segments. To start, we will compare two 2009 H_1N_1 virus strains - the reference strain designated A/California/07/2009 (H_1N_1) and the human seasonal H_1N_1 virus that was circulating at that time, A/Brisbane/59/2007 (H_1N_1). The term “reference strain” refers to a common strain of a virus that is selected for the manufacture of the seasonal Influenza vaccines developed by the Centers for Disease Control every year.

1. The files Influenza_A_Brisbane2007(H1N1)segment4.txt and Influenza_A_California2009(pandemicH1N1)segment4.txt have been uploaded to the lab6 directory in the shared repository. These files contain the DNA sequences of segment 4 of the each virus.

2. Go to the EBI-EMBL's EMBOSS webpage at www.ebi.ac.uk/Tools/emboss/ You should see a list of programs for pairwise alignment. Under the heading "Global Alignment", the program Needle is an implementation of the Needleman-Wunsch algorithm.

3. From the EMBOSS site, choose the version of Needle that compares nucleotide sequences and then paste the two viral segment 4 DNA sequences into the designated boxes. Leave the parameters set to the defaults and run Needle to align your two sequences.

Notice that you can set some parameters for comparison, most notably the gap penalty. Needle uses an ***affine*** *gap penalty, which means it imposes a larger penalty when a new gap is added and a smaller penalty when that gap is extended (our examples on Tuesday and Thursday used a* ***linear*** *gap penalty).*

At the top of your results, you will see parameters such as the gap penalty and two measures of similarity: the number and percentage of matching nucleotides (labeled "Identity") and an alignment score (based on the scoring matrix, in this case awarding a match bonus of 5). In the alignment itself, matching nucleotides are shown by a | character, mismatches by a dot (.), and gaps by a dash (-).

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

   1319 nucleotides match, giving an alignment score of 4934.5.

b. How many gaps were needed to align these sequences? Is there any particular pattern to where or how these gaps occur? If so, explain briefly.

   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.

Nearly all of segment 4 consists of coding sequence, so we would expect insertions/deletions, aka indels, especially one- or two-nucleotide indels to be mutations with serious consequences for HA proteins. Considering this, perhaps it would be valuable to consider strongly penalizing gaps.

1. Set the gap opening penalty to 50, rather than the default of 10.

d. When you align the two HA sequences using the higher gap opening penalty, does the percent identity change significantly relative to the previous alignment? 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.

e. Each of your alignments, the one with higher and the one with the lower gap opening penalties, are optimal alignments (the best alignments given the parameters), and they give quite similar scores. Which alignment do you feel is "better" biologically, and which 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.

Part II: Local Alignment with the Smith-Waterman Algorithm

Another way to use sequence alignment is to find one sequence within another. The influenza virus M2 gene, for example, is another key player in the biology of the virus: once the virus enters the cell, M2 is involved in the release of the virus genome subunits so they can travel to the nucleus and direct viral replication. Suppose we have sequenced segment 7 from the 2009 H_1N_1 pandemic virus but are uncertain what part of it represents the actual M2 coding region. To find out, we could align the well-characterized M2 coding sequence from the Brisbane strain with the full segment 7 sequence from the newly sequenced virus.

1. The files Influenza_A_California2009(pandemicH1N1)segment7.txt and Influenza_A_Brisbane2007(H1N1)M2.txt have been uploaded to the lab6 folder in the shared repository. Access these files and align them using Needle with a default gap opening penalty of 10.

    i. How good are the score and 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.

    j. 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.

Considering what you know about the Needleman-Wunsch algorithm, you should see why it might not be the best choice for aligning sequences that are so drastically different in length. Because the need to make alignments of this kind arises frequently, in 1981 Smith and Waterman published a modification of the Needleman-Wunsch algorithm that allows for **local** alignments.

A local alignment looks for optimal partial (subsequences) matches. EMBOSS includes an implementation of the Smith-Waterman algorithms called Water.

2. Choose the nucleotide version of the Water method and then set a gap opening penalty of 10 an a gap extension penalty of 0.1 and align the Influenza_A_California2009(pandemicH1N1)segment7.txt and Influenza_A_Brisbane2007(H1N1)M2.txt sequences.

   k. 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.

   l. There is an obvious difference between how the subsequences of the M2 coding region align with the 2009 segment 7 sequence in the local alignment. Can you suggest a hypotheses 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.

This alignment is very sensitive to the parameters used. Change the gap extension penalty (e.g. from 0.1 to 0.5). Although almost all bioinformatics programs come with default settings that are usable for many common purposes, this illustrates the importance of understanding the algorithm and the meaning of the parameters, as well as the value of considering what kind of alignment would be most appropriate for the sequences aligned.

Part III: Using Alignment to Investigate Virulence

Influenza viruses have received a great deal of study, and the ability to compare many strains has lead to significant advances in understanding what allows one strain to cause more severe disease than another. The H_5N_1 “bird flu” virus makes an interesting case in point. The virus causes severe influenza in birds and has become established in populations of domestic chickens and turkeys. Human cases occur sporadically, mostly in individuals heavily exposed to infected birds, such as poultry farmers, and H_5N_1 flu is severe for humans as well. Once a human case occurs, however, spread to another human is exceedingly rare, even among family members in close contact with the infected individual.

A 2006 article by van Riel et al. demonstrated that the avian H_5N_1 virus binds to a form of sialic acid receptor that in humans is only found far down in the lungs and lower respiratory system. Human viruses, in contrast, bind to a form of the receptor common in the upper respiratory tract. Thus, it is difficult for H_5N_1 to infect humans because our respiratory defenses normally prevent the virus from reaching the lungs. However, a mutant strain in which HA was altered to be able to bind to sialic acid receptors in the upper respiratory tract could be a very dangerous strain indeed.

So far, no such H_5N_1 strains that infect humans efficiently have been observed. However, we might ask whether the strains that do make it into humans tend to have altered HA genes –if so, that would suggest that either adaptive mutations could be occurring within the human host or that the viruses that cause human infections are subpopulations that are already better adapted. There are many avian H_5N_1 sequences available and a number of sequences of the H_5N_1 viruses isolated from infected humans, so we can use sequence alignment to see whether these have essentially the same HA or noticeable differences.

1. The files Influenza_A_Chicken_Vietnam2005(avianH5N1)segment4.txt, Influenza_avianHong_Kong2007(avianH5N1)segment4.txt, and Influenza_A_ChinaGD012006(humanH5N1isolate)segment4.txt have been have been uploaded to the lab6 directory in the shared repository. Perform pairwise comparisons of the three strains using the Needleman-Wunsch algorithm.

    m. What are the scores and sequence identities for the 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.

n. Based on your results (which of course are limited –it would be necessary to do many more comparisons in reality), do you believe there is evidence that human adaptation is occurring in H_5N_1 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.

Part IV: Implementation of Alignment Techniques

Required:

1. Although global alignments are useful, they don't solve all alignment problems. When you need to find the coding sequence for a gene within a longer DNA sequence, the Needleman-Wunsch algorithm will penalize not only the internal gaps but also the terminal gaps (gaps at the beginning and the end of the alignment). For example, in the alignment shown below, a global alignment finds all of these alignments to be optimal. However, the first alignment would be the best because it only includes terminal gaps used to position the short sequence. The alignment when we eliminate the gap penalty for terminal gaps is called a *semiglobal alignment*.

```
CGCTATAG    CGCTATAG        CGCTATAG

- - CTA - - -    C - - TA - - -        - - C - - TA -
```

Modify the Needleman-Wunsch program so it implements a semiglobal alignment by eliminating the gap penalty for terminal gaps. Make sure to rename the program as you copy it. (Hint: This requires just a few minor changes in the code given in class. Cosider what parts of the matrix represent the terminal gaps. Focus on what the outside rows and columns of the matrix represent and how they are used.) Test your modified implementation using short sequences from the example above first. Then, run your program on a real-world example using the sequence of 2009 H1N1 pandemic influenza virus segment 7 and the coding sequence for the 2009 H1N1 virus M1 gene (in the lab6 directory). Your program should successfully pick out the M1 coding sequence within the segment 7 sequence.

2. Local alignments solve the problem of finding and aligning conserved regions in otherwise dissimilar sequences by looking for optimal partial or sebsequence matches between the sequences. For example, in the sequences AAAGCTCCGATCTCG and TAAAGCAATTTTGGTTTTTTTCCGA, there are two similar regions AAAGC and TCCFA, separated by other very different regions. A global and semiglobal alignments will find AAAGC alignment but will not correctly align the sequences so that TCCGA

sequences also match up. Large gaps are expected when finding subregions of similarity and should not negatively affect the alignment score. This was the basis for Smith-Waterman algorithm, which requires only a few changes to a semi-global alignment problem.

Describe a modified algorithm that would find local alignments given two sequences. Include details on how the matrix is initialized, how the scores are calculated, and where the alignment path should start and end.

Optional:

1. Make necessary changes to your semiglobal alignment program to implement the local alignment algorithm you described in the previous part. Test your program with the short sequences given in the example above. Then, run your program with the segment 7 sequence for the 2009 H1N1 pandemic influenza virus and the coding region of the M2 gene from the Brisbane seasonal strain and see if your program gives you the same result as the EMBOSS implementation of the Smith-Waterman algorithm.

2. Modify your previous implementations to find all possible optimal alignments.