Title:	Pairwise Alignment Matrix Calculation
Date:	05/10/2023
By:	Ethan Corgatelli

1.0 Abstract

In this project I did the first step of creating a profile HMM from a set of unaligned sequences, which is to create a diagonal matrix of pairwise alignments of the set of new sequences using a scoring matrix. I tested two scoring matrices: 1) the scoring matrix that was estimated by my code in Project 3, and 2) the blosum50 scoring matrix.

My code can be found here: http://tiny.cc/cs415-corg-p4-code

2.0 Methods

This section contains a brief description of how my program generates the pairwise alignment matrix. First, there is a doubly-nested for-loop. The outside loop iterates over every sequence in the input, and the inside loop iterates from the next sequence after the outside loop's sequence to the end of the list of sequences. This way, we can compare every set of sequences exactly once (after comparing A to B, we don't want to compare B to A because alignment is symmetric).

For each pair of sequences, we create an alignment matrix and a direction matrix based on the lengths of the two sequences. Then, in another doubly-nested for-loop, we calculate the values that should go in the alignment and direction matrices using Needleman-Wunsch, starting in the top left, working right, then down. Then finally we add that alignment score (from the bottom right corner of the matrix) to a dictionary.

The last step is to backtrace the direction matrix to calculate what the optimal global alignment should be. In a while loop which breaks only once we've reached the top left corner of the matrix, we start in the bottom right and follow the symbols in the direction matrix. Each time a diagonal step is taken, characters from both sequences are added to the new aligned sequences variables. Each time a horizontal or vertical step is taken, one of the new sequence variables gets a character from its original string and the other gets a dash (a gap) inserted. After backtracing the direction matrix, the new aligned sequences are added to a dictionary.

3.0 Results

This section contains the resulting pairwise scores, both from my code's estimated scoring matrix and from the blosum50 scoring matrix, respectively. The sequences will be referred to 0-indexed in this section. Alignments are shown a line of each sequence at a time.

Using My Code's Estimated Scoring Matrix

Full results here: https://pastebin.com/raw/fn0wNMvg

Best scoring alignment (score: 639 – seq. 6 & seq. 7):

```
hvrlqdqap-m-ytl-vpvyspwsytydsnyyvyqmqkpvpwvwnyvtydvstwyywykwwpsgpntwsvgvyvahhgghwgttlymgtqpfmqklrm
hepqqdcapnfpwtlmtcvws-w-kwl-svgyanqmqnpvpwvwnyqtyfvstwyywykwwpsgpnthivyvyvkhhgghwgpvqymgfqpfmqgvrf

-iiapcdlqmrwcgelwnehlfq-dfc-adravhpanf-lnttrvnhmtmkvrnsikpqekeefrftts-cwm-arfmitc
pqgvggfyh-riifelw--tyrngadhwadravsgalfwl-llvknshtikviysq-pwppsgir-kycrnpyqtvhpgtm
```

Worst scoring alignment (score: 322 – seq. 1 & seq. 4):

```
vyvwtvwvwwwmgpwacdagldkyq-pyeryncynlp-nrstc-cvinvnkdyycdmhayvacqfqeaghylp---qevrsckypvfycwdwi--rcf
-yqrhkm-nwnyvpdtcpwsy-yylmnyym-slmmvpvpyqsctgtfcvstwyy---s-yf-sipsg-psvwpvgqqgadiavkkasflgdwmvnfmq

kgcsqr-ppkkpksrcwgqgyfwdtdtsgygsq
vfgwerhkpqmdrird-e-gcpnwffrhdm-ta
```

Pairwise scores:

```
{'01': 471, '02': 429, '03': 503, '04': 342, '05': 398, '06': 403, '07': 381,
            '12': 437, '13': 431, '14': 322, '15': 369, '16': 365, '17': 355,
                       '23': 417, '24': 337, '25': 393, '26': 392, '27': 393,
                                  '34': 322, '35': 364, '36': 373, '37': 357,
                                             '45': 428, '46': 405, '47': 369,
                                                        '56': 631, '57': 591,
                                                                   '67': 639}
```

Using The Blosum50 Scoring Matrix

Full results here: https://pastebin.com/raw/wXC4Hmgm

Best scoring alignment (score: 524 – seq. 6 & seq. 7):

```
hvrlqdqap-m-ytlvp-vyspwsytydsnyyvyqmqkpvpwvwnyvtydvstwyywykwwpsgpntwsvgvyvahhgghwgttlymgtqpfmqklrm
hepqqdcapnfpwtlmtcvws-wkw-l-svgyanqmqnpvpwvwnyqtyfvstwyywykwwpsgpnthivyvyvkhhgghwgpvqymgfqpfmqgvr-

iiap--c-dlqmrwcgelwn-ehlfqd-fcadravhpanf-lnttrvn-hmtmkvrnsikpqekeefr-fttscwmar--fm-i---tc
-f-pqgvggfyhriifelwtyrn-gadhw-adravsgalfwlllvk-nsh-tikviys-qpwppsgirky---c---rnpyqtvhpgtm
```

Worst scoring alignment (score: 57 – seq. 1 & seq. 6):

```
-v----y--vwt-vwvwwwmgpwa-c-dagldky--q---p----yeryn--c----yn-lpn-rstccv-inv-nk-dyy-cdmhayvacq-f-qea
hvrlqdqapmytlvpv--y-spwsytydsnyyvyqmqkpvpwvwnyvtydvstwyywykwwpsgpntwsvgvyvahhgghwgttl--ymgtqpfmqkl

ghyl-p-q-evrsc-k-y-p-vf--ycwd-wir-c-f----k-g-cs-q-rppkkp--k-s-r----cwgqgyfwdtdtsgygsq
rmiiapcdlqmrwcgelwnehlfqdfcadravhpanflnttrvnhmtmkvrnsikpqekeefrfttscw-marfm-i-t----c-
```

Pairwise scores:

```
{'01': 354, '02': 206, '03': 441, '04': 104, '05': 118, '06': 111, '07': 127,
            '12': 283, '13': 274, '14': 93,  '15': 77,  '16': 57,  '17': 114,
                       '23': 207, '24': 99,  '25': 58,  '26': 71,  '27': 103,
                                  '34': 111, '35': 85,  '36': 82,  '37': 88,
                                             '45': 282, '46': 197, '47': 181,
                                                        '56': 437, '57': 387,
                                                                   '67': 524}
```

4.0 Conclusion

In this section I'll discuss the following questions: 1) do the alignments seem plausible, 2) how did the choice of scoring matrix affect the results, and 3) does it seem that the sequences could be clustered into subsets?

How 'good' do you think the results were, i.e. do the alignments seem plausible?

Although I am no expert in the matter, the resulting alignments did seem 'good'/plausible in my opinion, as they looked quite similar to the alignments that can be found in the textbook. The majority of gaps are somewhere in the middle of the alignments rather than being all bunched up either side. My code's results do seem plausible.

Were the results from the two substitution matrices the same? If not, what was the difference and which seemed better?

Changing the scoring matrix used had only a slight effect on the results. I tried to choose gap penalties that would make the results as similar as possible, but the alignments from the blosum50 scoring matrix tended to have more gaps. However, most gaps from my scoring matrix's alignment are also represented in the blosum50 results (they are in roughly the same places). This leads me to believe that my results are on the right track. I'm guessing that the blosum50 results are better than mine, since it was created from much more data and with a much more rigorous process, but my results also seem reasonable.

Based on the alignment scores do you think the sequences could be clustered into two or more subsets?

Yes, it does seem like these alignments could be clustered. Here is very rough estimation of how these alignments might be clustered into two groups, based only on my working it out by hand based on alignment scores:

Group 1: Sequences 0, 1, 2, and 3.

Group 2: Sequences 4, 5, 6, and 7.