CSEP527 HW 1- Smith-Waterman Local Alignment

Isabelle Lee

October 31, 2020

1 Preparation

The functions on the given sequences for this assignment is summarized below in the table.

Ascension	Known function
P15172	promote muscle specific target genes and play a role in muscle differentiation for humans
P17542	plays role in blood cell differentiation for humans
P10085	muscle differentiation for mice
P16075	muscle differentiation for chicken
P13904	muscle differentiation for african clawed frog
Q90477	muscle differentiation for zebrafish
Q8IU24	muscle differentiation for invertebrate marine animal amphioxus
P22816	early muscle development for fruit flies
Q10574	development of peripheral sense organs for c elegans
O95363	responsible for charging of tRNA in mitochondrial transcription

Based on lecture notes, I suspect that amino acids that play similar roles would have similar sequences. So, the ascensions that play roles in muscle differentiation would be sequentially similar, and therefore have a longer substring match. This would imply that their BLOSUM scores would be higher. Based on course lecture slides, the function would play a more significant role than the species. For instance, P15172 would have a significant alignment to P10085. On the other hand, I wonder how speciation affects the sequences. Perhaps, similar species (i.e. vertebrates vs invertebrates, human and mice vs chicken) would generally have a more significant alignment, and therefore a higher score.

2 Implementations

Please take a look at `src/smith-waterman.py`! The commands to generate `output.txt` is organized in a shell script, `hw2.sh`.

3 Deliverable

3.1 Alignment and scores for the pair of strings x1=deadly and x2=ddgearlyk

First test case was an alignment between two short pairs of strings, and my output is pasted below. For this output, the number of random trials using Fisher Yates shuffle was 999.

```
Comparison of two string inputs:  ['D', 'E', 'A', 'D', 'L', 'Y']
and  ['D', 'D', 'G', 'E', 'A', 'R', 'L', 'Y', 'K']
Score Matrix:

[[ 0.  0.  0.  0.  0.  0.  0.  0.  0.  0.]
 [ 0.  6.  6.  2.  2.  0.  0.  0.  0.  0.]
```

```
 [ 0.  2.  8.  4.  7.  3.  0.  0.  0.  1.]
 [ 0.  0.  4.  8.  4. 11.  7.  3.  0.  0.]
 [ 0.  6.  6.  4. 10.  7.  9.  5.  1.  0.]
 [ 0.  2.  2.  2.  6.  9.  5. 13.  9.  5.]
 [ 0.  0.  0.  0.  2.  5.  7.  9. 20. 16.]]

maximum score:  20.0
substring of first input:  DE--ADLY
substring of second input:  DDGEARLY

empirical p-value:  0.044
```

3.2 Alignments and scores for each protein sequence in section 1.

Then, the second set of experiment was between two chosen alignments out of 10 FASTA sequences we downloaded for preparation. For these outputs, the number of random trials for all of them were 999. The table below summarizes the alignment scores calculated from BLOSUM and my script. As was the

	P15172	P17542	P10085	P16075	P13904	Q90477	Q8IU24	P22816	Q10574	O95363
P15172	NaN	143	1500	1020	978	893	428	368	118	56
P17542	143	NaN	128	129	128	112	144	123	156	66
P10085	1500	128	NaN	1043	1002	925	440	367	118	52
P16075	1020	129	1043	NaN	1147	1093	448	414	120	61
P13904	978	128	1002	1147	NaN	1104	450	410	120	72
Q90477	893	112	925	1093	1104	NaN	449	410	117	62
Q8IU24	428	144	440	448	450	449	NaN	446	125	45
P22816	368	123	367	414	410	410	446	NaN	124	74
Q10574	118	156	118	120	120	117	125	124	NaN	67
O95363	56	66	52	61	72	62	45	74	67	NaN

case with course slides, the sequences that dealt with similar functionality had higher alignment scores in general, in 1000s. Even the fruit fly sequences (Q10574) was more correlated with sequences related to muscle differentiation. One sequence that had nothing to do with muscle developments or differentiation was the last sequence (O95363) from humans, and it had the smallest scores out of all comparisons. Even with other human amino acid sequences, it had low alignment scores compared to amino acids from other species that shared similar functions. So, in terms of alignments, speciation doesn't matter as much after all, as long as the functions of amino acids are related.

3.3 Empirical p-values for the significance of the alignments of P15172 versus Q10574 and for P15172 versus O95363.

For the last experiment, the empirical p-values were calculated for two different sets of FASTA files. For this output, the number of random trials using Fisher Yates shuffle was 999. In summary, P-value for P15172 versus Q10574 was 0.001 (score 118), and meanwhile, p-value for P15172 versus O95363 was 0.532 (score = 56). The smaller the p-value, the more statistically significant the result. Therefore, the higher the alignment scores, the more significant the result, given this output. This intuitively lines up with our understanding of alignment scores; since the higher alignment score indicates the functionality of the sequences are significantly tied, the statistical significance of the alignment score bolsters this intuition.