Petra Wijngaard
Bio 131
May 7th, 2018

Verifying Sequence Similarity in Spectraplakins

Motivation and Introduction

My project aimed to get the global similarity of the Actin binding domains of the spectraplakins *Drosophila melanogaster* Shot isoform A and human ACF7 isoform 1. In Voelzmann *et al.*, the researchers claimed that *D. melanogaster* Shot and human ACF7 shared 76% sequence similarity in their Actin binding domains (AAs 1-300) (2017). [1] To test Voelzmann *et al.*'s claim, I made a program that used the Needleman-Wunsch algorithm to perform a global sequence alignment on the first 300 amino acids of Shot and ACF7. The program uses a constant gap penalty and not the affine gap penalty, which is when there are separate penalties for the creation and extension of idel gaps in an alignment. Additionally I calculated a local sequence alignment using the Smith-Waterman algorithm. To test to see if my outputs were accurate I compared them against the Needleman-Wunsch global alignment and Smith-Waterman local alignment on NCBI's BLAST (Altschul *et al.*, 1997).[2]

Data and Processing

The sequences for Shot and ACF7 were collected as FASTA sequences from UniProt (Bateman *et al.*, 2017).[3] I manually saved the FASTA outputs as text files and manually removed their headers. The sequences were stripped and cleaned using Python's .replace and .strip functions. Only the first 300 amino acids were aligned from each sequence.

[1] Voelzmann, A., Liew, Y.-T., Qu, Y., Hahn, I., Melero, C., Sánchez-Soriano, N., and Prokop, A. (2017). Drosophila Short stop as a paradigm for the role and regulation of spectraplakins. Seminars in Cell & Developmental Biology 69, 40–57.
[2] Altschul, S.F., Madden, T.L., Schäffer, A.A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D.J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 25, 3389–3402.
[3] Bateman, A., Martin, M.J., O'Donovan, C., Magrane, M., Alpi, E., Antunes, R., Bely, B., Bingley, M., Bonilla, C., Britto, R., et al. (2017). UniProt: the universal protein knowledgebase. Nucleic Acids Res 45, D158–D169.

Steps

The program takes the two sequences from specified text files as strings, `string1` and `string2`, and the program takes the scoring matrix, `pam` (which is not necessarily the PAM matrix), and the penalty value for indels, `indel`. `pam` can either be the PAM250 matrix or the BLOSUM62 matrix. `indel` is either `-8` if using PAM250 `or -4` if using BLOSUM62, according to references from Compeau and Pevzner (2015).[4]

The program initializes two tables, `table` and `backtrack`, with `len(string1)` columns and `len(string2)` rows. `table`, `pam` and `indel` are used for the Manhattan Tourist Problem, to find the longest common subsequence. `backtrack` starts at the end of the Manhattan Tourist Problem and with `pam` and `indel` calculates the global sequence alignment, `align1` and `align2`. The program then takes `align1` and `align2` and makes a three row display to show where the aligned sequences do and do not match.

The program also outputs summary statistics of the alignment to the console: score, identity and similarity. Score is the sum of the PAM/BLOSUM scores at the end of the alignment in `table`. Identity is the count of every time the two sequences had exactly the same amino acid in the same position, divided by the total length of the alignment. Similarity is the count of every time the PAM/BLOSUM score is positive divided by the total length of the alignment.

The program can also calculate a local alignment using the Smith-Waterman algorithm, using much the same process as the one described above. The difference in the local alignment is that the program instead only finds the substring that maximizes the alignment score instead of the the entire alignment. The user can switch between local and global alignments by editing a line in the code.

[4] Compeau, P., and Pevzner , P. (2015). Bioinformatics Algorithms: An Active Learning Approach. (La Jolla: Active Learning Publishers).

Results and Conclusion

The program outputted an alignment (Listing 1). To test to see if my output was accurate I compared it against the Needleman-Wunsch global alignment on NCBI's BLAST (Altschul *et al.*, 1997). BLAST's Needleman-Wunsch does not permit constant gap penalties and instead only uses the affine gap penalty with a limited set of preset options for penalties. Table 1 is a comparison of my program's global alignment with BLAST's global alignment set on its defaults.

Since neither my program nor BLAST gave similarities close to the 76% claimed by Voelzmann *et al.*, I suspected that Voelzmann *et al.*, used local alignment within the Actin binding domain rather than global alignment (2017). BLAST's local alignment for the first 300 amino acids of Shot and ACF7 was 75%. While I was not successfully able to modify my code to compute local alignment, I did find success modifying the model solution of Homework 6.2 to give the local alignment of a protein sequence (Listing 2). I again compared this result with BLAST (Table 2). These results were much closer to the 76% similarity claimed by Voelzmann *et al.,* and BLAST's similarity score was 75% (2017).

In both Table 1 and Table 2, my program's usage of a constant gap penalty rather than an affine gap penalty means that it produces different results than BLAST does. Using a constant gap penalty is insufficient for accurate alignments of biological sequences, and an affine gap penalty is a a better choice(Compeau and Pevzner, 2015). The differences between my program and BLAST are therefore evidence of the inaccuracy that using a constant gap penalty introduces.

Running a local alignment on the Actin binding domain of Shot and ACF7 through BLAST gave a similarity of 75%, close to the 76% similarity claimed by Voelzmann *et al.* (2017). In this project, assuming that Voelzmann *et al.* meant local alignment, I have independently verified the claim that Voelzmann *et al.* made (2017). This project has also illustrated the importance using an affine gap penalty, as my program, with its constant gap penalty, produces results for local and global alignment that are different from the reference BLAST alignments. To expand this program for general use, I would want to incorporate the affine gap penalty into it.

Table 1. Comparison of settings and outputs for the Needleman-Wunsch algorithm for global alignment performed by my program and NCBI BLAST.

	My Program:	NCBI Default Global Alignment:
Parameters	BLOSUM62 Constant gap penalty: -4	BLOSUM62 Affine gap penalty: - 11 existence -1 extension
Results	Align Length: 370 Score: 186 Similarity: 37.57 % Identity: 30.27 %	Alignment Length: 300 Score: 442 Similarity: 45% Identity: 34%

Table 2. Comparison of settings and outputs for the Smith-Waterman algorithm for local alignment performed by my program and NCBI BLAST.

	My Program:	NCBI Default Local Alignment:
Parameters	BLOSUM62 Constant gap penalty: -4	BLOSUM62 Affine gap penalty: -11 existence -1 extension
Results	Align Length: 125 Score: 481 Similarity: 83.2 % Identity: 71.2 %	Align Length: 232 Score: 591 Similarity: 75% Identity: 61%

Listing 1. Global alignment between Shot (top) and ACF7 (bottom) with comparison row in the middle, with a matching amino acid indicating identity, a + indicating similarity and a _ indicating an indel.

```
0000 MTSHSYYKDRLGFDPNEQQPGSNNSMKRSSSRQTTHHHQSYHHATTSSSQ 0050
     M+S_S___D____+__E++__S++S+_R_S+R_++++++SY_++++S+S+
0000 MSS-S---D----E--ETL--SERSC-R-SER-SCRSERSY-RSERSGSL 0050

0051 SPARISVSPGGNNGTLEYQQVQREQRDRELYSNNGSLHHHQHHHHHHRHS 0100
     SP+____+P_G_+_TL++________+__L_____+L_H+Q____++R_+
0051 SPC----PP-G-D-TLPW--------N--L-----PL-HEQ----KKR-K 0100

0101 TTGSASSPLYENSSSPAAPKKAKHSSTQAQPQGGYEDALTQFKDERDAIQ 0150
     ___S++S+L_+____P_A_++A____+_+__+__+__A_____DERD++Q
0101 ---SQDSVL-D----P-A-ERA----V-V--R--V--A-----DERDRVQ 0150

0151 KKTFTKWVNKHLKKANRRVVDLFEDLRDGHNLLSLLEVLSGEHLPREKGK 0200
     KKTFTKWVNKHL+K++++++DL+EDLRDGHNL+SLLEVLSG++LPREKG+
0151 KKTFTKWVNKHLMKVRKHINDLYEDLRDGHNLISLLEVLSGIKLPREKGR 0200

0201 MRFHMLQNAQMALDFLRYKKIKLVNIRAEDIVDGNPKLTLGLIWTIILHF 0250
     MRFH+LQN+Q+ALDFL+++++KLVNIR++DI+DGNPKLTLGLIWTIILHF
0201 MRFHRLQNVQIALDFLKQRQVKLVNIRNDDITDGNPKLTLGLIWTIILHF 0250

0251 QISDIVV-GKEDNVSAREALLRW-----A-----R--R--S--T------ 0300
     QISDI++_G+++++SA+E+LL+W_____A_____+__+__S__+______
0251 QISDIYISGESGDMSAKEKLLLWTQKVTAGYTGIKCTNFSSCWSDGKMFN 0300

0301 A---RY-PG-V---RV-------N-D--F--------T-----------S 0350
     A___RY_P+_V___RV_______N_+__F________T___________S
0301 ALIHRYRPDLVDMERVQIQSNRENLEQAFEVAERLGVTRLLDAEDVDVPS 0350

0351 ----S---WRDGL--AF-SA 0370
     ____S___+++++__AF_++
0351 PDEKSVITYVSSIYDAFPKV 0370
```

Listing 2. Local alignment between Shot (top) and ACF7 (bottom) with comparison row in the middle, with a matching amino acid indicating identity and a + indicating similarity.

```
0000 EDALTQFKDERDAIQKKTFTKWVNKHLKKANRRVVDLFEDLRDGHNLLSL 0050
     E+A+++++DERD++QKKTFTKWVNKHL+K++++++DL+EDLRDGHNL+SL
0000 ERAVVRVADERDRVQKKTFTKWVNKHLMKVRKHINDLYEDLRDGHNLISL 0050

0051 LEVLSGEHLPREKGKMRFHMLQNAQMALDFLRYKKIKLVNIRAEDIVDGN 0100
     LEVLSG++LPREKG+MRFH+LQN+Q+ALDFL+++++KLVNIR++DI+DGN
0051 LEVLSGIKLPREKGRMRFHRLQNVQIALDFLKQRQVKLVNIRNDDITDGN 0100

0101 PKLTLGLIWTIILHFQISDIVVGKE 0125
     PKLTLGLIWTIILHFQISDI++++E
0101 PKLTLGLIWTIILHFQISDIYISGE 0125
```