# VCU Bioinformatics and Bioengineering Summer Institute
# Scoring and protein alignments: Problem Set 3A

P3.1. Modify BlastN so that it no longer prints out a complete match but prints out instead only each initial exact match of a word.

P3.2. Examine BlastN and determine the values used for the following quantities:

| | | |
|---|---|---|
| **a.** Match reward | **c.** Gap open penalty | **e.** Word size |
| **b.** Mismatch penalty | **d.** Gap extension penalty | |

P3.3. Modify BlastN so that it prints out for each hit both the raw score and the score in bits. To do this you may need to find values for lambda and K. Do this by running ANY pairwise sequence comparison at the NCBI site, using the same parameters you use in local BlastN, and noting the values of lambda and K at the end of the output.

P3.4 Estimate how much more efficient BlastN is than a full Smith-Waterman algorithm. Proceed as follows.

A. Presume that the total time spent by each program is proportional to the number of cells in scoring tables each has to calculate (so your job is reduced to figuring out how many cells that is in each case).

B. Consider a specific case of a comparison of a 100-nucleotide query sequence with the E. coli genome. How big would the Smith-Waterman scoring matrix be?

*Don't know how big the E. coli genome is? Get BlastN to tell you! Note that*

*Len(variable$)*

*gives you the length of the variable*

C. OK, you got half the job done. Now you need to find out how many cells Blast would need to calculate. Easiest way is to just count one for every time BlastN calculates a score for a cell.

P3.5. Does local BlastN filter sequences? (test with a sequence you know is filtered by NCBI Blast)

P3.6. The subject arose as to how Blast handles gap penalties: should the first gap be included in the calculation of the gap extension penalty? For example:

**Table 1: Comparison of methods to calculate gap penalty**

| **Method** | **Formula for gap penalty** | **Example: penalty for**<br>`AGGC` open → -5<br>`T--G` ext → -2 |
|---|---|---|
| Smith-Waterman | Gap_opening + gap_extension · (gap_size-1) | 7 |
| Alternative (Blast?) | Gap_opening + gap_extension · gap_size | 9 |

a. Which method does local *BlastN* use? Answer this empirically and by looking at the program.

b. How about NCBI *BlastN*?

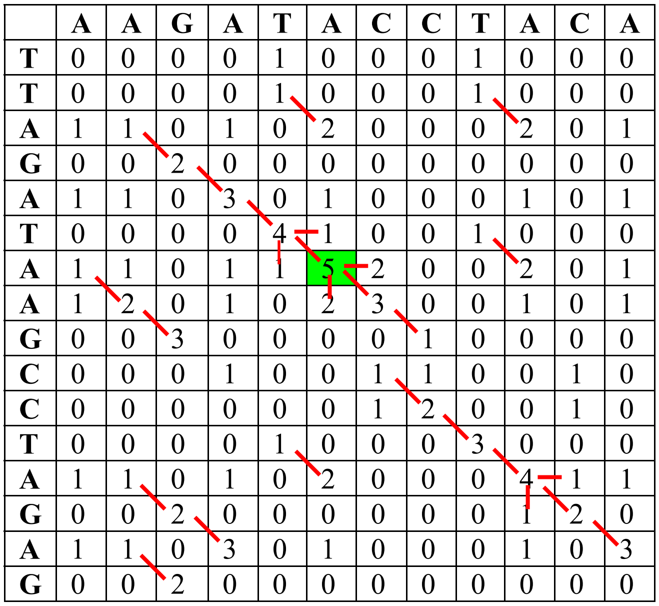
P3.7. How do you explain the fact that *BlastN* cannot find the evident similarity between DG47 and the *lef* gene?

P3.8. Consider Smith-Waterman scoring table below (equivalent to Table 2 of the notes from Tuesday).

a. Make just one change[1] in the query sequence so that the first long diagonal match (AGATA) is connected to the second long diagonal match (CCTA).

b. Make the minimum number of changes in the query sequence required to produce a score than any in the table.

c. What would be the effect on the original Figure 1 of changing the open-gap-penalty from 3 to 2? Based on this observation, how would you modify the claim that the Smith-Waterman algorithm finds "the best local alignment between two sequences"?



| | **A** | **A** | **G** | **A** | **T** | **A** | **C** | **C** | **T** | **A** | **C** | **A** |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **T** | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 1 | 0 | 0 | 0 |
| **T** | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 1 | 0 | 0 | 0 |
| **A** | 1 | 1 | 0 | 1 | 0 | 2 | 0 | 0 | 0 | 2 | 0 | 1 |
| **G** | 0 | 0 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| **A** | 1 | 1 | 0 | 3 | 0 | 1 | 0 | 0 | 0 | 1 | 0 | 1 |
| **T** | 0 | 0 | 0 | 0 | 4 | 1 | 0 | 0 | 1 | 0 | 0 | 0 |
| **A** | 1 | 1 | 0 | 1 | 1 | 5 | 2 | 0 | 0 | 2 | 0 | 1 |
| **A** | 1 | 2 | 0 | 1 | 0 | 2 | 3 | 0 | 0 | 1 | 0 | 1 |
| **G** | 0 | 0 | 3 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 |
| **C** | 0 | 0 | 0 | 1 | 0 | 0 | 1 | 1 | 0 | 0 | 1 | 0 |
| **C** | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 2 | 0 | 0 | 1 | 0 |
| **T** | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 3 | 0 | 0 | 0 |
| **A** | 1 | 1 | 0 | 1 | 0 | 2 | 0 | 0 | 0 | 4 | 1 | 1 |
| **G** | 0 | 0 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 2 | 0 |
| **A** | 1 | 1 | 0 | 3 | 0 | 1 | 0 | 0 | 0 | 1 | 0 | 3 |
| **G** | 0 | 0 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |

Scoring table based on match reward = 1, mismatch penalty = -2, gap opening penalty = -3, gap extension penalty = -2.

P3.9. Frequently sighted and aligns with the amino acid sequence DIVIT to give a score of 13 using BLOSUM62 as the scoring table. What is it? (see notes for Wednesday as a source of BLOSUM62)

[1] Insertion/deletion not allowed