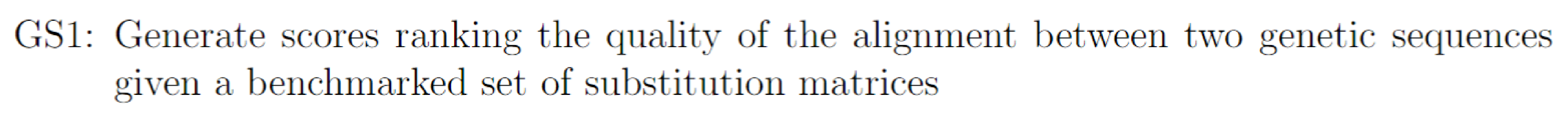
Proof of Concept

Uriel Garcilazo Cruz

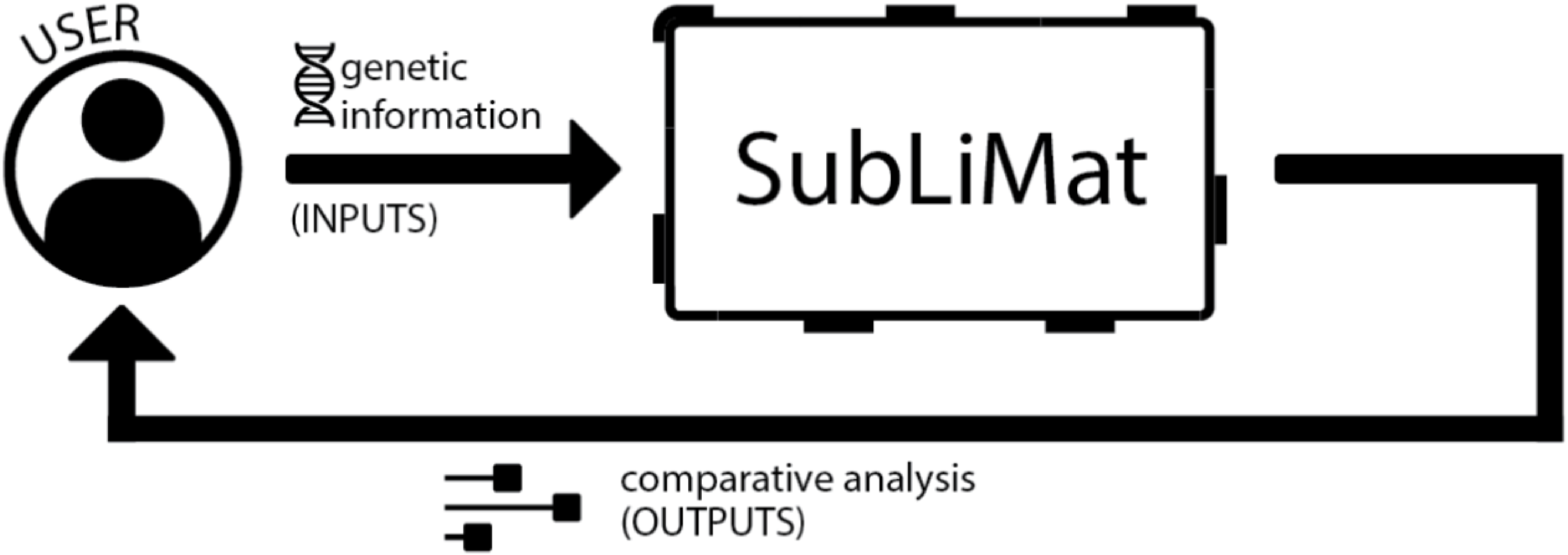
PROBLEM

Substitution matrices are critical assumptions that greatly impact studies in the area of comparative biology, yet, benchmarking these matrices is a laborious task.

GOALS (SRS)



GS1: Generate scores ranking the quality of the alignment between two genetic sequences given a benchmarked set of substitution matrices



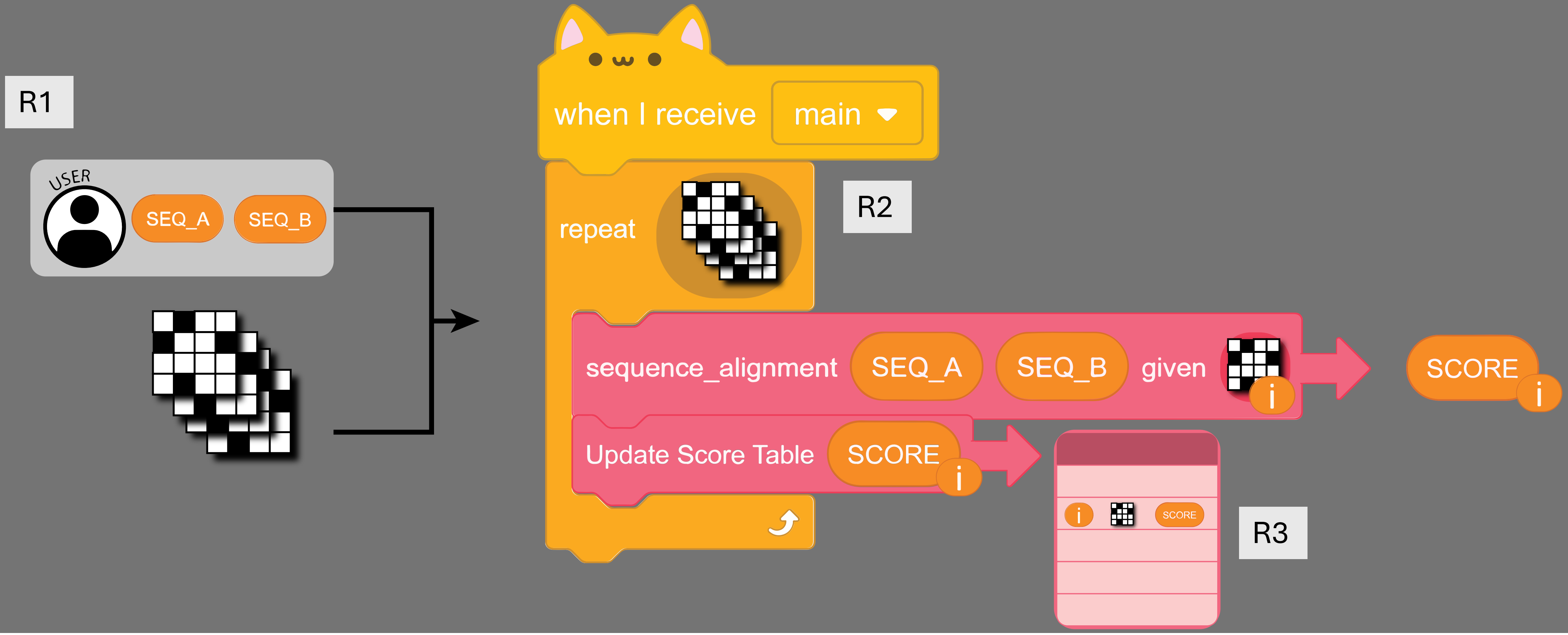
USER

genetic information

(INPUTS)

SubLiMat

comparative analysis
(OUTPUTS)

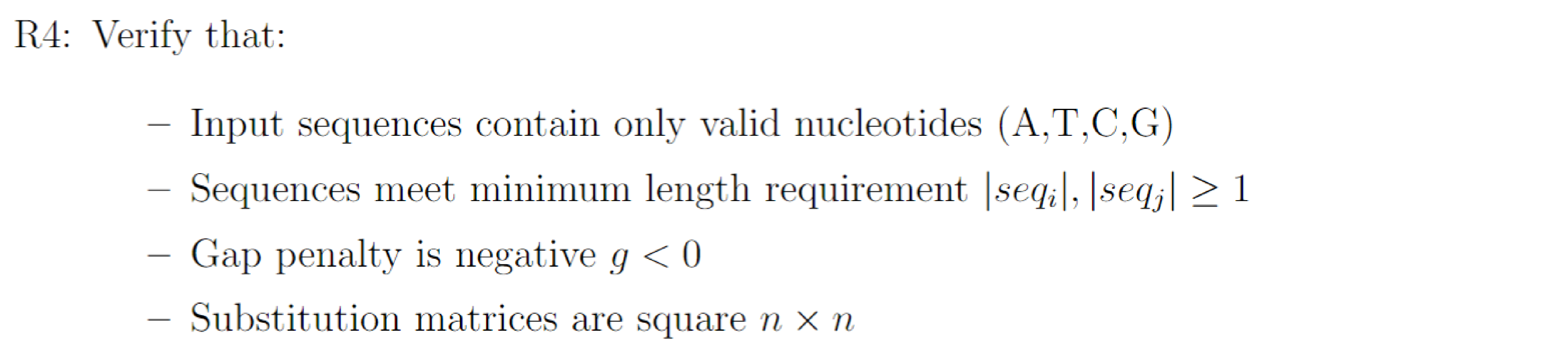


R1: Input SEQ_A, SEQ_B as strings of base pair units (bp), substitution matrix $S \in \mathbb{R}^{n \times n}$, and gap penalty $g \in \mathbb{R}_{<0}$.

R2: Use the inputs stated in IM1 to build a comparative matrix F^k for each substitution matrix S_k in $\mathbb{S}$.

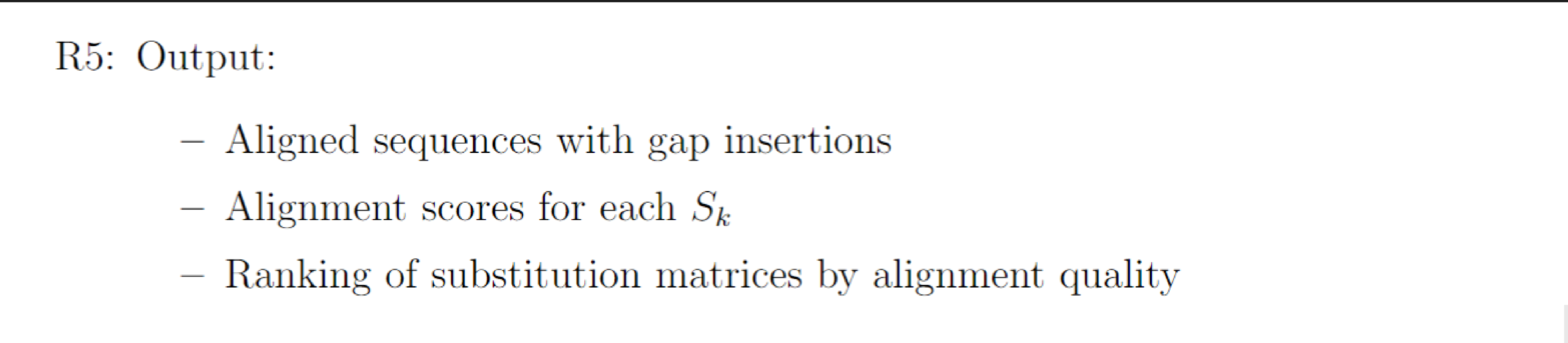
R3: Calculate optimal alignment scores IM1.

Looking for class feedback...



R4: Verify that:

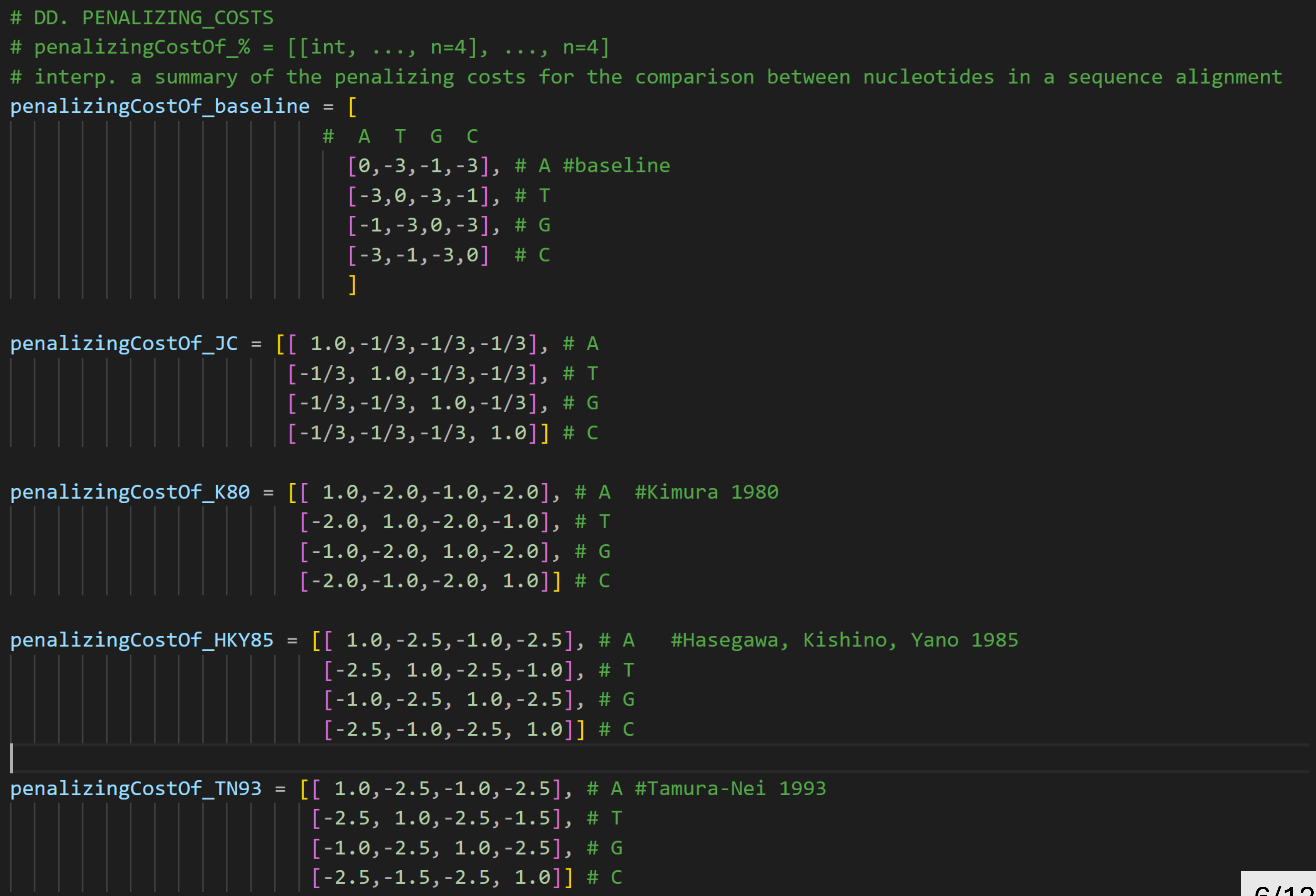
- Input sequences contain only valid nucleotides (A,T,C,G)
- Sequences meet minimum length requirement $|seq_i|, |seq_j| \geq 1$
- Gap penalty is negative $g < 0$
- Substitution matrices are square $n \times n$



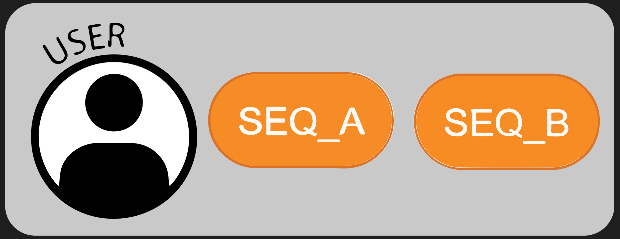
R5: Output:

- Aligned sequences with gap insertions
- Alignment scores for each S_k
- Ranking of substitution matrices by alignment quality

INPUTS

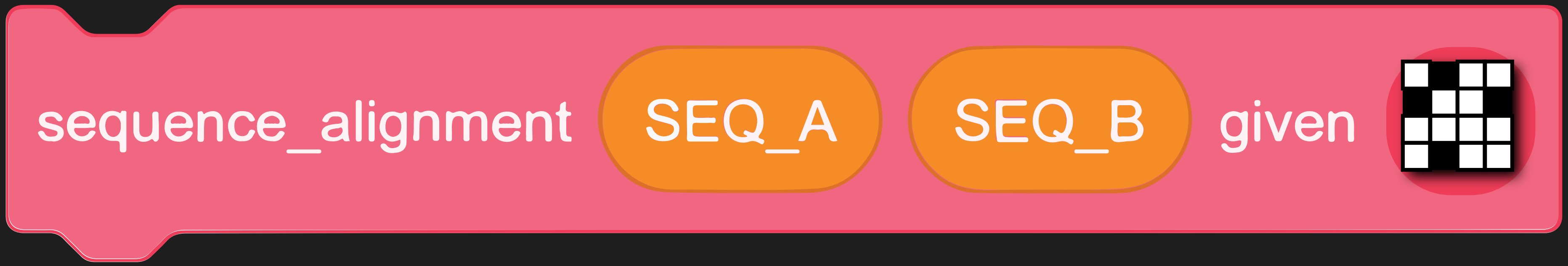


```
# DD. PENALIZING_COSTS
# penalizingCostOf_% = [[int, ..., n=4], ..., n=4]
# interp. a summary of the penalizing costs for the comparison between nucleotides in a sequence alignment
penalizingCostOf_baseline = [
                          #  A  T  G  C
                            [0,-3,-1,-3], # A #baseline
                            [-3,0,-3,-1], # T
                            [-1,-3,0,-3], # G
                            [-3,-1,-3,0]  # C
                            ]

penalizingCostOf_JC = [[ 1.0,-1/3,-1/3,-1/3], # A
                       [-1/3, 1.0,-1/3,-1/3], # T
                       [-1/3,-1/3, 1.0,-1/3], # G
                       [-1/3,-1/3,-1/3, 1.0]] # C

penalizingCostOf_K80 = [[ 1.0,-2.0,-1.0,-2.0], # A  #Kimura 1980
                        [-2.0, 1.0,-2.0,-1.0], # T
                        [-1.0,-2.0, 1.0,-2.0], # G
                        [-2.0,-1.0,-2.0, 1.0]] # C

penalizingCostOf_HKY85 = [[ 1.0,-2.5,-1.0,-2.5], # A   #Hasegawa, Kishino, Yano 1985
                          [-2.5, 1.0,-2.5,-1.0], # T
                          [-1.0,-2.5, 1.0,-2.5], # G
                          [-2.5,-1.0,-2.5, 1.0]] # C

penalizingCostOf_TN93 = [[ 1.0,-2.5,-1.0,-2.5], # A #Tamura-Nei 1993
                         [-2.5, 1.0,-2.5,-1.5], # T
                         [-1.0,-2.5, 1.0,-2.5], # G
                         [-2.5,-1.5,-2.5, 1.0]] # C
```

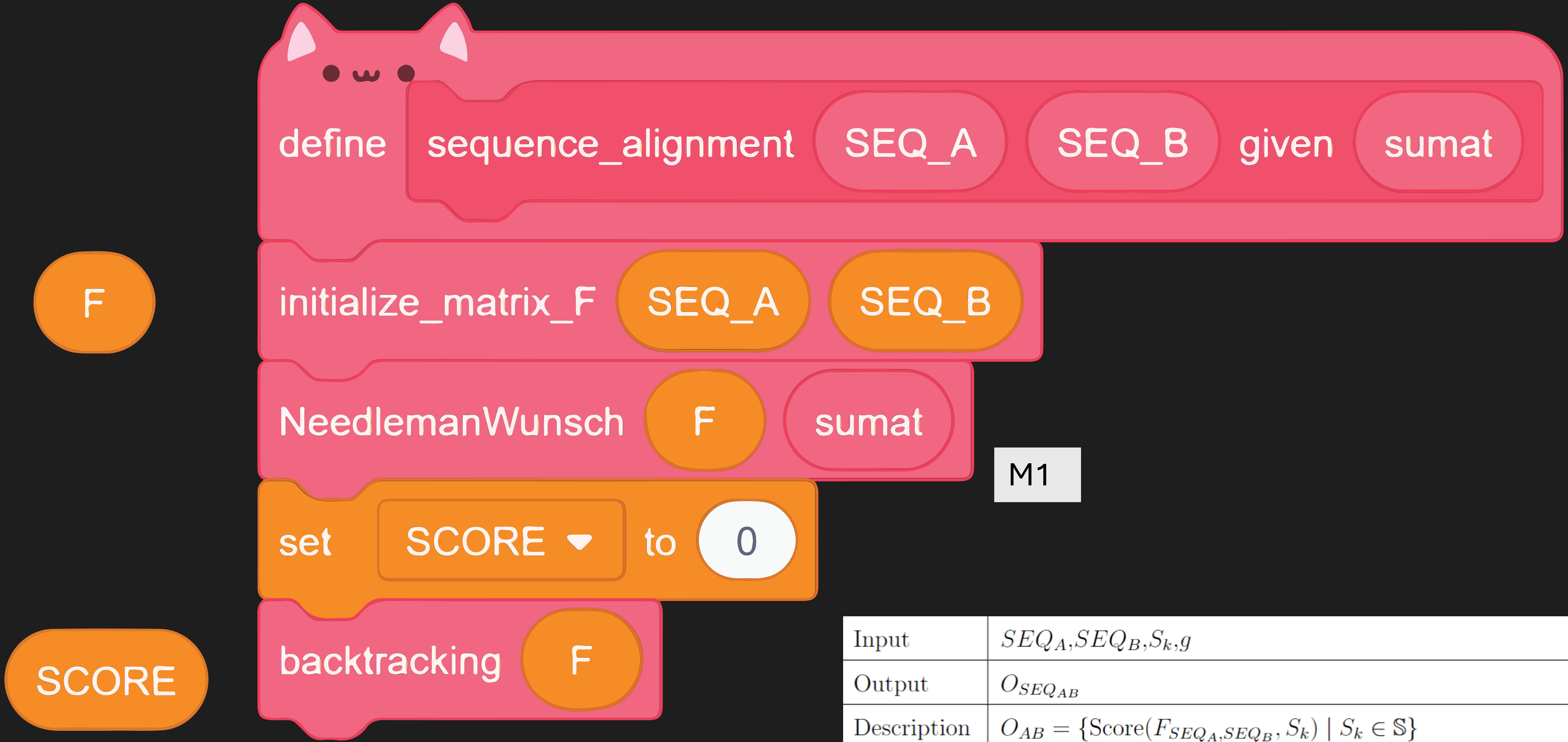


```
# DD. SEQUENCE
# seq = str
# interp. a string of characters representing nucleotides
seq0 = "TCCATCACCCTGGGCTGGCGGCGTGTGGCTATGGGGACGCTGGGCAGGGCTGGCCAGGAGGATGGCTGAGACACTGGAGTCCCAGCAGGCACGCGTCACCCCTGGCACATCCCCAGGCAGTGGGACTCCCTGTCCCCAGTT
seq1 = "GAGCAACACCACGGCCGGGGCCGGCGGCCCCTGGTGCCAGGGGCTCAACATCCCCAACGAGCTCTTCCTCACGCTGGGGCTGGTGAGCCTGGTGGAGAACCTGCTGGTGGTGGCTGCCATCCTGAAGAACAGGAACCTGCA
```

MODULES

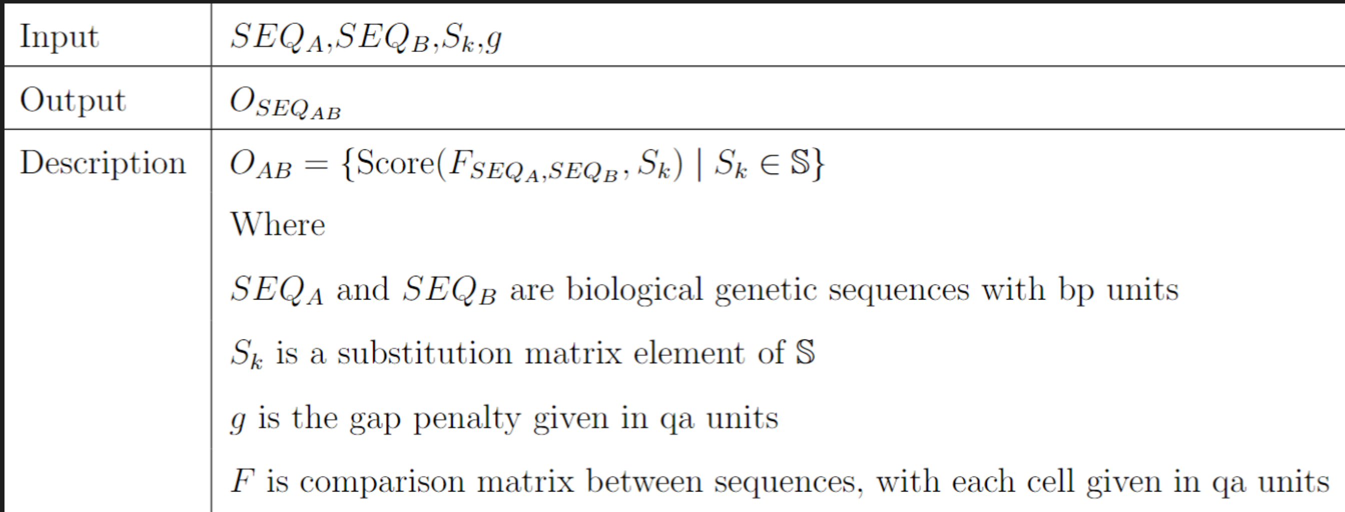


sequence_alignment SEQ_A SEQ_B given

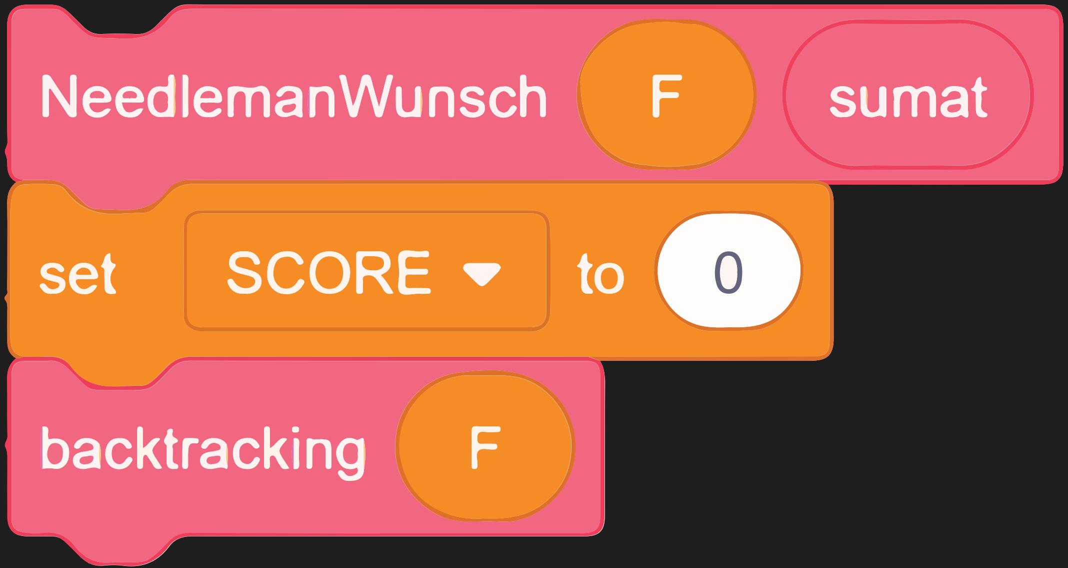


```
define sequence_alignment (SEQ_A) (SEQ_B) given (sumat)
initialize_matrix_F (SEQ_A) (SEQ_B)
NeedlemanWunsch (F) (sumat)
set [SCORE ▾] to (0)
backtracking (F)
```

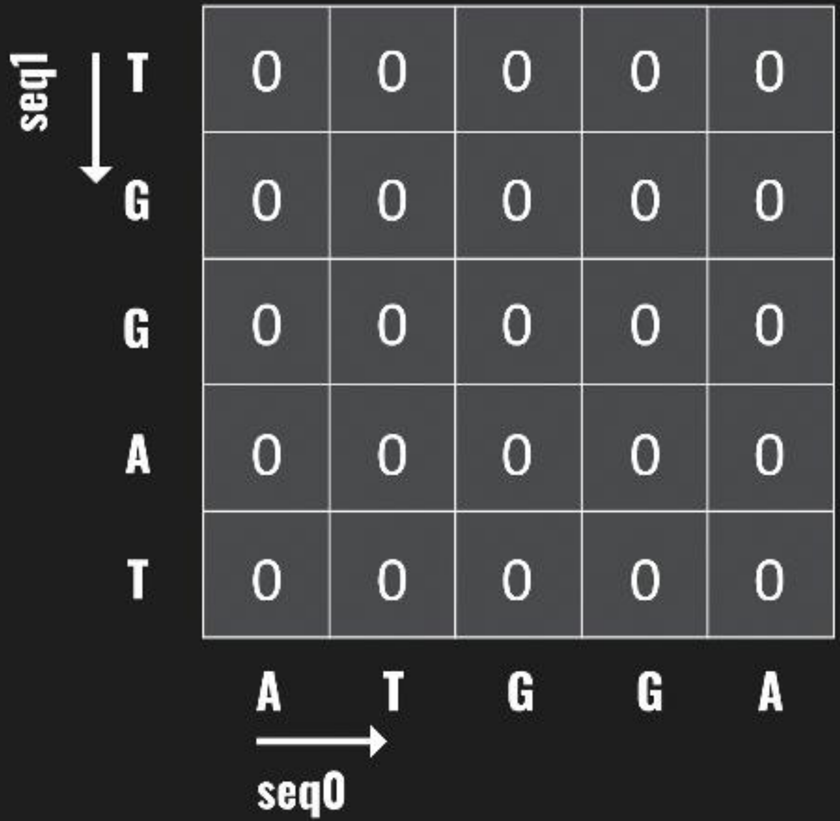
M1



Input	SEQ_A,SEQ_B,S_k,g
Output	$O_{SEQ_{AB}}$
Description	$O_{AB} = \{\text{Score}(F_{SEQ_A,SEQ_B}, S_k) \mid S_k \in \mathbb{S}\}$ Where SEQ_A and SEQ_B are biological genetic sequences with bp units S_k is a substitution matrix element of $\mathbb{S}$ g is the gap penalty given in qa units F is comparison matrix between sequences, with each cell given in qa units



```
NeedlemanWunsch F sumat
set SCORE ▾ to 0
backtracking F
```

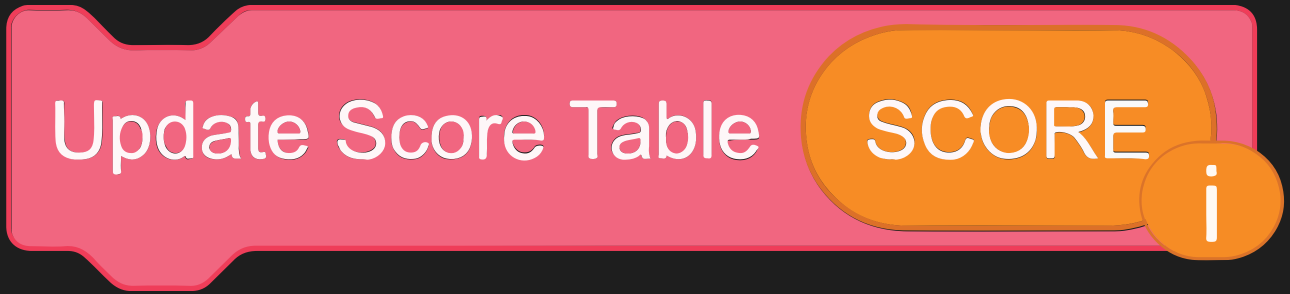


	A	T	G	G	A
T	0	0	0	0	0
G	0	0	0	0	0
G	0	0	0	0	0
A	0	0	0	0	0
T	0	0	0	0	0

```
 A  T  G  C
[0,-3,-1,-3], # A
[-3,0,-3,-1], # T
[-1,-3,0,-3], # G
[-3,-1,-3,0]  # C
"GAP":2
```

TGGAT

ATGGA



	matrix	score
1	JC	1.67
3	HKY85	-3.00
4	TN93	-3.00
2	K80	-4.00
0	baseline	-12.00