Individual Project -- Multiple Sequence Alignment (MSA) Problems

MSA Problem

$\{S^1, S^2, ..., S^k\}$ is a set of sequences over the same alphabet. The goal of MSA is to find an alignment for this set that minimizes some cost functions. As for the pairwise alignment, MSA implies a pairwise alignment between every pair of sequences.

Given cost matrix as below:

	MATCH	MISMATCH	GAP
COST	0	5	3

*Please note that **GAP-GAP** alignment is considered as a **MATCH** with cost 0. But you should avoid this condition to achieve shorter alignment.*

For two-sequence alignment with sequences of $ABCDE$ and $BBCD$, the best alignment is:

$$\begin{matrix} ABCDE \\ BBCD- \end{matrix}$$

and the cost of this alignment is:

$$5 + 0 + 0 + 0 + 3 = 8$$

For multiple sequence alignment with size larger than 2, the cost is computed using sum-of-pairs (sum up all pairwise cost).
Taking three-sequence alignment with sequences of $ABCD$, ACD and BCD as an example, the best alignment is:

$$\begin{matrix} ABCD \\ A-CD \\ -BCD \end{matrix}$$

and the cost of this alignment is:

$$(0+3+3)+(3+0+3)+(0+0+0)+(0+0+0)=12$$

Requirement

There are 5 queries for two-sequence alignment and 2 queries for three-sequence alignment in `MSA_query.txt`. Database is stored in `MSA_database.txt` with 100 sequences.

Using cost matrix shown above, you need to:

- Implement A-star (A*) algorithm to find the **optimal** solution.
    - Can you construct **another heuristic function** for your algorithm? Please detail these two methods and compare them.
- Implement genetic algorithm to find the **optimal/suboptimal** solution.
    - How do you denote an **individual**? How do you perform **crossover** and **mutation**? Can you improve your algorithm for better solution?
- Please describe and analyze your implementation, results(alignments and costs must be included), running time and time complexity thoroughly in the report.
- Please submit a file in ZIP or RAR format containing your report (Chinese or English) and codes on Canvas before **2023-10-11, 23 : 59**. Name it as **StudentID_Name**.
- **Do not cheat**.