CS3317: Artificial Intelligence

Shanghai Jiao Tong University - 2022 Fall

Project 1: Multiple Sequence Alignment

Introduction

Multiple sequence alignment (MSA) may refer to the process or the result of sequence alignment of three or more biological sequences, generally protein, DNA, or RNA. MSA can reveal the potential information in biological sequences, such as function, evolution and structure. Here we also consider pair-wise alignment. An example is shown below.

$$\left\{\begin{array}{r} HEAGAWGHEE \\ HDACAWGHEE \\ HDACWGHEE \\ HDCSTGHEE \\ PAWHEAE \end{array}\right. \longrightarrow \left\{\begin{array}{r} H\,E\,AG\,A\,WGHE-E \\ H\,D\,A\,C\,A\,WGHE-E \\ HD\,A\,C-WGHE-E \\ HD-C\,S\,T\,GHE-E \\ --P-A\,WGHE\,A\,E \end{array}\right.$$

Requirements

- Implement dynamic programming (DP) algorithm to find the optimal solution.
- Implement A-star (A*) algorithm to find the optimal solution.
- Implement genetic algorithm to find the optimal/suboptimal solution.

For each query, you need to find the best alignment (another one or two sequences) that yields the least cost. Cost matrix is revealed in Table 1.

Table 1: Cost matrix.

	MATCH	**MISMATCH**	**GAP**
COST	0	4	3

For multiple sequence alignment with size larger than 2, the cost is computed using sum-of-pairs (sum up all pairwise cost). Note that **GAP-GAP** alignment is considered as a **MATCH** with cost 0. For example, given one query $ABCD$, the cost of the alignment

$$\begin{cases} ABCD \\ A-CD \\ B-CD \end{cases}$$

is $(0+4+4)+(3+3+0)+(0+0+0)+(0+0+0)=14$.
There are **5** queries for pairwise alignment and **2** queries for three-sequence alignment in $MSA\_query.txt$. Database is stored in $MSA\_database.txt$ with 100 sequences.

Submission

1. Please submit a file in *ZIP* or *RAR* format containing your **report** (Chinese or English) and **codes** on Canvas before **2022-10-07**, **23** : **59**. Name it as *StudentID_Name* (eg. 5180xxxxxxxx_张三).

2. Describe and analyze your implementation, results (alignment and cost), running time and time complexity (O, Ω, etc.) thoroughly in the report.

3. Do not cheat. Please construct DP and A* algorithms from scratch. If you need to utilize existing libraries or codes for GA algorithm, please state the source and your own understanding in the report.