2460 - Searching Sequence Database in Molecular Biology

Asia - Singapore - 2001/2002

Molecular biologists frequently compare bio-sequences to see if any similarities can be found in the hope that what is true of one sequence is also true of its analogue. In this problem, we focus on nucleic acid sequences that are composed of four symbols `A', `C', `G' or `T'. Generally, such comparisons involve aligning sections of the two sequences in a way that exposes the similarities between them. Given a query sequence and a set of sequences stored in a database, you are asked to write a program that searches the database and finds the sequence having the largest similarity score with the query sequence.

The similarity score between the query sequence and a database sequence is the sum of the alignment scores of the aligned pairs of symbols from an alignment of the two sequences. Two identical symbols that are aligned are given a score of +5 while a mismatched pair of symbols is assigned a score of -4. A gap is introduced into an alignment if one symbol in one sequence is not aligned with symbols in the other. The penalty for a gap is a score of -7. For example, given a query sequence *m*=`GAAGGCA' and a database sequence *n*=`GCAGAGCA', the following alignment between them (aligned pairs of symbols are written one above the other) has a similarity score of 5+(-4)+5+5+(-7)+5+5+5=21.

```
    Sequence m: G A A G - G C A

Sequence n:             G C A G A G C A
```

Note that a gap in the above alignment is represented by the symbol `-'. The dynamic programming algorithm provides a rigorous mathematical approach towards this alignment problem.

Input

The input consists a query sequence (on the first two lines) and a set of data sequences (each occupying two lines). Each sequence has the following format:

```
> sequence name
 sequence data
```

There is a blank line between two adjacent sequences.

Output

The output indicates the sequence having the highest similarity score with the query sequence. The result should be printed out in the following format:

```
The query sequence is:
```
(the query sequence data)

```
The most similar sequences are:
```

(sequence 1 data)
`The similarity score is: (`*`similarity score`*`)`

(sequence 2 data)
`The similarity score is: (`*`similarity score`*`)`

Sample Input

```
>query
ACGGG

>seq1
ACGGT

>seq2
ACGGGG

>seq3
TCCGGTT

>seq4
TCGGG

>seq5
AACGGG
```

Sample Output

```
The query sequence is:
ACGGG

The most similar sequences are:

ACGGGG
The similarity score is: 18

AACGGG
The similarity score is: 18
```

Singapore 2001-2002