CSE3800/BME4800/CSE5800: BIOINFORMATICS
Programming Assignment #2

Due Friday, Oct. 7, 2011

Multiple sequence alignment is a critical tool for extracting and representing biologically important, yet faint commonalities between large numbers of sequences. Such commonalities may reveal distant evolutionary history and critical conserved motifs, providing clues about the biological function of the strings. Although exact solutions for multiple sequence alignment can be obtained by dynamic programming, such algorithms are not practical since their running time is exponential in the number of sequences. For the sum-of-pairs (SP) objective with a scoring function that satisfies the triangle inequality, the *center star algorithm* is guaranteed to find a multiple sequence alignment whose SP score is no more than twice larger than the optimum.

Project description

Implement the center star algorithm for minimizing the sum-of-pairs objective under edit distance scores (i.e., cost of 1 for indels and mismatches, and cost of 0 for matches, including matches between 2 spaces). As described in class (see also section 14.6.2 in the attached excerpt from Gusfield's book on "Algorithms on Strings, Trees, and Sequences"), the algorithm has two main steps. Given n strings $s_1,\ldots,s_n$, in the first step one must find the index $c\in\{1,\ldots,n\}$ such that the star "cost" $\sum_{i\neq c} edit\_dist(s_c, s_i)$ is minimized. In the second step, a multiple alignment is constructed by successively aligning each string s_i, $i\neq c$, to s_c, using the modified edit distance algorithm described in class (see also the proof of Theorem 14.6.1 in Gusfield's chapter).

Input

Your program should read from the standard input an integer n followed by n lines, each containing a string of at most 1000 upper case letters.

Output

Using a format similar to that in the example below, print to the standard output the minimum star cost identified in the first step of the algorithm, the sum-of-pairs score of the final alignment, and the final alignment itself, using dashes to represent alignment gaps.

Sample input

```
4
AXZ
AXXZ
AYXYZ
AYZ
```

Sample output

```
min star cost: 4
total cost: 11
A-X-Z
AXX-Z
AYXYZ
A-Y-Z
```