CSci 421
Introduction to Algorithms

Homework Assignment 4
Due: Wednesday, July 25, 2007

Summer 2007

W. L. Ruzzo July 21, 2007

Homework:

1. Text, Chapter 6, #1, page 312. Note that for this problem, the next, and #4, the question is to find the optimal *solution*, not just its cost (i.e., include "trace-back"). Also, as always, "given an algorithm" means algorithm, correctness argument and run time analysis.

2. Text, Chapter 6, #6, page 317.

3. Run the string alignment algorithm (section 6.6) on strings $S = tcatag$ and $T = tataag$. Build the cost matrix and traceback pointers as in the example given in lecture. Assume that aligning two identical letters gives a score of $+2$, whereas aligning a letter with a mismatched letter or a gap ("–") gives a score of -1.

4. You are given a binary tree with n leaves, with the ith leaf labeled by a letter S_i from a fixed alphabet Σ. You are also given a function c, which assigns a cost $c(S, S') \geq 0$ to each pair of letters S, S' in Σ. Assume $c(S, S') = c(S', S)$ and $c(S, S) = 0$ for all $S, S' \in \Sigma$. The problem is to give each internal node x of the tree a label $S_x \in \Sigma$ so as to minimize the total over all tree edges of the cost of that edge, where the cost of an edge whose end points are labeled U and V is $c(U, V)$. Give an efficient algorithm to solve this problem, explain why it is correct, and analyze its running time as a function of n and $|\Sigma|$. Assume that each evaluation of the cost function c costs $O(1)$ operations; e.g., via table lookup using a table built in to your algorithm. Hint: use dynamic programming. As we've seen before, you might need a stronger induction hypothesis; i.e., you'll calculate more at each internal node than just whether to put letter "S" there. For instance, it might help to know the total cost of the subtree rooted there, assuming that it's labeled "S".

   [Although it doesn't matter for purposes of this homework, this algorithm is sometimes used in estimating evolutionary trees. The S_i are letters at corresponding positions in DNA or protein sequences from n different modern species, the tree is their presumed evolutionary tree, and the goal is to try to reconstruct the corresponding letters in the ancestral sequences. For example, with the cost function $c(S, S') = (\text{ if } S == S' \text{ then } 0 \text{ else } 1)$, the min cost is simply counting the minimum number of "mutation" events in the evolutionary tree necessary to explain the observed variability in the modern species with respect to their presumed common ancestors.]