CS 325 Report 2

Chunxiao Wang

1 Pseudo-code

Input: sequence1(seq1) with length m, sequence2 (seq2) with length n, costmatrix (costm)
Output: minimum edit distance, aligned sequence1, aligned sequence2
Base condition:

$$
\begin{aligned}
&D(0,0) = 0, \quad D(i,0) = D(i-1,0) + costm('-', seq1(i)),\\
&D(0,j) = D(0,j-1) + costm(seq2(j), '-')\\
&ptr(0,0) = (0,0,0), \quad ptr(i,0) = (1,0,0), \quad ptr(0,j) = (0,1,0),
\end{aligned}
$$

for $ptr(i,j)$, the first position is for deletion, the second position for insertion and the third position for substitution.

Recurrence relation:
For i from 1 to m
For j from 1 to n
$del = D(i-1,j) + costm(seq1(i), '-')$, $ins = D(i,j-1) + costm('-', seq2(j))$, $subs = D(i-1,j-1) + costm(seq1(i), seq2(j))$
$D(i,j) = min(del, ins, subs)$, $ptr(i,j) = (del == 1, ins == 1, subs == 1)$.

Traceback:
From ptr matrix, figure out the backtrace and edit distance operations. Then align seq1 and seq2 based on the backtrace and the edit distance operations.

Return:
$D(m,n)$, aligned seq1, aligned seq2

2 Asymptotic analysis of run time

To build the D matrix, we have to go through m base of sequence1 and inside each, go through n base of sequence2, so time complexity is $O(mn)$, for saving space, the matrix is $m \times n$, so space is $O(mn)$, for backtrace, the time complexity is $O(m+n)$. So the asymptotic running time should be $max(O(mn), O(m+n))$, usually it is $O(mn)$.

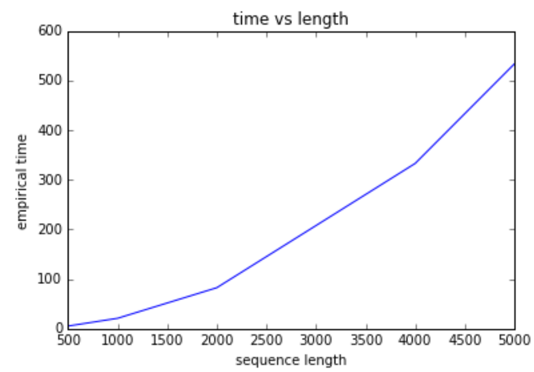
3 Runtime

To get the running time, I use the default time function in python, and store running time for each implementation, and then compute the average of the 10 running time. I used a mac pro with 2.6 GHz and Core i5 processor. I include all the parts in running time measurements including output of alignment. The empirical running time is

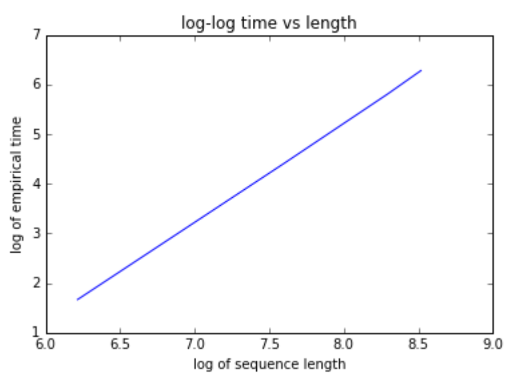
Table 1: Runtime Table

length	500	1000	2000	4000	5000
average runtime(s)	5.2957	20.9122	82.5379	332.9839	533.5867

The original running time plot is



The log-log scale running time plot is



Using $\frac{y_i - y_j}{x_i - x_j}$, I get a slope of 1.98. So the line is of the form $O(x^1.98)$ in the linear plot, which corresponds to the asymptotic running time $O(mn)$.

4 Interpretation and discussion

Because there are only 5 cases, the original plot is not exactly $y = x^2$. But from the log-log running time plot and calculation, the power 1.98 is very closed to 2. So the growth curve of empirical running time does match with the asymptotic running time $O(x^2)$.