Abstract

Longest common subsequence (LCS) is a problem of finding longest subsequence common to the given sequences. Motivated by applications in genomics and computational molecular biology, different variants of LCS were introduced for more accurate measure of similarity. This project studies following variants of LCS: Variable gapped LCS, k-substring LCS and Constrained LCS: multiple exclusion constraints.

Motivation

Information in various fields such as bioinformatics is expressed as sequence of characters. LCS is a similarity measure of character sequences. LCS is used in evidence search for species common origin, gene discovery and evolutionary tree construction.

- M-STR-EC-LCS: LCS with exclusion of multiple given substring constraints [1]. It can be used in homology computation of biological sequences which requires specific or putative structure (constraints) on the LCS.

- LCS-k: LCS with maximal k-length substrings [2]. In order to improve the accuracy of similarity measure, LCS-k was introduced by ensuring adjacency of the characters picked for LCS.

- VGLCS: LCS with variable gap constraints [3]. It can be used in motif pattern discovery when all the positions of desired pattern are not important and in long multiple repeats extraction in DNA sequences.

Algorithms

M-STR-EC-LCS

\[ f[i,j,k] = \begin{cases} \max\{f[i+1,j+1,k], f[j+1,i+1,0]\} & \text{if } \tau_i = \tau_j \text{ and } q < k; \\ \max\{f[i+1,j,k], f[j+1,i,k]\} & \text{otherwise}; \\ \end{cases} \]

VGLCS

\[ Gaps_{A}(i) = \begin{cases} 2 & \text{if } i = 0 \text{ or } j = 0; \\ 1 & \text{if } A[i] \neq B[j]; \\ 0 & \text{otherwise}; \\ \end{cases} \]

VGLCS dynamic approach takes \( O(nm^2) \). Incremental maximum suffix query (ISMQ) technique is used to find maximum of all possible gap options to improve the running time to \( O(nm) \). Union-find problem is used to arrange the elements of ISMQ to find the maximum element in constant time and VGLCS is computed in \( O(nm) \) time.

Evaluations

For the experiment, M-STR-EC-LCS and LCS-k are implemented on random strings of length 5 to 5000 over alphabets of size 4 and compared with the LCS dynamic solution. Measured M-STR-EC-LCS with different values of \(|P|\) (number of constraints) and LCS-k with different values of \(k\) (length of substring).

References