An FPT-Algorithm for Longest Common Subsequence Parameterized by the Number of Deletions

Laurent Bulteau, Mark Jones, Rolf Niedermeier, Till Tantau

An FPT-Algorithm for Longest Common Subsequence Parameterized by the Number of Deletions

Laurent Bulteau[1]
LIGM, CNRS, Université Gustave Eiffel, F77454 Marne-la-vallée France

Mark Jones
Delft Institute of Applied Mathematics, Delft University of Technology, The Netherlands

Rolf Niedermeier
Algorithmics and Computational Complexity, Faculty IV, TU Berlin, Germany

Till Tantau
Institute of Theoretical Computer Science, University of Lübeck, Germany

Abstract

In the NP-hard Longest Common Subsequence problem (LCS), given a set of strings, the task is to find a string that can be obtained from every input string using as few deletions as possible. LCS is one of the most fundamental string problems with numerous applications in various areas, having gained a lot of attention in the algorithms and complexity research community. Significantly improving on an algorithm by Irving and Fraser [CPM'92], featured as a research challenge in a 2014 survey paper, we show that LCS is fixed-parameter tractable when parameterized by the maximum number of deletions per input string. Given the relatively moderate running time of our algorithm (linear time when the parameter is a constant) and small parameter values to be expected in several applications, we believe that our purely theoretical analysis could finally pave the way to a new, exact and practically useful algorithm for this notoriously hard string problem.

Keywords NP-hard string problems, multiple sequence alignment, parameterized complexity, search tree algorithms, enumerative algorithms

2012 ACM Subject Classification Theory of computation → Fixed parameter tractability

Keywords and phrases NP-hard string problems, multiple sequence alignment, parameterized complexity, search tree algorithms, enumerative algorithms

Funding *Mark Jones*: Supported by Netherlands Organization for Scientific Research (NWO) through grants NETWORKS and OCENW.KLEIN.125.
Rolf Niedermeier: Supported by the DFG, NI 369/16, FPTinP.

Acknowledgements This work was initiated during Dagstuhl Seminar 19443, *Algorithms and Complexity in Phylogenetics*, held at Schloss Dagstuhl, Germany, in October 2019.

1 Introduction

With its numerous applications in bioinformatics, data compression, computational linguistics, etc. the NP-hard Longest Common Subsequence (LCS) problem is among the best studied algorithmic string problems. Suiting our parameterized analysis purposes, we formally define the problem as follows.

Longest Common Subsequence
Input: A set of k strings $\mathcal{S} = \{S_1, \ldots, S_k\}$, each of length at most n, an integer ℓ.
Parameter: $\Delta = n - \ell$.

[1] Corresponding author

Question: Is there a string S of length at least ℓ that is a (not necessarily contiguous) subsequence of each S_i?

With standard dynamic programming, LCS can be solved in $O(n^k)$ time; on the contrary, it is known to be W[1]-hard [12] respectively W[2]-hard for parameter k and it has no $O(n^{k-\epsilon})$ algorithm under SETH [1]. Indeed, LCS is *the* string problem receiving most attention when the field of parameterized complexity [8] started. Unfortunately, so far parameterized complexity analysis beyond trivial algorithmic observations mainly contributed computational hardness results. We refer to some surveys [2, 6, 7] for an overview on research results (and open questions) for LCS.

We remark that the special case of two input strings (that is, $k = 2$) recently attracted much attention, particularly motivated by the theoretical challenge of breaking the straightforward time bound of $O(n^2)$ [3, 5, 9]. Notably, Bringmann and Künnemann [5] (the corresponding arXiv paper has around 60 pages) also discuss the "maximum number of deletions" parameter we focus on here. Indirectly, this parameter already appears in the work of Irving and Fraser [10], who provided two algorithms for LCS with three or more input strings.

Irving and Fraser [10] in their 1992 paper provided an algorithm for LCS running in time $O(kn(n-\ell)^{k-1}$, implying fixed-parameter tractability with respect to the combined parameter k and $n-\ell$, where the latter coincides with our parameter Δ. We are not aware of any improvement since then and this is also reflected by a corresponding challenge featured in a 2014 survey [6, Challenge 9]. Answering positively the research challenge posed there, we improve Irving and Fraser's result to fixed-parameter tractability only with respect to Δ. More specifically, our algorithm runs in time $O((\Delta+1)^{\Delta+1}kn)$, which means linear time when Δ is a constant. In addition, we can enumerate *all* longest common subsequences within this time. Given that it seems natural to assume that in many applications the seeked common subsequence is fairly close to every input string (which would mean small Δ), this promises to be of also practical relevance. However, the focus of this work is purely theoretical. We mention in passing that our result holds for arbitrary alphabet size.

Figure 1, at a very high level, for three input strings presents an example for LCS and the main idea behind our recursive approach towards achieving our result, the FPT-algorithm for parameter Δ.

2 LCS Algorithm Using Maximal Common Subsequences

In this section, we present a linear-time algorithm for LCS when the number of deletions is a constant. Note that it is not incompatible with the quadratic lower bound for this problem, since this lower bound only applies to the general case where the number of deletions is unbounded. In particular, the $O(\delta n)$ algorithm by Nakatsu et al. [11] (with $\delta = \min\{|S_i|\} - \ell$) remains better than our algorithm for the two-string case. Furthermore, it is not clear if a smaller (typically constant) alphabet could be exploited in the algorithm or its analysis to obtain a better running time.

2.1 Definitions

Strings

The set of strings on an alphabet Σ is denoted Σ^*. The empty string is ϵ, $|S|$ denotes the length of S. We write $\cdot$ for the concatenation. We write $u \cdot T := S$ as a short-hand for "let u

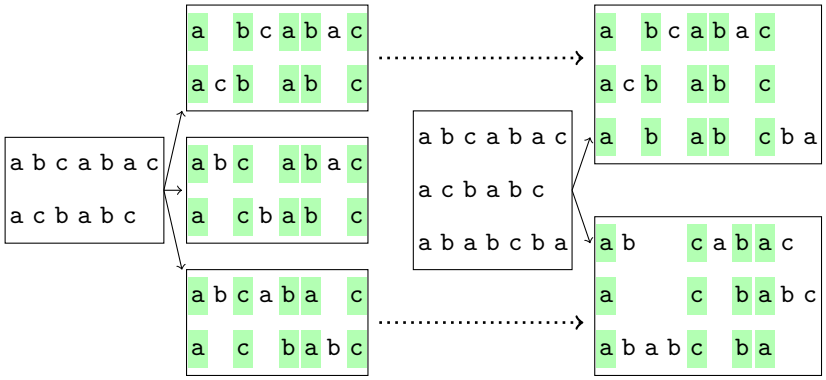

Figure 1 Our approach towards computing the LCS of three strings `abcabac`, `acbabc`, `ababcba`. Left: compute maximal common subsequences of the first two strings (all three subsequences and their alignment with input strings are depicted). Right: compute maximal common subsequences of all three strings by comparing those obtained at the first step with the third input string (only two strings remain after filtering non-maximal common subsequences). The longest result, `ababc` is the LCS of the input strings. Filtering out strings that are shorter than a threshold prevents the number of intermediate strings from growing too fast, yielding our FPT-algorithm.

be the first character of S and T be the suffix of S starting from the second character (or $u = T = \epsilon$ if S is empty)".

Given two strings S_1, S_2, we write $S_1 \preceq S_2$ (resp. $S_1 \prec S_2$) if S_1 is a (strict) subsequence of S_2 (formally, $\epsilon \preceq S$ for any S and, if $S \preceq S'$, then for any u, $S \preceq u \cdot S \preceq u \cdot S'$).

Longest and Maximal Common Subsequences

Given $\mathcal{S}$ and ℓ, let $\mathrm{CS}_\ell(\mathcal{S})$ denote the set of all common subsequences of $\mathcal{S}$ that have length at least ℓ. Let L be the largest integer such that CS_L is not empty, and let $\mathrm{LCS}(\mathcal{S})$ denote an arbitrary string in CS_L, i.e. a longest common subsequence of $\mathcal{S}$.

Let $\mathrm{MCS}_\ell(\mathcal{S})$ denote the set of all *maximal* common subsequences of $\mathcal{S}$ with length at least ℓ; that is, $S \in \mathrm{MCS}_\ell(\mathcal{S})$ iff $S \in \mathrm{CS}_\ell(\mathcal{S})$ and there is no $S' \in CS_\ell(\mathcal{S})$ such that $S \prec S'$. Note that, if ℓ is small enough ($\ell \leq L$), then $\mathrm{LCS}(\mathcal{S}) \in \mathrm{MCS}_\ell(\mathcal{S})$, otherwise $\mathrm{MCS}_\ell(\mathcal{S})$ is empty. A set of strings M is an *extended MCS* of $(\mathcal{S}, \ell)$ if $\mathrm{MCS}(\mathcal{S}, \ell) \subseteq M \subseteq \mathrm{CS}(\mathcal{S}, \ell)$.

String parameters

Let $\mathcal{S} = \{S_1, \ldots, S_k\}$ be a set of strings. We write $n(\mathcal{S}) = \max_{S \in \mathcal{S}} |S|$, $m(\mathcal{S}) = \min_{S \in \mathcal{S}} |S|$. Given an integer ℓ, we write $\Delta(\mathcal{S}, \ell) = n(\mathcal{S}) - \ell$ and $\delta(\mathcal{S}, \ell) = m(\mathcal{S}) - \ell$. We omit dependencies on $\mathcal{S}$ and ℓ when the context is clear (e.g., they are given in the lemma statement). Note that $\delta \leq \Delta$.

2.2 Main results

▶ **Theorem 1.** *Let $\mathcal{S} = \{S_1, \ldots, S_k\}$ be a set of strings and ℓ be an integer. Then an extended MCS of $(\mathcal{S}, \ell)$ with size at most $(\Delta + 1)^\delta$ can be computed in time $O(2^{\delta+\Delta}(\Delta + 1)^\delta kn)$.*

Theorem 1 directly yields an algorithm for LCS, since it suffices to test if an extended MCS of $(\mathcal{S}, \ell)$ is non-empty. Note that the algorithm can be adapted for the optimization formulation of LCS, i.e., when ℓ is not part of the input, with a constant factor in the time complexity (taking δ and Δ with respect to $\ell = |\mathrm{LCS}(\mathcal{S})|$). Indeed, apply Theorem 1 for

decreasing values of ℓ starting with $\ell = m$, until a non-empty set is obtained. Then, the resulting set contains the common substrings of $\mathcal{S}$ of size $\mathrm{LCS}(\mathcal{S})$ (indeed, $\mathrm{MCS}_\ell(\mathcal{S}) = \mathrm{CS}_\ell(\mathcal{S})$ for this value of ℓ), so it contains all longest common subsequences of $\mathcal{S}$. The time complexity of the i-th call, $1 \leq i \leq \delta$, is upper-bounded by $O(2^{i+\Delta}(\Delta+1)^\delta kn)$. Using $\sum_{i=1}^{\delta} 2^i = O(2^\delta)$, we get the following corollary.

▶ **Corollary 2.** *All longest common subsequcences of $\mathcal{S}$ (and a fortiori the value $\mathrm{LCS}(\mathcal{S})$) can be computed in time $O(2^{\delta+\Delta}(\Delta+1)^\delta kn)$.*

The remainder of the section is dedicated to proving Theorem 1. We first compute the number of strings and their size distribution in the MCS of two strings, then build up on this result to bound the size of the MCS of k strings.

2.3 Extended MCS for two strings

Algorithm 1 allows us to compute an extended MCS of two strings. Its correctness is proven using the main recursive relation for MCS given in Lemma 3, while its time complexity is analyzed in Lemmas 6 and 8.

▶ **Lemma 3.** *For any two non-empty strings S, S' and any ℓ, let $u \cdot T := S$ and $u' \cdot T' := S'$.*
If $u = u'$, then $\mathrm{MCS}_\ell(\{S, S'\}) \subseteq \{u \cdot X \mid X \in \mathrm{MCS}_{\ell-1}(\{T, T'\})\}$.
If $u \neq u'$, then $\mathrm{MCS}_\ell(\{S, S'\}) \subseteq \mathrm{MCS}_\ell(\{S, T'\}) \cup \mathrm{MCS}_\ell(\{T, S'\})$.

Proof. Let $R \in \mathrm{MCS}_\ell(\{S, S'\})$, and $r \cdot X := R$.

For the first case ($u = u'$), we show that $r = u$ and X is a maximal common subsequence of $\{T, T'\}$ of length at least $\ell - 1$. Indeed, $r = u$, as otherwise the concatenation $u \cdot r \cdot X$ would also be a common subsequence of $\{S, S'\}$, with $R \prec u \cdot r \cdot X$ (contradicting the maximality of R). Note that X is a subsequence both of T and T'. Moreover X is maximal, as otherwise, if $X \prec X'$ with X' a common subsequence of T, T', then $u.X'$ would be a common subsequence of S, S' with $R \prec u.X'$ (again, contradicting the maximality of R).

For the second case ($u \neq u'$), we show that R is either in $\mathrm{MCS}_\ell(\{T, S'\})$, or in $\mathrm{MCS}_\ell(\{S, T'\})$ (or both). Indeed, if $r \neq u$, then $R \prec S$ implies $R \prec T$, and R is a common subsequence of $\{T, S'\}$. Otherwise, $r = u \neq u'$, and R is a common subsequence of $\{S, T'\}$. In both cases, R is maximal, since for any R', if R' is a common subsequence of (say) $\{T, S'\}$ with $R \prec R'$, then R' is also a common subsequence of $\{S, S'\}$ which contradicts the maximality of R for $\{S, S'\}$. ◀

Algorithm 1 follows the recursive relation of Lemma 3, along with trivial base cases ($\ell > \min\{|S|, |S'|\}$ or one of S, S' is a substring of the other). It also clearly returns only common substrings of S and S' of length at least ℓ, so it is correct.

▶ **Corollary 4.** *Let S, S' be two strings and ℓ be an integer. Then* `xMCS2`(ℓ, S, S') *returns an extended MCS of $(\{S, S'\}, \ell)$*

▶ Remark 5. The first inclusion in Lemma 3 (case $u = u'$) is actually an equality, but we only need this direction for the algorithm to be correct. The second inclusion however may be strict: for example with $S = ABCD$ and $S' = DABC$, string $R = BC$ is a maximal common subsequence of $T = BCD$ and S', but not of S, S' since $R \prec ABC$. Such "extra" strings are actually returned by our algorithm, motivating the naming of *extended* MCS (although they could be filtered out, see Remark 7).

We now focus on the time complexity of Algorithm 1

Algorithm 1 Compute a bounded-size extended MCS of two strings.

```
xMCS2(ℓ, S, S′):
  if ℓ > |S| or ℓ > |S′|:
    return {}
  if S (resp S′) is a substring of S′ (resp S):
    return {S} (resp {S′})
  u·T := S
  u′·T′ := S′
  if u = u′:
    return {u·X | X in xMCS2(ℓ-1, T, T′)}
  else:
    return xMCS2(ℓ, S, T′) ∪ xMCS2(ℓ, T, S′)
```

▶ **Lemma 6.** *Let S, S' be strings of length respectively $\ell+\delta$ and $\ell+\Delta = n$. Then* $\texttt{xMCS2}(\ell, S, S')$ *terminates in time $O(2^{\delta+\Delta}n)$.*

Proof. The time complexity is achieved using a precomputed substring table: for every pair i, i' with $1 \leq i \leq |S|$, $1 \leq i' \leq |S'|$, and $|i - i'| \leq \Delta$, $\texttt{sub}[i, i']$ contains `True` if $S[i...|S|]$ is a substring of $S'[i'...|S'|]$. The entries of this table can be computed in time $O(n\Delta)$ by straightforward dynamic programming.

Note that during recursive calls, the values of Δ and δ are non-increasing, and $\Delta + \delta$ decreases by 1 in the case where two recursive calls are performed. In particular, if $\ell \leq \min\{|S|, |S'|\}$ in a recursive call then $||S|-|S'|| \leq \Delta$, which enables us to use the precomputed table for the substring test. So the total number of leaves in the tree of recursive calls is at most $2^{\delta+\Delta}$, each call taking constant time, and the height of this tree is at most $\ell+d+\Delta \leq 2n$. Thus the algorithm takes overall time $O(2^{\delta+\Delta}n)$. ◀

▶ Remark 7. Algorithm 1 can be adapted to output only the set of maximal common substrings, rather than an extended version of it, by simply removing non-maximal strings (which can be done in quadratic time in the size of the output set). However, this does not improve the theoretical size of the returned set since in the worst case it does not filter out any string, but adds a quadratic running time to the complexity. It should however be an important step in an implementation of the algorithm, since an additive quadratic computation would probably be quickly compensated by pruning a possibly exponential search-tree.

Lemma 6 gives a first bound on the number of strings returned by `xMCS2` (precisely, at most $2^{\delta+\Delta}$). We know that all strings have lengths between ℓ and m. However, we will need an additional information for a more precise analysis of our algorithm on k instead of two strings. Namely, the fact that there cannot be many strings of length almost m. Intuitively, a long string in the returned set corresponds to a leaf in the search tree with few branching nodes among its ancestors, which actually helps reducing the size of the search tree. On the other hand, a short string in the returned set will cause less branchings in our next algorithm. Thus, the following lemma describes the repartition of the number of maximal common subsequences of two strings based on their lengths. Note that we would obtain the same bound if we used the filtering step from Remark 7 (i.e., the same formula applies to the set $\mathrm{MCS}_\ell(\{S, S'\})$).

▶ **Lemma 8.** *Let ℓ, d, d' be integers, and S, S' be two strings of length respectively $\ell + d$ and $\ell+d'$ (note that $\{\delta, \Delta\} = \{d, d'\}$). Moreover, let N_i be the number of strings in* $\texttt{xMCS2}(\ell, S, S')$ *of length exactly $\ell + d' - i$. Then*

$$\sum_{i=0}^{d'} \frac{N_i}{(d+1)^i} \leq 1.$$

Proof. We prove this property by induction on $|S| + |S'|$.

If $\ell > \min\{|S|, |S'|\}$, then $\texttt{xMCS2}(\ell, \{S, S'\})$ is empty, and the inequality is valid. If one of S, S' is a substring of the other, then $|\texttt{xMCS2}(\ell, S, S')| = 1$, so we have $N_i = 1$ for some i and $N_j = 0$ for $j \neq i$. The formula follows in this case as well. Note that this includes the cases where S or S' are empty.

In the remaining cases, S and S' are not substrings of each other, so in particular they are not empty. Let $u \cdot T := S$, $u' \cdot T' := S'$.

If $u = u'$, then N_i is upper-bounded by the number of strings of length $(\ell - 1) + d' - i$ in $\texttt{xMCS2}(\ell - 1, T, T')$, so we can directly apply the property by induction to get $\sum_{i=0}^{d'} \frac{N_i}{(d+1)^i} \leq 1$.

Otherwise ($u \neq u'$), let N_i^a (resp. N_i^b) be the number of strings of length $\ell + d' - i$ in $\texttt{xMCS2}(\ell, S, T')$ (resp. $\texttt{xMCS2}(\ell, T, S')$). We have $N_i \leq N_i^a + N_i^b \leq 2N_i$ (accounting for the fact that a string counted in N_i must be counted once one of N_i^a, N_i^b, and at most twice in total). Note that $N_0 = 0$ (otherwise, S and S' have a common substring of length $\ell + d' = |S'|$, which implies S' is a substring of S). Thus $N_0^a = N_0^b = 0$.

We apply the induction property first on pair $\{S, T'\}$. Note that d' decreases by 1 and indices of N_i are shifted by 1, which gives $\sum_{i=0}^{d'-1} \frac{N_{i+1}^a}{(d+1)^i} \leq 1$, so

$$\sum_{i=0}^{d'} \frac{N_i^a}{(d+1)^i} = N_0^a + \frac{1}{d+1} \sum_{i=1}^{d'} \frac{N_i^a}{(d+1)^{i-1}} \leq \frac{1}{d+1}.$$

Then the induction property on $\{T, S'\}$ (where d decreases by 1) gives $\sum_{i=0}^{d'} \frac{N_i^b}{d^i} \leq 1$, so

$$\begin{aligned}
\sum_{i=0}^{d'} \frac{N_i^b}{(d+1)^i} &= N_0^b + \sum_{i=1}^{d'} \frac{N_i^b}{(d+1)^i} \\
&= \frac{d}{d+1} \sum_{i=1}^{d'} \frac{d^{i-1}}{(d+1)^{i-1}} \frac{N_i^b}{d^i} \\
&\leq \frac{d}{d+1} \sum_{i=1}^{d'} \frac{N_i^b}{d^i} \leq \frac{d}{d+1}.
\end{aligned}$$

Combining both inequalities yields:

$$\begin{aligned}
\sum_{i=0}^{d'} \frac{N_i}{(d+1)^i} &\leq \sum_{i=0}^{d'} \frac{N_i^a}{(d+1)^i} + \sum_{i=0}^{d'} \frac{N_i^b}{(d+1)^i} \\
&\leq \frac{d}{d+1} + \frac{1}{d+1} \\
&= 1
\end{aligned}$$

◀

Lemma 8 yields an upper bound of Δ^δ on the size of $\mathrm{MCS}_\ell(S, S')$. Examples (see Figure 2 and Proposition 9) indicate that this bound is close to being tight, since there exist instances where $|\mathrm{MCS}_\ell(S, S')| = \big((1 + \frac{\Delta}{\delta})^\delta$.

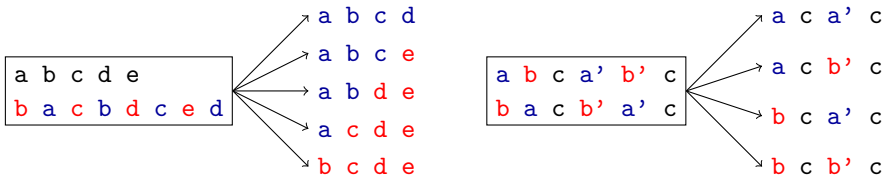

Figure 2 Examples of pairs of strings $\{S, S'\}$ with large $|\mathrm{MCS}_\ell(S, S')|$. Left: a pair with $\delta = 1, \Delta = 4$, and $|\mathrm{MCS}_\ell(S, S')| = 5 = 1 + \frac{\Delta}{\delta}$, showing that a dependency on Δ is unavoidable. Right: a pair with 2^δ maximal common subsequences, with $\delta = \Delta = 2$. Proposition 9 is a generalization of both examples that yields strings with $|\mathrm{MCS}_\ell(S, S')| = (1 + \frac{\Delta}{\delta})^\delta$.

▶ **Proposition 9.** *For any integers u and v, there exist some ℓ and two strings S, S' of length respectively $\ell + u$ and $\ell + uv$ such that $|\mathrm{MCS}_\ell(S, S')| \geq (u+1)^v = (1 + \frac{\Delta}{\delta})^\delta$.*

Proof. Let $\ell = uv$, and $\Sigma = \{x_{i,j} \mid 1 \leq i \leq v, 1 \leq j \leq u+1\}$ be an alphabet of size $(u+1)v$. Using $\prod$ as the concatenation operator, let $S = \prod_{i=1}^{v} S_i$ and $S' = \prod_{i=1}^{v} S'_i$ with

$$S_i = \prod_{j=1}^{u+1} x_{i,j}$$

$$S'_i = \prod_{j=1}^{u} x_{i,j+1}\ x_{i,j}.$$

Note that the length of S is indeed $|\Sigma| = \ell + v$ and the length of S' is $2uv = \ell + uv$. Since S_i and $S'_{i'}$ only have common characters for $i = i'$, a common substring T of S, S' is of the form $T = \prod_{i=1}^{v} T_i$ where T_i is a common substring of S_i and S'_i. Each T_i has length at most u (since S_i is not a substring of S'_i, any common substring has length at most $|S_i| - 1 = u$). If T has length at least $\ell = uv$, then each T_i has length exactly u. There are precisely $u + 1$ such common substrings for each i (all proper substrings of S_i are also substrings of S'_i). Counting all combinations of strings T_i, there are a total of $(u+1)^v$ common substrings of S, S' of length ℓ, and they are all maximal. So $|\mathrm{MCS}_\ell(S, S')| = (u+1)^v$. ◀

2.4 Extended MCS of k strings

We now present our algorithm computing an extended MCS for any number k of strings, using `xMCS2` as a subroutine, see Algorithm 2. We first give the recurrence relation on MCS on which the algorithm is based.

▶ **Lemma 10.** *Let $\mathcal{S} = \{S_1, \ldots, S_k\}$ be a set of strings ($k \geq 2$) and ℓ be an integer. Let $M' = \mathrm{MCS}_\ell(\{S_1, \ldots, S_{k-1}\})$, then*

$$\mathrm{MCS}_\ell(\mathcal{S}) \subseteq \bigcup_{S' \in M'} \mathrm{MCS}_\ell(\{S', S_k\}).$$

Proof. Consider some string $S \in \mathrm{MCS}_\ell(\mathcal{S})$. Then in particular S is a common subsequence of $\{S_1, \ldots, S_{k-1}\}$ of length at least ℓ, and so $S \in \mathrm{CS}_\ell(\{S_1, \ldots, S_{k-1}\})$. By definition of MCS, there exists a string S' in $\mathrm{MCS}_\ell(\{S_1, \ldots, S_{k-1}\})$ such that $S \preceq S'$

Since S is a subsequence of both S' and S_k, we have that $S \in \mathrm{CS}_\ell(\{S', S_k\})$. To see that S is also in $\mathrm{MCS}_\ell(\{S', S_k\})$, assume that $S'' \in \mathrm{CS}_\ell(\{S', S_k\})$ and $S \preceq S''$. Then S'' is in $\mathrm{CS}_\ell(\mathcal{S})$ (since $S' \in \mathrm{CS}_\ell(\{S_1, \ldots, S_i\})$), and since S is maximal in $\mathrm{CS}_\ell(\mathcal{S})$, then $S = S''$.

Thus S is in $\mathrm{MCS}_\ell(\{S', S_{i+1}\})$ for some $S' \in \mathrm{MCS}_\ell(\{S_1, \ldots, S_{k-1}\})$, which gives the desired inclusion. ◀

Algorithm 2 Compute a bounded-size extended MCS of k strings.

```
xMCSk(ℓ, S_1, ..., S_k):
  assert(∀i, |S_i| ≥ |S_1| )
  if k = 1:
    if |S_1| ≥ ℓ: return {S_1}
    else: return {}
  else:
    M' := xMCSk(ℓ, S_1, ..., S_{k-1})
    M := {}
    for S' in M':
      M := M ∪ xMCS2(ℓ,S_k,S')
    return M
```

▶ Remark 11. We note that the containment in Lemma 10 may sometimes be strict, as can be seen in the following example with $\ell = 1$. Take $S_1 = ABC$ and $S_2 = ACB$. Then $\mathrm{MCS}_\ell(\{S_1, S_2\}) = \{AB, AC\}$. Combining strings AB and AC with $S_3 = AAB$ yields respectively $\mathrm{MCS}_\ell(\{S_3, AB\}) = \{AB\}$ and $\mathrm{MCS}_\ell(\{S_3, AC\}) = \{A\}$. However, only AB (and not A) is part of $\mathrm{MCS}_\ell(\{S_1, S_2, S_3\})$. As for `xMCS2`, `xMCSk` outputs these extra strings to avoid a costly filtering step without any gain in the worst case.

▶ **Corollary 12.** *Given $\mathcal{S}$ and ℓ, Algorithm 2 correctly computes an extended MCS of $(\mathcal{S}, \ell)$.*

We now bound the number of strings at any point in the set M of the algorithm. The key point here is that this bound does not depend on k or n. This may seem counter-intuitive, compared to the following upper bound: the algorithm starts with a single string, and each recursive call may replace any string by up to $2^{\delta+\Delta}$ strings (cf. the complexity of `xMCS2`). There are k recursive calls so this would give a bound of $2^{k(\delta+\Delta)}$ strings in total. The key argument here is that whenever a string is replaced, new strings are strictly shorter than the former. Since we only allow for at most δ deletions (starting from a minimal length input string), this gives a bound depending on δ and Δ only. Our more precise analysis in Lemma 13 allows us to shrink this quantity from $2^{O(\Delta\delta)}$ to $2^{O(\log(\Delta)\delta)}$.

▶ **Lemma 13.** *Let $\mathcal{S} = \{S_1, \ldots, S_k\}$ be a set of strings with S_1 of minimal length (i.e. $|S_1| = m$), and ℓ be an integer. Then*

$$|\texttt{xMCSk}(\ell, \mathcal{S})| \leq (\Delta + 1)^\delta.$$

Proof. We prove the following property by induction on k: let $d \geq \Delta$, and N_i be the number of strings in $\texttt{xMCSk}(\ell, \mathcal{S})$ of length $\ell + d - i$. Then

$$\sum_{i=0}^{\delta} \frac{N_i}{(d+1)^i} \leq 1.$$

Note that the lemma statement follows easily from this property for $d = \Delta$:

$$\frac{|\texttt{xMCSk}(\ell, \mathcal{S})|}{(\Delta+1)^\delta} = \sum_{i=0}^{\delta} \frac{N_i}{(\Delta+1)^\delta} \leq \sum_{i=0}^{\delta} \frac{N_i}{(\Delta+1)^i} \leq 1$$

For $k = 1$, we have a single string in $\texttt{xMCSk}(\ell, \mathcal{S})$, namely, S_1, so $N_i = 1$ for exactly one value of i and 0 otherwise, and the formula is satisfied.

For $k \geq 2$, we have $M' = \texttt{xMCSk}(\ell, \{S_1, \ldots, S_{k-1}\})$. Consider the for-loop in Lines 9–10. We assume that when we iterate with $S' \in M'$, the string S' is immediately removed from M'. At any point of the loop, we write σ for the quantity $\sum_{i=0}^{\delta} \frac{N_i}{(d+1)^i}$ where N_i denotes the number of strings of length $\ell + \delta - i$ in $M' \cup M$. Note that by induction, before the first iteration of the loop, $\sigma \leq 1$ (using the fact that $\delta(\{S_1, \ldots, S_{k-1}\}, \ell) = \delta$ since $|S_1|$ is minimal, and $d \geq \Delta \geq \Delta(\{S_1, \ldots, S_{k-1}\}, \ell)$).

We show that σ may only decrease after each iteration. Consider the iteration for string S', let $d' = |S'| - \ell$ and $j = \delta - d'$ (since S' is a substring of it has length at most $\ell + \delta$, so $d' \leq \delta$ and $j \geq 0$).

First, removing S' from M' makes N_j decrease by one, so σ decreases by $\frac{1}{(d+1)^j}$. Then, we insert strings from $\texttt{xMCS2}(\{S_k, S'\})$ in M. Write D_i for the number of such strings of length $\ell + d' - i$. Note that for each pair i, j with $j \leq i \leq \delta$, N_i increases by D_{i-j}. By Lemma 8, $\sum_{i=0}^{d'} \frac{D_i}{(|S_k|-\ell+1)^i} \leq 1$. Since $d \geq \Delta \geq |S_k| - \ell$, $\sum_{i=0}^{d'} \frac{D_i}{(d+1)^i} \leq 1$. Then σ increases by $\sum_{i=0}^{d'} \frac{D_i}{(d+1)^{j+i}} = \frac{1}{(d+1)^j} \sum_{i=j}^{\delta} \frac{D_{i-j}}{(d+1)^i} \leq \frac{1}{(d+1)^j}$. Overall, σ may not increase between two steps, so at the end of the for loop, $\sum_{i=0}^{\delta} \frac{N_i}{(d+1)^i} \leq 1$. ◀

We can now conlude with the proof of Theorem 1.

Proof of Theorem 1. Given $\mathcal{S}$ and ℓ, Algorithm 2 computes an extended MCS of $(\mathcal{S}, \ell)$ (Corollary 12) of size at most $(\Delta + 1)^\delta$ (Lemma 13). Its running time is bounded by k times the complexity of the for loop, which requires at most $(\Delta + 1)^\delta$ calls to `xMCS2`, each taking time $O(2^{\delta+\Delta} n)$ (Lemma 6). This gives the overall complexity of $O(2^{\delta+\Delta}(\Delta + 1)^\delta kn)$. ◀

3 Conclusion

Regarding LCS, we have proposed an FPT algorithm for parameter Δ, i.e., the maximum number of deletions per input string. It is open whether the complexity could be improved, e.g. using only parameter δ, i.e., the smallest number of deletions per input strings. In other words, the goal is to find an LCS of size ℓ in a set of strings where *one* string has size at most $\ell + \delta$ (and other strings might be arbitrarily long). Such an algorithm may not compute and store explicitely each MCS, since the number of maximal common subsequenes, even with only two input strings, can grow in $(1 + \frac{\Delta}{\delta})^\delta$. Also, it is open whether any improvement can be obtained when the alphabet size is bounded, or when each character has a bounded number of occurrences in each string.

A longest common subsequence can be seen as a string that can be obtained with a minimal number of edits (deletions only) from all input strings. Generalizing this notion to other edits (insertions and substitutions) yields the Center String problem, which is highly related to the problem of Multiple Sequence Alignment in bioinformatics. In future work, we aim at extending our approach in order to design an FPT algorithm for Center String, parameterized by the maximum distance to input strings. Allowing for a small number of outliers (input strings that are discarded in order to obtain a better solution [4]) would also be a useful extension of our algorithm.

Finally, a more practical objective towards algorithm engineering would be to design an efficient data structure to store all maximal common subsequences of any number of strings, in order to reduce the memory footprint of our algorithm.

References

1. Amir Abboud, Arturs Backurs, and Virginia Vassilevska Williams. Tight hardness results for LCS and other sequence similarity measures. In *IEEE 56th Annual Symposium on Foundations of Computer Science, FOCS 2015*, pages 59–78. IEEE Computer Society, 2015.
2. Lasse Bergroth, Harri Hakonen, and Timo Raita. A survey of longest common subsequence algorithms. In *Seventh International Symposium on String Processing and Information Retrieval, SPIRE 2000*, pages 39–48. IEEE Computer Society, 2000.
3. Mahdi Boroujeni, Masoud Seddighin, and Saeed Seddighin. Improved algorithms for edit distance and LCS: beyond worst case. In *Proceedings of the 2020 ACM-SIAM Symposium on Discrete Algorithms, SODA 2020*, pages 1601–1620. SIAM, 2020.
4. Christina Boucher and Bin Ma. Closest string with outliers. *BMC Bioinformatics*, 12(1):1–7, 2011.
5. Karl Bringmann and Marvin Künnemann. Multivariate fine-grained complexity of longest common subsequence. In *Proceedings of the Twenty-Ninth Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2018*, pages 1216–1235. SIAM, 2018.
6. Laurent Bulteau, Falk Hüffner, Christian Komusiewicz, and Rolf Niedermeier. Multivariate algorithmics for NP-hard string problems. *Bulletin of the EATCS*, 114, 2014.
7. Laurent Bulteau and Mathias Weller. Parameterized algorithms in bioinformatics: An overview. *Algorithms*, 12(12):256, 2019.
8. Rodney G. Downey and Michael R. Fellows. *Parameterized Complexity*. Springer, 1999.
9. MohammadTaghi Hajiaghayi, Masoud Seddighin, Saeed Seddighin, and Xiaorui Sun. Approximating LCS in linear time: Beating the $\sqrt{n}$ barrier. In *Proceedings of the Thirtieth Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2019*, pages 1181–1200. SIAM, 2019.
10. Robert W. Irving and Campbell Fraser. Two algorithms for the longest common subsequence of three (or more) strings. In *Combinatorial Pattern Matching, Third Annual Symposium, CPM 1992*, volume 644 of *Lecture Notes in Computer Science*, pages 214–229. Springer, 1992.
11. Narao Nakatsu, Yahiko Kambayashi, and Shuzo Yajima. A longest common subsequence algorithm suitable for similar text strings. *Acta Informatica*, 18:171–179, 1982.
12. Krzysztof Pietrzak. On the parameterized complexity of the fixed alphabet shortest common supersequence and longest common subsequence problems. *Journal of Computer and System Sciences*, 67(4):757–771, 2003.