Polynomial-Time Equivalences and Refined Algorithms for Longest Common Subsequence Variants

Yuichi Asahiro
Kyushu Sangyo University, Fukuoka, Japan

Jesper Jansson
Kyoto University, Japan

Guohui Lin
University of Alberta, Edmonton, Canada

Eiji Miyano
Kyushu Institute of Technology, Iizuka, Japan

Hirotaka Ono
Nagoya University, Japan

Tadatoshi Utashima
Kyushu Institute of Technology, Iizuka, Japan

Abstract

The problem of computing the longest common subsequence of two sequences (LCS for short) is a classical and fundamental problem in computer science. In this paper, we study four variants of LCS: the Repetition-Bounded Longest Common Subsequence problem (RBLCS) [2], the Multiset-Restricted Common Subsequence problem (MRCS) [11], the Two-Side-Filled Longest Common Subsequence problem (2FLCS), and the One-Side-Filled Longest Common Subsequence problem (1FLCS) [5, 6]. Although the original LCS can be solved in polynomial time, all these four variants are known to be NP-hard. Recently, an exact, $O(1.44225^n)$-time, dynamic programming (DP)-based algorithm for RBLCS was proposed [2], where the two input sequences have lengths n and $poly(n)$. We first establish that each of MRCS, 1FLCS, and 2FLCS is polynomially equivalent to RBLCS. Then, we design a refined DP-based algorithm for RBLCS that runs in $O(1.41422^n)$ time, which implies that MRCS, 1FLCS, and 2FLCS can also be solved in $O(1.41422^n)$ time. Finally, we give a polynomial-time 2-approximation algorithm for 2FLCS.

2012 ACM Subject Classification Theory of computation → Design and analysis of algorithms

Keywords and phrases Repetition-bounded longest common subsequence problem, multiset restricted longest common subsequence problem, one-side-filled longest common subsequence problem, two-side-filled longest common subsequence problem, exact algorithms, and approximation algorithms

Digital Object Identifier 10.4230/LIPIcs.CPM.2022.15

Funding This work is partially supported by NSERC Canada, JST CREST JPMJR1402, and JSPS KAKENHI Grant Numbers JP20H05967, JP21K11755, JP21K19765, JP22H00513, and JP22K11915.

1 Introduction

1.1 Longest common subsequence problems with occurrence constraints

The problem of computing the longest common subsequence of two sequences (LCS for short) is a classical and fundamental problem in computer science [3, 4, 9, 16]. Indeed, many polynomial-time algorithms have been published for LCS [8, 9, 14, 15, 16]. A natural extension of LCS is to impose constraints on the occurrences of the symbols in the solution. It has been shown that even very simple constraints may make the problem computationally

much harder. As an example, the Repetition-Free Longest Common Subsequence problem (RFLCS), introduced by Adi et al. [1] is: Given two sequences X and Y over an alphabet Σ, the goal of RFLCS is to find a "*repetition-free*" longest common subsequence of X and Y, where each symbol appears at most once in the obtained subsequence. Adi et al. [1] proved that RFLCS is APX-hard even if each symbol appears at most twice in each of the given sequences. On the positive side, they showed that RFLCS admits a polynomial-time occ_{max}-approximation algorithm, where occ_{max} is defined as follows: Let $occ(W, \sigma)$ be the number of occurrences of a symbol σ in a sequence W. Then occ_{max} is the maximum of $\min\{occ(X, \sigma), occ(Y, \sigma)\}$ taken over all σ's in two sequences X and Y.

Mincu and Popa [11] introduced a general form of RFLCS, called the Multiset Restricted Common Subsequence problem (MRCS): Given two sequences X and Y, and a multiset $\mathcal{M}$ over the alphabet Σ, the goal of MRCS is to find a common subsequence $Z_{\mathcal{M}}$ of X and Y, that contains the maximum number of symbols from $\mathcal{M}$. If $\mathcal{M} = \Sigma$, then MRCS is essentially equivalent to RFLCS. Therefore, MRCS is also APX-hard. In [11], the authors showed that there exists an exact algorithm solving MRCS with running time $O(|X||Y|(t+1)^{|\Sigma|})$, where t is the maximum multiplicity of symbols in $\mathcal{M}$. Also, they provided a polynomial-time $2\sqrt{\min\{|X|, |Y|\}}$-approximation algorithm for MRCS [11].

Recently, Asahiro et al. [2] introduced a slightly different generalization of RFLCS, called the Repetition-Bounded Longest Common Subsequence problem (RBLCS for short): Let $\Sigma = \{\sigma_1, \sigma_2, \ldots, \sigma_k\}$ be an alphabet of k symbols and C_{occ} be an occurrence constraint $C_{occ} : \Sigma \to \mathbb{N}$, assigning an upper bound on the number of occurrences of each symbol in Σ. Given two sequences X and Y over the alphabet Σ and an occurrence constraint C_{occ}, the goal of RBLCS is to find a "*repetition-bounded*" longest common subsequence of X and Y, where each symbol σ_i appears at most $C_{occ}(\sigma_i)$-times in the obtained subsequence for $i = 1, 2, \ldots, k$. In [2], Asahiro et al. provided a dynamic programming (DP) based algorithm for RBLCS and proved that its running time is $O(1.44225^{|X|})$ for any occurrence constraint C_{occ}, assuming $|X| \leq |Y|$ and $|Y| = O(poly(|X|))$, and even less in certain special cases. In particular, for RFLCS, their DP-based algorithm runs in $O(1.41422^{|X|})$ time. NP-hardness and APX-hardness results for RBLCS on restricted instances were also shown in [2].

1.2 Longest common subsequence problems on incomplete sequences

The comparison of biological sequences is a widely investigated field of bioinformatics, in which the genomic features including DNA sequences and genes of different organisms are compared in order to identify biological differences and similarities. In genomic analyses, however, the considered genomes are usually not complete and thus there are cases where we have to reconstruct complete genomes from incomplete genomes (so-called *scaffolds*) by filling in missing genes. For this purpose, Muñoz et al. [13] formulated the following combinatorial optimization problem, called the One-Sided Scaffold Filling problem (1SF): Given an incomplete genome Y, a multiset $\mathcal{M}$ of missing genes, and a reference genome X, the goal of 1SF is to insert the missing genes into Y so that the number of common adjacencies between the resulting Y^* and X is maximized. Subsequently, Jiang et al. [10] proposed the *Two-Sided Scaffold Filling* problem (2SF): Given two scaffolds (incomplete genomes), the goal of 2SF is to fill the missing genes into those two scaffolds respectively to result in such two genomes that the number of common adjacencies between them is maximized.

Inspired by methods for genome comparison based on LCS and by 1SF/2SF, Castelli et al. [5] introduced a new variant of LCS, called the One-Side-Filled Longest Common Subsequence problem (1FLCS), which aims to compare a complete sequence with an incomplete one, i.e., with some missing elements: Given a complete sequence X, an incomplete

sequence Y, and a multiset $\mathcal{M}_Y$ of symbols missing in Y, 1FLCS asks for a sequence Y^+ obtained by inserting a subset of the symbols of $\mathcal{M}_Y$ into Y so that Y^+ induces a common subsequence with X of maximum length. The authors proved the APX-hardness of 1FLCS and designed a polynomial-time $\frac{5}{3}$-approximation algorithm for 1FLCS. They also presented an exponential-time exact algorithm for 1FLCS. (However, they did not analyze its time complexity in detail.) In [6], Castelli et al. showed that if the alphabet size $|\Sigma|$ is a constant, then there is a polynomial-time algorithm for 1FLCS, and concluded by introducing the Two-Side-Filled Longest Common Subsequence problem (2FLCS), i.e., LCS on two incomplete sequences and two multisets of missing symbols: Given two incomplete sequences X and Y, and two multisets $\mathcal{M}_X$ and $\mathcal{M}_Y$, 2FLCS asks for two sequences X^+ and Y^+ obtained by inserting subsets of the symbols of $\mathcal{M}_X$ and $\mathcal{M}_Y$ into X and Y, respectively, so that X^+ and Y^+ induce a common subsequence of maximum length. They conjectured that 2FLCS can be approximated within a constant factor in polynomial time, and that the following simple method gives a 2-approximation: (1) First find a longest common subsequence Z_1 of input two sequences X and Y. Then, (2) obtain a sequence Z_2 that maximizes the number of symbols matched by inserting symbols of $\mathcal{M}_X$ and $\mathcal{M}_Y$. Finally, (3) output the longest of Z_1 and Z_2. Moreover, they conjectured that 2FLCS can be solved in polynomial time if the alphabet size is a constant.

1.3 Our contributions

Suppose that there exist an $O(T_A)$-time algorithm for an optimization problem P_A and an $O(T_B)$-time algorithm for another optimization problem P_B. In this paper, we say that two problems P_A and P_B are *polynomially equivalent*, or that *polynomial-time equivalence* between P_A and P_B holds, if an optimal solution for an instance I_A of P_A can be obtained in $O(T_B) + O(poly(|I_A|))$ time and an optimal solution for an instance I_B of P_B can be obtained in $O(T_A) + O(poly(|I_B|))$ time. Our contributions are:

1. We establish that MRCS is polynomially equivalent to RBLCS by showing the following: (i) From an input $(X, Y, \mathcal{M})$ of MRCS, we construct an input (X, Y, C_{occ}) of RBLCS in $O(poly(|(X, Y, \mathcal{M})|))$ time. Then, from an optimal solution Z_R of RBLCS on (X, Y, C_{occ}), we construct an optimal solution Z_M of MRCS on $(X, Y, \mathcal{M})$ in $O(poly(|(X, Y, \mathcal{M})|))$ time. Conversely, (ii) from an input (X, Y, C_{occ}) of RBLCS, we construct an input $(X, Y, \mathcal{M})$ of MRCS in $O(poly(|(X, Y, C_{occ})|))$ time. Then, from an optimal solution Z_M of MRCS on $(X, Y, \mathcal{M})$, we construct an optimal solution Z_R in $O(poly(|(X, Y, C_{occ})|))$ time. It is important to note that our constructions between two inputs are "input-sequences preserving reductions", i.e., X and Y in $(X, Y, \mathcal{M})$ and (X, Y, C_{occ}) are identical.
2. Similarly to the above, we show the polynomial-time equivalence between 1FLCS and RBLCS: (i) From an input $(X, Y, \mathcal{M}_Y)$ of 1FLCS, we construct an input (X, Y, C_{occ}) of RBLCS in $O(poly(|(X, Y, \mathcal{M}_Y)|))$ time. Then, from an optimal solution Z_R of RBLCS on (X, Y, C_{occ}), we construct an optimal solution Z_{1F} of 1FLCS on $(X, Y, \mathcal{M}_Y)$ in $O(poly(|(X, Y, \mathcal{M}_Y)|))$ time. Conversely, (ii) from an input (X, Y, C_{occ}) of RBLCS, we construct an input $(X, Y, \mathcal{M}_Y)$ of 1FLCS in $O(poly(|(X, Y, C_{occ})|))$ time. Then, from an optimal solution Z_{1F} of 1FLCS on $(X, Y, \mathcal{M}_Y)$, we construct an optimal solution Z_R of RBLCS on (X, Y, C_{occ}) in $O(poly(|(X, Y, C_{occ})|))$ time.
3. We prove the polynomial-time equivalence between 2FLCS and RBLCS. Due to the second contribution and 1FLCS being a special case of 2FLCS, we only need to show one direction: (i) From an input $(X, Y, \mathcal{M}_X, \mathcal{M}_Y)$ of 2FLCS, we construct an input (X, Y, C_{occ}) of RBLCS in $O(poly(|(X, Y, \mathcal{M}_X, \mathcal{M}_Y)|))$ time. Then, from an optimal solution Z_R of RBLCS on (X, Y, C_{occ}), we construct an optimal solution Z_{2F} of 2FLCS on $(X, Y, \mathcal{M}_X, \mathcal{M}_Y)$ in $O(poly(|Z_R|))$ time.

4. We design a refined DP-based algorithm that runs in $O(1.41422^n)$ time for RBLCS on two sequences X of length n and Y of length m (assuming that $n \leq m$ and $m = O(poly(n))$), while the previously known running time was $O(1.44225^n)$ in [2].
5. We give a simple polynomial-time 2-approximation algorithm for 2FLCS, thus resolving one of the conjectures in [6].

▶ Remark 1. One sees that 1FLCS on $(X, Y, \mathcal{M}_Y)$ is equivalent to 2FLCS on $(X, Y, \emptyset, \mathcal{M}_Y)$; 1FLCS can be solved by using an algorithm for 2FLCS. From (ii) in the second contribution, RBLCS can also be solved by using the algorithm for 2FLCS with some extra polynomial-time calculations. Therefore, the one-way equivalence in the third contribution demonstrates the "two-way" polynomial-time equivalence between 2FLCS and RBLCS. Furthermore, interestingly, an algorithm for 1FLCS can solve 2FLCS within an extra polynomial-time factor.

▶ Remark 2. None of the constructions between inputs described above change the sequences X and Y. In particular, $|X|$ and $|Y|$ remain the same, so the above polynomial-time equivalences imply that MRCS, 1FLCS, and 2FLCS can also be solved in $O(1.41422^n)$ time.

▶ Remark 3. We also remark that the polynomial-time equivalence between 1FLCS and 2FLCS gives an affirmative answer to the conjecture on the polynomial-time solvability of 2FLCS for a constant size alphabet in [6] since we do not change Σ.

2 Preliminaries

2.1 Notation

An *alphabet* $\Sigma = \{\sigma_1, \sigma_2, \ldots, \sigma_k\}$ is a set of k *symbols*. Let X be a sequence over the alphabet Σ and $|X|$ be the length of the sequence X. Throughout the paper, a sequence X is often regarded as a *multiset* of the same symbols. For example, $X = \langle x_1, x_2, \ldots, x_n \rangle$ is a sequence of length n, where $x_i \in \Sigma$ for $1 \leq i \leq n$, i.e., $|X| = n$. A *subsequence* of X is obtained by deleting zero or more symbols from X. Then, we say that a sequence Z is a *common subsequence* of X and Y if Z is a subsequence of both X and Y. Given two sequences X and Y as input, the goal of the Longest Common Subsequence problem (LCS) is to find a *longest* common subsequence of X and Y, which is denoted by $LCS(X, Y)$. Let $L(X, Y)$ denote the length of $LCS(X, Y)$.

For the sequence X, the *consecutive subsequence*, i.e., *substring* $\langle x_i, x_{i+1}, \ldots, x_j \rangle$ is denoted by $X_{i..j}$. Then, we define the ith *prefix* of X, for $i = 1, \ldots, n$, as $X_{1..i} = \langle x_1, x_2, \ldots, x_i \rangle$. Also, we define the ith *suffix* of X, for $i = 1, \ldots, n$, as $X_{i..n} = \langle x_i, x_{i+1}, \ldots, x_n \rangle$. $X_{1..n}$ is X.

Let $X = \langle x_1, x_2, \ldots, x_n \rangle$ and $Y = \langle y_1, y_2, \ldots, y_m \rangle$ be the given two sequences of length n and length m, respectively. Assume that $n \leq m$ and $m = O(poly(n))$ throughout the paper. Suppose that $Z = \langle z_1, z_2, \ldots, z_p \rangle$ is a common subsequence with length p of X and Y. Then, we can consider two strictly increasing sequences $I_X = \langle i_1, i_2, \ldots, i_p \rangle$ of indices of X and $I_Y = \langle j_1, i_2, \ldots, j_p \rangle$ of indices of Y such that $z_\ell = x_{i_\ell} = y_{j_\ell}$ holds for each $\ell = 1, 2, \ldots, p$. We call the pair (I_X, I_Y) of such sequences an *index-expression* of the common sequence Z of X and Y. A pair (x_{i_ℓ}, y_{j_ℓ}) is called the ℓth *match*. Also, we say that the ℓth match is z_ℓ, x_{i_ℓ}, or y_{i_ℓ}.

For two sequences $A = \langle a_1, \ldots, a_i \rangle$ of length i and $B = \langle b_1, \ldots, b_j \rangle$ of length j, let $A \oplus B$ be the *concatenation* of A and B, i.e., the sequence $A \oplus B = \langle a_1, \ldots, a_i, b_1, \ldots, b_j \rangle$ of length $i+j$. For $X = \langle x_1, x_2, \ldots, x_n \rangle$ of length n, let $X \setminus \langle i \rangle$ denote the sequence obtained by deleting

the ith symbol x_i from X, i.e., $X \setminus \langle i \rangle = X_{1..i-1} \oplus X_{i+1..n} = \langle x_1, x_2, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n \rangle$. Similarly, for $1 \leq i_1 < i_2 < \cdots < i_p \leq n$, let $X \setminus \langle i_1, i_2, \ldots, i_p \rangle$ be the sequence obtained by deleting p symbols x_{i_1}, x_{i_2}, $\ldots$, x_{i_p} from X.

Let $\mathcal{M}$ be a multiset of symbols in Σ and let $|\mathcal{M}|$ be the cardinality of $\mathcal{M}$. Let $occ(\mathcal{M}, \sigma)$ denote the occurrences (i.e., the *multiplicity*) of a symbol $\sigma \in \Sigma$ in a multiset $\mathcal{M}$. Let $\mathcal{M} \setminus \{\sigma^\ell\}$ be the multiset obtained by removing ℓ σ's from a multiset $\mathcal{M}$. Let $\mathcal{M} \setminus \{\sigma^*\}$ be the multiset obtained by removing all σ's from a multiset $\mathcal{M}$.

Consider a multiset $\mathcal{M}$ of cardinality ℓ and obtain an arbitrarily fixed sequence $M = \langle \mu_1, \mu_2, \ldots, \mu_\ell \rangle$ of ℓ symbols in $\mathcal{M}$, called a *sequence-expression* of the multiset $\mathcal{M}$. In the following, the multiset $\mathcal{M}$ is often regarded as its sequence-expression M; $\mathcal{M}$ and M are used interchangeably. Similarly to the above, for $1 \leq i_1 < i_2 < \cdots < i_p \leq \ell$, let $M \setminus \langle i_1, i_2, \ldots, i_p \rangle$ be the sequence obtained by deleting p symbols μ_{i_1}, μ_{i_2}, $\ldots$, μ_{i_p} from M.

An algorithm `ALG` is called an α-approximation algorithm and `ALG`'s approximation ratio is α if $OPT(x)/ALG(x) \leq \alpha$ holds for every input x of an LCS-type problem, where $ALG(x)$ and $OPT(x)$ are the length of solutions obtained by `ALG` and an optimal algorithm in polynomial-time.

2.2 Repetition-bounded longest common subsequence

Recall that $occ(W, \sigma)$ is the number of occurrences of $\sigma \in \Sigma$ in a sequence W. Without loss of generality, we assume that two input sequences X and Y have all k symbols in Σ, and thus $occ(X, \sigma_i) \geq 1$ and $occ(Y, \sigma_i) \geq 1$ for every symbol σ_i. Let C_{occ} be an occurrence constraint, i.e., a function $C_{occ} : \Sigma \to \mathbb{N}$ assigning an upper bound on the number of occurrences of each symbol in Σ. The Repetition-Bounded Longest Common Subsequence problem (RBLCS) can be formally defined as follows [2]:

> Repetition-Bounded Longest Common Subsequence problem (RBLCS)
> **Input:** A pair of sequences X and Y, and an occurrence constraint C_{occ}.
> **Goal:** Find a longest common subsequence Z of X and Y such that $occ(Z, \sigma) \leq C_{occ}(\sigma)$ is satisfied for every $\sigma \in \Sigma$.

We call Z a *repetition-bounded* longest common subsequence. Let $LCS(X, Y, C_{occ})$ denote the repetition-bounded longest common subsequence for the input triple (X, Y, C_{occ}). Also, $L(X, Y, C_{occ})$ denotes the length of $LCS(X, Y, C_{occ})$.

▶ **Example 4.** Let (X, Y, C_{occ}) be an instance of RBLCS defined by:

$$X = \langle t, g, t, c, a, c, g, t, g, a, a, g \rangle,\ Y = \langle a, t, g, c, a, t, g, g, a, c, a, g, c \rangle;\ \text{and}$$
$$C_{occ}(a) = 1, C_{occ}(c) = 1, C_{occ}(g) = 2, C_{occ}(t) = 1.$$

$Z = \langle g, c, t, g, a \rangle$ of length five is an optimal solution of RBLCS since $occ(Z, a) = 1$, $occ(Z, c) = 1$, $occ(Z, g) = 2$, $occ(Z, t) = 1$, and $\sum_{\sigma \in \{a,c,g,t\}} C_{occ}(\sigma) = 5$, i.e., $L(X, Y, C_{occ}) = 5$. As a side note, $\langle t, g, c, a, t, g, a, a, g \rangle$ of length nine is an optimal solution of the original LCS.

Consider an input triple (X, Y, C_{occ}) of RBLCS and a feasible solution Z_R for (X, Y, C_{occ}). Then, for every $\sigma \in \Sigma$, the number of occurrences $occ(Z_R, \sigma)$ of σ must be bounded above by $C_{occ}(\sigma)$. If $C_{occ}(\sigma') > \min\{occ(X, \sigma'), occ(Y, \sigma')\}$ for some σ', then the constraint C_{occ} is somewhere redundant. Therefore, if the input (X, Y, C_{occ}) of RBLCS satisfies $C_{occ}(\sigma) \leq \min\{occ(X, \sigma), occ(Y, \sigma)\}$ for every $\sigma \in \Sigma$, then we call (X, Y, C_{occ}) the *standard* input. Without loss of generality, we assume that every input of RBLCS is standard in the following.

2.3 Multiset restricted common subsequence

The formal definition of the Multiset Restricted Common Subsequence problem (MRCS) is as follows [11]:

Multiset Restricted Common Subsequence problem (MRCS)
Input: A pair of sequences X and Y, and a multiset $\mathcal{M}$.
Goal: Find a common subsequence Z of X and Y such that Z contains the maximum number of symbols from $\mathcal{M}$.

That is, the goal of MRCS is to maximize $|\mathcal{M} \cap Z|$ as a multiset intersection or, equivalently, to minimize $|\mathcal{M} \setminus Z|$ as a multiset difference (if Z is regarded as the corresponding multiset). The optimal solution Z is denoted by $LCS(X, Y, \mathcal{M})$ in the following. The length of $LCS(X, Y, \mathcal{M})$ is denoted by $L(X, Y, \mathcal{M})$.

▶ **Example 5.** Consider the following input triple $(X, Y, \mathcal{M})$ of MRCS:

$$X = \langle t, g, t, c, a, c, g, t, g, a, a, g\rangle,\ Y = \langle a, t, g, c, a, t, g, g, a, c, a, g, c\rangle,\ \mathcal{M} = \{a, c, g, g, t\}$$

One sees that a common subsequence $\langle g, c, t, g, a\rangle$ of X and Y is an optimal solution of MRCS since $|\mathcal{M}| = 5$ and solutions of length five with all the symbols in $\mathcal{M}$ are equally as good as longer solutions. For example, the objective function value of a longer common subsequence $Z = \langle g, c, t, g, a, a, g\rangle$ is also five since $|\mathcal{M} \cap Z| = 5$.

2.4 Filled longest common subsequence

Let $\mathcal{M}_X$ ($\mathcal{M}_Y$, resp.) be a multiset of symbols in Σ. Then, we denote the cardinality of the multiset $\mathcal{M}_X$ ($\mathcal{M}_Y$, resp.) by $|\mathcal{M}_X|$ ($|\mathcal{M}_Y|$, resp.), i.e., $\sum_{\sigma\in\mathcal{M}_X} occ(\mathcal{M}_X, \sigma)$ ($\sum_{\sigma\in\mathcal{M}_Y} occ(\mathcal{M}_Y, \sigma)$, resp.). A *filling* X^+ (Y^+, resp.) of the sequence X (Y, resp.) is defined as a sequence obtained from X (Y, resp.) by inserting a subset of the symbols from $\mathcal{M}_X$ ($\mathcal{M}_Y$, resp.) into X (Y, resp.). That is, for some $0 \le p \le |\mathcal{M}_X|$ and $\mathcal{M}'_X = \{\chi_1, \ldots, \chi_p\} \subseteq \mathcal{M}_X$, the filling X^+ obtained by inserting $\mathcal{M}'_X$ into X is the following concatenation of $2p + 1$ subsequences (some might be a *null* sequence):

$$X^+ = X_{1..j_1} \oplus \langle\chi_{i_1}\rangle \oplus X_{j_1+1..j_2} \oplus \langle\chi_{i_2}\rangle \oplus \cdots \oplus \langle\chi_{i_p}\rangle \oplus X_{j_{p+1}..n},$$

where $X = X_{1..j_1} \oplus X_{j_1+1..j_2} \oplus \cdots \oplus X_{j_{p+1}..n}$ and $\{i_1, \ldots, i_p\} = \{1, \ldots, p\}$. For some $0 \le q \le |\mathcal{M}_Y|$ and $\mathcal{M}'_Y = \{\psi_1, \ldots, \psi_q\} \subseteq \mathcal{M}_Y$, the filling Y^+ obtained by inserting $\mathcal{M}'_Y$ into Y is similarly defined. Let X^* and Y^* be fillings such that the length of $LCS(X^*, Y^*)$ is the longest among the length of $LCS(X^+, Y^+)$ over all pairs of X^+ and Y^+. The Two-Side-Filled Longest Common Subsequence problem (2FLCS) is defined as follows [6]:

Two-Side-Filled Longest Common Subsequence problem (2FLCS)
Input: A pair of sequences X and Y, and a pair of multisets $\mathcal{M}_X$ and $\mathcal{M}_Y$.
Goal: Find two fillings X^* and Y^* such that the length of $LCS(X^*, Y^*)$ is the longest among the lengths of $LCS(X^+, Y^+)$ over all pairs of X^+ and Y^+.

In the following, the longest common subsequence $LCS(X^*, Y^*)$ of two fillings X^* and Y^* is written as $LCS(X, Y, \mathcal{M}_X, \mathcal{M}_Y)$. The length of $LCS(X, Y, \mathcal{M}_X, \mathcal{M}_Y)$ is denoted by $L(X, Y, \mathcal{M}_X, \mathcal{M}_Y)$. As a special case, if $\mathcal{M}_X = \emptyset$, then the problem is called the One-Side-Filled Longest Common Subsequence problem (1FLCS) [6]:

One-Side-Filled Longest Common Subsequence problem (1FLCS)
Input: A pair of sequences X and Y, and a multiset $\mathcal{M}_Y$.
Goal: Find a filling Y^* such that the length of $LCS(X, Y^*)$ is the longest among the length of $LCS(X, Y^+)$ over all fillings Y^+.

Let $LCS(X, Y, \mathcal{M}_Y)$ and $L(X, Y, \mathcal{M}_Y)$ be the longest common subsequence $LCS(X, Y^*)$ and its length, respectively.

▶ **Example 6.** Now we consider the following example, two sequences X and Y, and two multisets $\mathcal{M}_X$ and $\mathcal{M}_Y$, as input to 2FLCS:

$$X = \langle g, t, c, a, c, t, g, a\rangle,\ Y = \langle g, a, t, c, c, g, t, g\rangle,\ \mathcal{M}_X = \{g, t\},\ \text{and } \mathcal{M}_Y = \{c, t, t\}$$

Here, for example, $occ(X, c) = 2$ and $occ(\mathcal{M}_Y, c) = 1$. One sees that for the input quadruple $(X, Y, \mathcal{M}_X, \mathcal{M}_Y)$, an optimal pair of fillings is as follows:

$$X^* = \langle \underline{t}, g, t, c, a, c, \underline{g}, t, g, a\rangle \quad \text{and} \quad Y^* = \langle \underline{t}, g, \underline{t}, \underline{c}, a, t, c, c, g, t, g\rangle.$$

That is, the leftmost $\underline{t}$ and the seventh $\underline{g}$ in X^* are inserted into the original X from $\mathcal{M}_X$. For Y^*, the first, third, and fourth symbols ($\underline{t}$, $\underline{t}$, and $\underline{c}$, respectively) are inserted into Y from $\mathcal{M}_Y$. Then, the longest common subsequence $LCS(X^*, Y^*)$ of those fillings X^* and Y^* is $\langle t, g, t, c, a, c, g, t, g\rangle$. Note that $I_{X^*} = \langle 1, 2, 3, 4, 5, 6, 7, 8, 9\rangle$ and $I_{Y^*} = \langle 1, 2, 3, 4, 5, 7, 9, 10, 11\rangle$. One can verify that, for example, the first symbol t in $LCS(X^*, Y^*)$ originally comes from $\mathcal{M}_X$ and $\mathcal{M}_Y$, but the second symbol g comes from X and Y.

Now let $X^+ = \langle x_1, x_2, \ldots, x_n\rangle$ and $Y^+ = \langle y_1, y_2, \ldots, y_m\rangle$ be two fillings of X and Y, respectively. Let (I_{X^+}, I_{Y^+}) be an index-expression of a common subsequence of two fillings X^+ and Y^+. Then, the ℓth match (x_{i_ℓ}, y_{j_ℓ}) is one of the following four types of matches:

- $\mathcal{M}_X\mathcal{M}_Y$-match: x_{i_ℓ} and y_{i_ℓ} are inserted from $\mathcal{M}_X$ and $\mathcal{M}_Y$, respectively.
- $\mathcal{M}_X Y$-match: x_{i_ℓ} is inserted from $\mathcal{M}_X$ but y_{i_ℓ} is originally in Y.
- $X\mathcal{M}_Y$-match: x_{i_ℓ} is originally in X but y_{j_ℓ} is inserted from $\mathcal{M}_Y$.
- XY-match: x_{i_ℓ} and y_{i_ℓ} are originally in X and Y, respectively.

Let X^* and Y^* denote optimal fillings for the quadruple $(X, Y, \mathcal{M}_X, \mathcal{M}_Y)$ of 2FLCS. If there exists at least one symbol, say, σ, in $\mathcal{M}_Y$ that does not appear in an optimal filling Y^*, then the length of $LCS(X^*, Y^* \oplus \langle\sigma\rangle)$ is equal to one of $LCS(X^*, Y^*)$, which implies that $Y^* \oplus \langle\sigma\rangle$ is another optimal filling. Similarly, if $\sigma' \in \mathcal{M}_X$ does not appear in X^*, then $X^* \oplus \langle\sigma'\rangle$ is another optimal filling. Therefore, without loss of generality, we assume that all the symbols in $\mathcal{M}_X$ and $\mathcal{M}_Y$ are inserted to the optimal fillings.

2.5 Known results on exact/approximation algorithms

Here, we summarize the previously known results on exact and approximation algorithms. For RBLCS, the following exact exponential-time algorithm is known:

▶ **Proposition 7** ([2]). *There is an $O(1.44225^n)$-time algorithm for RBLCS on two sequences X and Y, where $|X| = n$, $|Y| = m$, and $n \le m$, assuming that $m = O(poly(n))$.*

If $C_{occ}(\sigma) = 1$ for every symbol $\sigma \in \Sigma$, then we can design a faster exact algorithm:

▶ **Proposition 8** ([2]). *There is an $O(1.41422^n)$-time algorithm for RFLCS on two sequences X and Y, where $|X| = n$, $|Y| = m$, and $n \le m$, assuming that $m = O(poly(n))$.*

Furthermore, the following approximation algorithm is known for RFLCS:

▶ **Proposition 9** ([1]). *There is a polynomial-time occ_{max}-approximation algorithm for RFLCS on two sequences X and Y, where $occ_{max} = \max_{\sigma\in\Sigma}\{\min\{occ(X,\sigma), occ(Y,\sigma)\}\}$.*

For MRCS, the following exact exponential-time algorithm and the polynomial-time approximation algorithm are proposed in [11]:

▶ **Proposition 10** ([11]). *There is an $O(nm(t+1)^k)$-time algorithm for MRCS on two sequences X and Y, and a multiset $\mathcal{M}$, where t and k are the maximum multiplicity of $\mathcal{M}$ and the alphabet size $|\Sigma|$, respectively* [1].

▶ **Proposition 11** ([11]). *There is a polynomial-time $2\sqrt{\min\{n,m\}}$-approximation algorithm for MRCS on two sequences X and Y, and a multiset $\mathcal{M}$, where $|X| = n$ and $|Y| = m$.*

For 1FLCS, an FPT-algorithm parameterized by the number k of $X\mathcal{M}_Y$-matches in the optimal subsequence is known [6]. Note that k may be as large as the length of X, i.e., n.

▶ **Proposition 12** ([6]). *There is an $O(2^{O(k)}poly(n+m+|\mathcal{M}_Y|))$-time algorithm for 1FLCS on an input triple $(X, Y, \mathcal{M}_Y)$ if the number of $X\mathcal{M}_Y$-matches in $LCS(X, Y^*)$ is k.*

The following algorithm for 1FLCS runs in polynomial time if $|\Sigma|$ is a constant [6]:

▶ **Proposition 13** ([6]). *There is an $O(n^{|\Sigma|+2}m)$-time algorithm for 1FLCS on $(X, Y, \mathcal{M}_Y)$.*

The following approximability result is also known for 1FLCS:

▶ **Proposition 14** ([6]). *There is a polynomial-time $\frac{5}{3}$-approximation algorithm for 1FLCS.*

3 Polynomial-time equivalence of RBLCS and MRCS

In this section we show the polynomial-time equivalence between RBLCS and MRCS. First consider any optimal solution $Z_{\mathcal{M}}$ for an input $(X, Y, \mathcal{M})$ of MRCS. Recall that the objective function value of MRCS is $|\mathcal{M} \cap Z_{\mathcal{M}}|$. Hence, $|\mathcal{M} \cap Z_{\mathcal{M}}|$ can be regarded as the summation of occurrences of all the symbols in the solution. Furthermore, intuitively, the number $occ(\mathcal{M}, \sigma)$ of occurrences of every symbol $\sigma \in \mathcal{M}$ can be regarded as the occurrence constraint $C_{occ}(\sigma)$ of the solution for RBLCS, and vice versa. One sees that we can transform from/to a multiset $\mathcal{M}$ of symbols in Σ to/from an occurrence constraint C_{occ} of symbols in Σ such that $C_{occ}(\sigma) = occ(\mathcal{M}, \sigma)$ for every $\sigma \in \Sigma$ clearly in polynomial time; all we have to do is count the multiplicity/occurrences of every symbol in $\mathcal{M}$. Then, we can obtain the following theorem (see the journal version of this paper for its proof):

▶ **Theorem 15.** *Consider a pair of a multiset $\mathcal{M}$ in an input for MRCS and an occurrence constraint C_{occ} of symbols in Σ in an input for RBLCS such that $C_{occ}(\sigma) = occ(\mathcal{M}, \sigma)$ for every $\sigma \in \Sigma$. Then, the followings hold: (1) Given an optimal solution Z_R for an input (X, Y, C_{occ}) of RBLCS, we can obtain an optimal solution for an input $(X, Y, \mathcal{M})$ of MRCS in polynomial time. (2) Given an optimal solution $Z_{\mathcal{M}}$ for an input $(X, Y, \mathcal{M})$ of MRCS, we can obtain an optimal solution for an input (X, Y, C_{occ}) of RBLCS in polynomial time.*

[1] We remark that the time complexity shown in Theorem 3 of [11] is $O(nmt^k)$, but the correct one must be $O(nm(t+1)^k)$ because the algorithm has to store $t+1$ values from 0 through t for the maximum multiplicity. As described before, if $\mathcal{M} = \Sigma$, i.e., $t = 1$, then MRCS is essentially equivalent to RFLCS and thus MRCS is NP-hard. If we can solve MRCS with $t = 1$ in $O(nmt^k) = O(nm)$ time, then we can obtain P = NP.

4 Polynomial-time equivalence of RBLCS, 1FLCS, and 2FLCS

4.1 Proof tools

In this subsection we give some proof tools. The first tool reduces the numbers of XY-matches and $\mathcal{M}_X\mathcal{M}_Y$-matches in an output subsequence (see the journal version of this paper for its proof):

▶ **Lemma 16.** *Suppose that $(X, Y, \mathcal{M}_X, \mathcal{M}_Y)$ is an input for 2FLCS, and X^* and Y^* are optimal fillings of $(X, Y, \mathcal{M}_X, \mathcal{M}_Y)$. Also, suppose that the numbers of XY-matches, $\mathcal{M}_X\mathcal{M}_Y$-matches, $X\mathcal{M}_Y$-matches, and $\mathcal{M}_XY$-matches of some σ in the index-expression (I_{X^*}, I_{Y^*}) of X^* and Y^* are $\alpha > 0$, $\beta > 0$, $\zeta \geq 0$, and $\eta \geq 0$, respectively. Then, we can obtain in polynomial time another pair of optimal fillings X^{**} and Y^{**} such that (i) the numbers of XY-matches, $\mathcal{M}_X\mathcal{M}_Y$-matches, $X\mathcal{M}_Y$-matches, and $\mathcal{M}_XY$-matches of σ in the index-expression $(I_{X^{**}}, I_{Y^{**}})$ of X^{**} and Y^{**} are $\alpha - 1$, $\beta - 1$, $\zeta + 1$, and $\eta + 1$, respectively, and (ii) all the matches of any different symbol $\sigma' \neq \sigma$ do not change.*

If we use the above tool iteratively α-times for $\alpha \leq \beta$ (β-times for $\beta \leq \alpha$, resp.), then we can obtain so-called an "XY-match-free" ("$\mathcal{M}_X\mathcal{M}_Y$-match-free", resp.) output subsequence.

▶ **Lemma 17.** *Suppose that an input quadruple $(X, Y, \mathcal{M}_X, \mathcal{M}_Y)$ satisfies $occ(X, \sigma) > 0$ and $occ(\mathcal{M}_Y, \sigma) > 0$ for some $\sigma \in \Sigma$. Let $X = \langle x_1, \ldots, x_n \rangle$ and $\mathcal{M}_Y = \langle \psi_1, \ldots, \psi_\ell \rangle$. Then,*

$$L(X, Y, \mathcal{M}_X, \mathcal{M}_Y) = \max_{\sigma = x_i = \psi_j} L(X \setminus \langle i \rangle, Y, \mathcal{M}_X, \mathcal{M}_Y \setminus \langle j \rangle) + 1.$$

We can apply very similar arguments to the pair Y and $\mathcal{M}_X$, which gives:

▶ **Corollary 18.** *Suppose that an input quadruple $(X, Y, \mathcal{M}_X, \mathcal{M}_Y)$ satisfies $occ(Y, \sigma) > 0$ and $occ(\mathcal{M}_X, \sigma) > 0$ for some $\sigma \in \Sigma$. Let $Y = \langle y_1, \ldots, y_m \rangle$ and $\mathcal{M}_X = \langle \chi_1, \ldots, \chi_\ell \rangle$. Then,*

$$L(X, Y, \mathcal{M}_X, \mathcal{M}_Y) = \max_{\sigma = y_i = \chi_j} L(X, Y \setminus \langle i \rangle, \mathcal{M}_X \setminus \langle j \rangle, \mathcal{M}_Y) + 1.$$

The following lemma and corollary deal with the symbol additions to multisets:

▶ **Lemma 19.** *Let X^+ be a filling of X and $\mathcal{M}_X$, and let Y^+ be a filling of Y and $\mathcal{M}_Y$. Suppose that a common subsequence Z of X^+ and Y^+ satisfies $occ(Z, \sigma) < occ(Y^+, \sigma)$ for some symbol $\sigma \in \Sigma$. Then, we can find in polynomial time a new filling X^{++} of X and $\mathcal{M}_X \cup \{\sigma\}$ and a common subsequence Z' of X^{++} and Y^+ satisfying the following conditions: (1) $occ(Z, \sigma) + 1 = occ(Z', \sigma)$, and (2) for every σ' except for σ $occ(Z, \sigma') = occ(Z', \sigma')$.*

▶ **Corollary 20.** *Let X^+ be a filling of X and $\mathcal{M}_X$, and let Y^+ be a filling of Y and $\mathcal{M}_Y$. Suppose that a common subsequence Z of X^+ and Y^+ satisfies $occ(Z, \sigma) < occ(X^+, \sigma)$ for some symbol $\sigma \in \Sigma$. Then, we can find in polynomial time a new filling Y^{++} of Y and $\mathcal{M}_Y \cup \{\sigma\}$ and a common subsequence Z' of Y^{++} and X^+ satisfying the following conditions: (1) $occ(Z, \sigma) + 1 = occ(Z', \sigma)$, and (2) for every σ' except for σ, $occ(Z, \sigma') = occ(Z', \sigma')$.*

4.2 RBLCS and 1FLCS

In this subsection we show that 1FLCS is polynomially equivalent to RBLCS. Consider an input triple $(X, Y, \mathcal{M}_Y)$ of 1FLCS. In [12], Mincu and Popa observed that a filling-procedure of a symbol $\sigma \in \mathcal{M}_Y$ into Y to match some σ in X can be seen as a deleting-procedure of the matched σ from X [12]. Our basic ideas are based on their observation: Every symbol

$\sigma \in \mathcal{M}_Y$ can be matched to σ at any position in X without restrictions. After all σ's in $\mathcal{M}_Y$ are matched, the number of remaining unmatched σ's in X is $occ(X,\sigma) - occ(\mathcal{M}_Y,\sigma)$, which can be seen as the occurrence constraint $C_{occ}(\sigma)$ of the input (X, Y, C_{occ}) for RBLCS. In the following, we show that (i) from the input $(X, Y, \mathcal{M}_Y)$ for 1FLCS, we can construct the input (X, Y, C_{occ}) for RBLCS such that $C_{occ}(\sigma) = occ(X,\sigma) - occ(\mathcal{M}_Y,\sigma)$ for every $\sigma \in \Sigma$ in polynomial time, and vice versa; (ii) from an optimal solution of the former problem, we can construct an optimal solution of the latter problem in polynomial time, and vice versa.

Consider an input triple $(X, Y, \mathcal{M}_Y)$ of 1FLCS and a feasible solution Z_{1F}. Then, for every symbol σ, $occ(Z_{1F},\sigma) \leq occ(X,\sigma)$ holds. If $occ(X,\sigma) < occ(\mathcal{M}_Y,\sigma)$, then $occ(\mathcal{M}_Y,\sigma) - occ(X,\sigma)$ σ's are clearly redundant. If the input $(X, Y, \mathcal{M}_Y)$ of 1FLCS satisfies $occ(X,\sigma) \geq occ(\mathcal{M}_Y,\sigma)$ for every $\sigma \in \Sigma$, then we call $(X, Y, \mathcal{M}_Y)$ the *standard* input. Without loss of generality, we assume that every input of 1FLCS is standard.

▶ **Lemma 21.** *Suppose that a triple $(X, Y, \mathcal{M}_Y)$ is a standard input for 1FLCS, Y^* is an optimal filling, and Z is the longest common subsequence of X and Y^*. Then, for every σ in Σ, $occ(Z,\sigma) \geq occ(\mathcal{M}_Y,\sigma)$ is satisfied.*

Proof. Let $X = \langle x_1, \ldots, x_n\rangle$, $Y^* = \langle y^*_1, \ldots, y^*_m\rangle$, and $Z = \langle z_1, \ldots, z_\ell\rangle = \langle x_{i_1}, \ldots, x_{i_\ell}\rangle = \langle y^*_{j_1}, \ldots, y^*_{j_\ell}\rangle$ (i.e., $i_p < i_{p+1}$ and $j_p < j_{p+1}$ hold for every $1 \leq p \leq \ell - 1$). Since the input is standard, for every σ, $occ(\mathcal{M}_Y,\sigma) \leq occ(X,\sigma)$ holds.

Now suppose for the purpose of obtaining a contradiction that there exists at least one symbol, say, σ', $occ(Z,\sigma') < occ(\mathcal{M}_Y,\sigma') \leq occ(X,\sigma')$ holds. Since $occ(Z,\sigma') < occ(X,\sigma')$ holds, we can find an index q such that the qth symbol x_q in X is σ' but q is not in $I_X = \langle i_1, i_2, \ldots, i_\ell\rangle$. First, we assume that $i_p < q < i_{p+1}$ holds for some p where $1 \leq p \leq \ell - 1$. Then, we construct a new sequence $Z' = \langle x_{i_1}, \ldots, x_{i_p}\rangle \oplus \langle\sigma'\rangle \oplus \langle x_{i_{p+1}} \ldots, x_{i_\ell}\rangle$ of length $\ell + 1$. If $q < i_1$ ($i_\ell < q$, resp.), then we insert σ' to the head position, i.e., $Z' = \langle\sigma'\rangle \oplus \langle x_{i_1}, \ldots, x_{i_\ell}\rangle$ (to the tail position, i.e., $Z' = \langle x_{i_1}, \ldots, x_{i_\ell}\rangle \oplus \langle\sigma'\rangle$, resp.). Moreover, since $occ(Z,\sigma') < occ(\mathcal{M}_Y,\sigma')$, we can find an index q' such that the q'th symbol $y_{q'}$ inserted into Y^* is σ' but q' is not in $I_{Y^*} = \langle j_1, j_2, \ldots, j_\ell\rangle$. Then we construct a new filling Y^{**} as follows: (1) First remove the q'th symbol $y_{q'}$ ($= \sigma'$) from Y^*, and then (2) insert $y_{q'}$ right after y_{j_p} of Y^*. Note that the $(p+1)$st symbol in the new sequence Z' is σ'. It follows that $LCS(X, Y^{**}) = Z'$ and thus we can obtain the sequence of length $\ell + 1$ from $(X, Y, \mathcal{M}_Y)$, which is a contradiction. Therefore, for all σ in Σ, $occ(Z,\sigma) \geq occ(\mathcal{M}_Y,\sigma)$ holds. ◀

Consider an input triple $(X, Y, \mathcal{M}_Y)$ of 1FLCS and its optimal solution Z_{1F}. Suppose that there is a symbol σ such that $occ(X,\sigma) > occ(Y,\sigma) + occ(\mathcal{M}_Y,\sigma)$. Let $\ell = occ(X,\sigma) - (occ(Y,\sigma) + occ(\mathcal{M}_Y,\sigma)) \geq 0$. Then, at least ℓ σ's in X do not appear in Z_{1F}. Let $\mathcal{S}_\sigma$ be a multiset of ℓ σ's. Now, suppose that for a new triple $(X, Y, \mathcal{M}_Y \cup \mathcal{S}_\sigma)$, we can obtain an optimal solution Z. Then, the length of Z must be equal to $|Z_{1F}| + \ell$. Moreover, by removing ℓ σ's in $\mathcal{S}_\sigma$ from Z, we can easily find the original optimal solution Z_{1F} for $(X, Y, \mathcal{M}_Y)$. For every symbol σ' in Σ satisfying $occ(X,\sigma') > occ(Y,\sigma') + occ(\mathcal{M}_Y,\sigma')$, the similar discussion as the above can be applied. Let $\mathcal{S} = \bigcup_{\sigma' : occ(X,\sigma') > occ(Y,\sigma') + occ(\mathcal{M}_Y,\sigma')} \mathcal{S}_{\sigma'}$. If we are given the triple $(X, Y, \mathcal{M}_Y \cup \mathcal{S})$, then by finding its optimal solution Z' first, and then removing all the symbols in $\mathcal{S}$ from Z', we obtain Z_{1F}. In the following we call the triple $(X, Y, \mathcal{M}_Y \cup \mathcal{S})$ by merging $\mathcal{S}$ to $\mathcal{M}_Y$ an *extended triple*. If the triple $(X, Y, \mathcal{M}_Y)$ of 1FLCS is extended and satisfies $occ(X,\sigma) \geq occ(\mathcal{M}_Y,\sigma)$ for every $\sigma \in \Sigma$ then it is called *ex-standard*. To simplify the discussion, we assume that every input triple $(X, Y, \mathcal{M}_Y)$ of 1FLCS is always ex-standard.

The following lemma is quite trivial but plays an important role:

▶ **Lemma 22.** *(1) Suppose that an input triple $(X, Y, \mathcal{M}_Y)$ for 1FLCS is ex-standard. Then, we can construct a standard input triple (X, Y, C_{occ}) for RBLCS satisfying $C_{occ}(\sigma) = occ(X, \sigma) - occ(\mathcal{M}_Y, \sigma)$ for every $\sigma \in \Sigma$ in polynomial time. (2) Suppose that an input triple (X, Y, C_{occ}) for RBLCS is standard. Then, we can construct an ex-standard input triple $(X, Y, \mathcal{M}_Y)$ for 1FLCS satisfying $occ(\mathcal{M}_Y, \sigma) = occ(X, \sigma) - C_{occ}(\sigma)$ for every $\sigma \in \Sigma$ in polynomial time.*

Proof. (1) Since the triple $(X, Y, \mathcal{M}_Y)$ is ex-standard, $occ(X, \sigma) - occ(\mathcal{M}_Y, \sigma) \geq 0$ for every σ. Therefore, we can always obtain the valid occurrence constraint such that $C_{occ}(\sigma) = occ(X, \sigma) - occ(\mathcal{M}_Y, \sigma)$ for every σ. Furthermore, since $(X, Y, \mathcal{M}_Y)$ is ex-standard, $C_{occ}(\sigma) = occ(X, \sigma) - occ(\mathcal{M}_Y, \sigma) \leq occ(Y, \sigma)$. It follows that $C_{occ}(\sigma) \leq \min\{occ(X, \sigma), occ(Y, \sigma)\}$. Hence, the triple (X, Y, C_{occ}) must be standard for RBLCS. (2) Since the triple (X, Y, C_{occ}) is standard, $C_{occ}(\sigma) \leq \min\{occ(X, \sigma), occ(Y, \sigma)\}$. Therefore, we can always obtain the valid multiset $\mathcal{M}_Y$ such that $occ(\mathcal{M}_Y, \sigma) = occ(X, \sigma) - C_{occ}(\sigma) \geq 0$ for every σ. ◀

▶ **Lemma 23.** *Consider an ex-standard input $(X, Y, \mathcal{M}_Y)$ for 1FLCS and a standard input (X, Y, C_{occ}) for RBLCS such that $C_{occ}(\sigma) = occ(X, \sigma) - occ(\mathcal{M}_Y, \sigma)$ holds for every $\sigma \in \Sigma$. Let $Z_F = LCS(X, Y, \mathcal{M}_Y)$ and Y^* be an optimal filling for 1FLCS. Also, let $Z_R = LCS(X, Y, C_{occ})$ be an optimal solution for RBLCS. Then, $|Z_R| + |\mathcal{M}_Y| = |Z_F|$ holds.*

Proof. First, from Lemma 22, we always find a pair of triples $(X, Y, \mathcal{M}_Y)$ and (X, Y, C_{occ}) such that the former and the latter are the ex-standard input for 1FLCS and the standard input for RBLCS satisfying $C_{occ}(\sigma) = occ(X, \sigma) - occ(\mathcal{M}_Y, \sigma)$ for every $\sigma \in \Sigma$, respectively.

(1) We first show that $|Z_F| \leq |Z_R| + |\mathcal{M}_Y|$ holds. Let $X = \langle x_1, \ldots, x_n \rangle$, $Y = \langle y_1, \ldots, y_m \rangle$, and $M_Y = \langle \psi_1, \ldots, \psi_\ell \rangle$, where M_Y is the sequence-expression of $\mathcal{M}_Y$. By the assumption that $(X, Y, \mathcal{M}_Y)$ is ex-standard, there exists a sequence $\langle i_1, i_2, \ldots, i_\ell \rangle$ of indices of X satisfying $L(X, Y, \mathcal{M}_Y) = L(X \setminus \langle i_1, \ldots, i_\ell \rangle, Y, \emptyset) + \ell$, by regarding $L(X, Y, \emptyset, \mathcal{M}_Y)$ as $L(X, Y, \mathcal{M}_Y)$, and by using the formula in Lemma 17 recursively. Since $\mathcal{M}_Y = \emptyset$, $L(X \setminus \langle i_1, \ldots, i_\ell \rangle, Y, \emptyset)$ is clearly equal to the length of the longest common subsequence Z' of $X \setminus \langle i_1, \ldots, i_\ell \rangle$ and Y. Therefore, $|Z_F| = |Z'| + |\mathcal{M}_Y|$. Note that Z' is a common subsequence of the original X and Y and satisfies the following for every σ:

$$C_{occ}(\sigma) = occ(X, \sigma) - occ(\mathcal{M}_Y, \sigma) = occ(X \setminus \langle i_1, \ldots, i_\ell \rangle, \sigma) \geq occ(Z', \sigma).$$

That is, every symbol in Z' satisfies the occurrence constraint C_{occ} of RBLCS, which implies that $|Z'| \leq |Z_R|$. As a result, $|Z_F| = |Z'| + |\mathcal{M}_Y| \leq |Z_R| + |\mathcal{M}_Y|$ holds.

(2) Next, we show that $|Z_R| + |\mathcal{M}_Y| \leq |Z_F|$. Recall that for every σ, $occ(Z_R, \sigma) \leq C_{occ}(\sigma) = occ(X, \sigma) - occ(\mathcal{M}_Y, \sigma)$ is satisfied. Here, from the viewpoint of 1FLCS, we can obtain a longer sequence than Z_R by filling symbols of $\mathcal{M}_Y$ into Y. Suppose that Z_R is a common subsequence for RBLCS on (X, Y, C_{occ}) and $(X, Y, \emptyset)$ is an input triple for 1FLCS. From Lemma 19, by setting a multiset $\mathcal{M}'_Y = \{\sigma\}$ and filling σ into Y as matched with some σ in X, we can obtain a common subsequence Z_1 such that $|Z_1| = |Z_R| + 1$, $occ(Z_1, \sigma) = occ(Z_R, \sigma) + 1$, and $occ(Z_1, \sigma') = occ(Z_R, \sigma')$ for every σ' except for σ. By repeating the merge $\mathcal{M}'_Y \cup \{\sigma\}$ and the filling of σ $occ(\mathcal{M}_Y, \sigma)$-times for every $\sigma \in \Sigma$, we can eventually obtain $\mathcal{M}_Y$, the filling of Y and $\mathcal{M}_Y$, and a common subsequence Z satisfying $|Z| = |Z_R| + \sum_{\sigma \in \Sigma} occ(\mathcal{M}_Y, \sigma) = |Z_R| + |\mathcal{M}_Y|$. Since Z_F is the longest, $|Z| \leq |Z_F|$. Hence, $|Z_R| + |\mathcal{M}_Y| = |Z| \leq |Z_F|$ holds.

From (1) and (2), $|Z_R| + |\mathcal{M}_Y| = |Z_F|$. This completes the proof. ◀

▶ **Theorem 24.** *Consider an ex-standard input $(X, Y, \mathcal{M}_Y)$ for 1FLCS and a standard input (X, Y, C_{occ}) for RBLCS such that $C_{occ}(\sigma) = occ(X, \sigma) - occ(\mathcal{M}_Y, \sigma)$ holds for every $\sigma \in \Sigma$. Let $Z_F = LCS(X, Y, \mathcal{M}_Y)$ and Y^* be an optimal filling for 1FLCS. Also, let $Z_R = LCS(X, Y, C_{occ})$ be an optimal solution for RBLCS. Then, the followings hold: (1) Given an optimal solution Z_R for RBLCS, we can obtain an optimal solution for 1FLCS in polynomial time. (2) Given an optimal filling Y^* for 1FLCS, we can obtain an optimal solution for RBLCS in polynomial time.*

Proof. Consider two sequences X and Y, a multiset $\mathcal{M}_Y$, and an occurrence constraint C_{occ} such that $C_{occ}(\sigma) = occ(X, \sigma) - occ(\mathcal{M}_Y, \sigma)$ holds for every $\sigma \in \Sigma$.

(1) Suppose that the optimal solution Z_R for RBLCS is now given. From Lemma 23, every optimal solution for 1FLCS is of length $|Z_R| + |\mathcal{M}_Y|$. Hence, it is enough to prove that we can obtain an optimal filling Y^* of Y and $\mathcal{M}_Y$ from Z_R and a common subsequence Z_F of X and Y^* such that $|Z_R| + |\mathcal{M}_Y| = |Z_F|$ in polynomial time. As seen in the proof of Lemma 23, by repeating the merge $\mathcal{M}'_Y = \mathcal{M}'_Y \cup \{\sigma\}$ and the filling of σ $occ(\mathcal{M}_Y, \sigma)$-times for every $\sigma \in \Sigma$, we eventually obtain Y^* and Z_F satisfying $|Z_F| = |Z_R| + \sum_{\sigma \in \Sigma} occ(\mathcal{M}_Y, \sigma) = |Z_R| + |\mathcal{M}_Y|$. The total number of iterations is $|\mathcal{M}_Y|$. Since each iteration works in polynomial time as shown in Lemma 23, Y^* and Z_F of 1FLCS can be obtained in polynomial time.

(2) Suppose that the optimal filling Y^* is now given. The longest common subsequence Z_F of X and Y^*, and its index-expression (I_X, I_{Y^*}) can be obtained in polynomial time. From Lemma 21, $occ(Z_F, \sigma) \geq occ(\mathcal{M}_Y, \sigma)$ holds for every $\sigma \in \Sigma$. Therefore, we can find $|Z_F| - |\mathcal{M}_Y|$ XY-matches in (I_X, I_{Y^*}). Letting z_ℓ be the symbol of the ℓth XY-match $(1 \leq \ell \leq |Z_F| - |\mathcal{M}_Y|)$, we construct the sequence $Z_F^- = \langle z_1, z_2, \ldots, z_{|Z_F| - |\mathcal{M}_Y|} \rangle$ of length $|Z_F| - |\mathcal{M}_Y|$. Note that Z_F^- must be a common subsequence of X and Y. Moreover, Z_F^- satisfies the occurrence constraint $C_{occ}(\sigma) = occ(X, \sigma) - occ(\mathcal{M}_Y, \sigma) \geq occ(Z_F, \sigma) - occ(\mathcal{M}_Y, \sigma)$ for every $\sigma \in \Sigma$. Since $|Z_F^-| = |Z_F| - |\mathcal{M}_Y|$, Z_F^- is an optimal solution for RBLCS from Lemma 23. The construction of Z_F^- can be easily executed by scanning the index-expression (I_X, I_{Y^*}) and thus it can be done in polynomial time. ◀

4.3 RBLCS and 2FLCS

In this subsection we consider the polynomial-time equivalence between 2FLCS and RBLCS. Since 1FLCS on $(X, Y, \mathcal{M}_Y)$ is equivalent to 2FLCS on $(X, Y, \emptyset, \mathcal{M}_Y)$, 1FLCS can be solved by using any algorithm for 2FLCS. From the polynomial-time equivalence between 1FLCS and RBLCS in the previous subsection, RBLCS can also be solved by the same algorithm with some extra polynomial-time calculations. Therefore, to establish the equivalence between RBLCS and 2FLCS, only one direction remains to be proved. To do so, we first give a pair of two inputs $(X, Y, \mathcal{M}_X, \mathcal{M}_Y)$ for 2FLCS and (X, Y, C_{occ}) for RBLCS. Then, we show that given an optimal solution Z_R of RBLCS on (X, Y, C_{occ}), we can obtain optimal fillings X^* and Y^* of 2FLCS on $(X, Y, \mathcal{M}_X \mathcal{M}_Y)$ in polynomial time.

▶ **Lemma 25.** *Suppose that an input $(X, Y, \mathcal{M}_X, \mathcal{M}_Y)$ of 2FLCS satisfies $occ(X, \sigma) = p < occ(\mathcal{M}_Y, \sigma) = q$ and $\min\{occ(\mathcal{M}_X, \sigma), occ(Y, \sigma) + q - p)\} = \lambda \geq 0$ for some positive integers p and q. Then the following holds:*

$$L(X, Y, \mathcal{M}_X, \mathcal{M}_Y) \leq L(X, Y, \mathcal{M}_X \setminus \{\sigma^*\}, \mathcal{M}_Y \setminus \{\sigma^{q-p}\}) + \lambda$$

Proof. Suppose that an input quadruple $(X, Y, \mathcal{M}_X, \mathcal{M}_Y)$ of 2FLCS satisfies $occ(X, \sigma) = p < occ(\mathcal{M}_Y, \sigma) = q$. If we set $X = \langle x_1, \ldots, x_n \rangle$ and $\mathcal{M}_Y = \langle \psi_1, \ldots, \psi_\ell \rangle$, and apply the recursive formula in Lemma 17 recursively, then there exist two sequences of $\langle i_1, \ldots, i_p \rangle$ and $\langle j_1, \ldots, j_p \rangle$ of indices such that $\sigma = x_{i_r} = \psi_{j_r}$ for every $1 \leq r \leq p$. Therefore, we obtain

$$L(X, Y, \mathcal{M}_X, \mathcal{M}_Y) = L(X \setminus \langle i_1, \ldots, i_p \rangle, Y, \mathcal{M}_X, \mathcal{M}_Y \setminus \langle j_1, \ldots, j_p \rangle) + p.$$

Suppose that X^+ and Y^+ are optimal fillings of $(X \setminus \langle i_1, \ldots, i_p \rangle, Y, \mathcal{M}_X, \mathcal{M}_Y \setminus \langle j_1, \ldots, j_p \rangle)$. Then, $occ(X^+, \sigma) \leq occ(\mathcal{M}_X, \sigma)$ since $\sigma \notin X \setminus \langle i_1, \ldots, i_p \rangle$ and $occ(Y^+, \sigma) \leq occ(Y, \sigma) + q - p$. Therefore, we obtain $occ(Z, \sigma) \leq \min\{occ(\mathcal{M}_X, \sigma), occ(Y, \sigma) + q - p\}$ for $Z = LCS(X^+, Y^+)$. Now, we set $\min\{occ(\mathcal{M}_X, \sigma), occ(Y, \sigma) + q - p\} = \lambda$. Then, we have:

$$L(X \setminus \langle i_1, \ldots, i_p \rangle, Y, \mathcal{M}_X \setminus \{\sigma^*\}, \mathcal{M}_Y \setminus \{\sigma^*\}) + \lambda \geq L(X \setminus \langle i_1, \ldots, i_p \rangle, Y, \mathcal{M}_X, \mathcal{M}_Y \setminus \langle j_1, \ldots, j_p \rangle).$$

Suppose that a sequence $J^+ = \langle j_1, \ldots, j_q \rangle$ of indices satisfies that $\psi_{j_{r'}} = \sigma$ for every $1 \leq r' \leq q$. Then, we obtain:

$$\begin{aligned} L(X, Y, \mathcal{M}_X, \mathcal{M}_Y) =& L(X \setminus \langle i_1, \ldots, i_p \rangle, Y, \mathcal{M}_X, \mathcal{M}_Y \setminus \langle j_1, \ldots, j_p \rangle) + p \\ \leq& L(X \setminus \langle i_1, \ldots, i_p \rangle, Y, \mathcal{M}_X \setminus \{\sigma^*\}, \mathcal{M}_Y \setminus \{\sigma^*\}) + \lambda + p \\ =& L(X, Y, \mathcal{M}_X \setminus \{\sigma^*\}, \mathcal{M}_Y \setminus \langle j_{p+1}, \ldots, j_q \rangle) + \lambda. \end{aligned}$$

This completes the proof. ◀

▶ **Theorem 26.** *Suppose that an input quadruple $(X, Y, \mathcal{M}_X, \mathcal{M}_Y)$ of 2FLCS satisfies $occ(X, \sigma) = p < occ(\mathcal{M}_Y, \sigma) = q$ for some positive integers p and q, and optimal fillings X_1^+ and Y_1^+ of $(X, Y, \mathcal{M}_X \setminus \{\sigma^*\}, \mathcal{M}_Y \setminus \{\sigma^{q-p}\})$ are given. Then, optimal fillings X_2^+ and Y_2^+ of an input quadruple $(X, Y, \mathcal{M}_X, \mathcal{M}_Y)$ can be obtained in polynomial time.*

Proof. Suppose that Z_1 is the longest common subsequence of X_1^+ and Y_1^+ such that the index-expression of Z_1 is (I, J), where $I = \langle i_1, \ldots, i_k \rangle$ and $J = \langle j_1, \ldots, j_k \rangle$. Also suppose that Z_2 is the longest common subsequence of X_2^+ and Y_2^+. From Lemma 25, $|Z_1| + \lambda \geq |Z_2|$ holds, where $\min\{occ(\mathcal{M}_X, \sigma), occ(Y, \sigma) + q - p\} = \lambda$.

Now suppose that $X_1^+ = \langle x_1, \ldots, x_n \rangle$ and $Y_1^+ = \langle y_1, \ldots, y_m \rangle$. Also suppose that $Y_2^+ = Y_1^+ \oplus \langle \overbrace{\sigma, \ldots, \sigma}^{q-p} \rangle$. One can see that $occ(Y_2^+, \sigma) = occ(Y, \sigma) + q$, $occ(Z_1, \sigma) \leq p < occ(\mathcal{M}_Y, \sigma) = q$, and Z_1 is a common subsequence of X_1^+ and Y_2^+. Therefore, by applying the formula in (1) of Lemma 19 $\min\{occ(\mathcal{M}_X, \sigma), occ(Y, \sigma) + q - p\}$-times, we can get the target sequence X_2^+ in polynomial time. ◀

It is important to note that $(X, Y, \mathcal{M}_X \setminus \{\sigma^*\}, \mathcal{M}_Y \setminus \{\sigma^{q-p}\})$ does not satisfy both $occ(X, \sigma) < occ(\mathcal{M}_Y, \sigma)$ and $occ(Y, \sigma) < occ(\mathcal{M}_X, \sigma)$. For Y and $\mathcal{M}_X$, we have:

▶ **Corollary 27.** *Suppose that an input quadruple $(X, Y, \mathcal{M}_X, \mathcal{M}_Y)$ of 2FLCS satisfies $occ(Y, \sigma) = p < occ(\mathcal{M}_X, \sigma) = q$ for some positive integers p and q, and optimal fillings X_1^+ and Y_1^+ of $(X, Y, \mathcal{M}_X \setminus \{\sigma^{q-p}\}, \mathcal{M}_Y \setminus \{\sigma^*\})$ are given. Then, optimal fillings X_2^+ and Y_2^+ of an input quadruple $(X, Y, \mathcal{M}_X, \mathcal{M}_Y)$ can be obtained in polynomial time.*

From Theorem 26 and Corollary 27, any input can be reduced to the quadruple $(X, Y, \mathcal{M}_X, \mathcal{M}_Y)$ such that for every σ, both $occ(X, \sigma) \geq occ(\mathcal{M}_Y, \sigma)$ and $occ(Y, \sigma) \geq occ(\mathcal{M}_X, \sigma)$ are satisfied. Therefore, if the input $(X, Y, \mathcal{M}_X, \mathcal{M}_Y)$ of 2FLCS satisfies both $occ(X, \sigma) \geq occ(\mathcal{M}_Y, \sigma)$ and $occ(Y, \sigma) \geq occ(\mathcal{M}_X, \sigma)$ for every $\sigma \in \Sigma$, then we call $(X, Y, \mathcal{M}_X, \mathcal{M}_Y)$ the *standard* input.

▶ **Theorem 28.** *For a standard input quadruple $(X, Y, \mathcal{M}_X, \mathcal{M}_Y)$, consider an occurrence constraint C_{occ} such that $C_{occ}(\sigma) = \min\{occ(X,\sigma) - occ(\mathcal{M}_Y,\sigma), occ(Y,\sigma) - occ(\mathcal{M}_X,\sigma)\}$ holds for every $\sigma \in \Sigma$. Then, the triple (X, Y, C_{occ}) must be standard for RBLCS. If an optimal solution Z_R of RBLCS on (X, Y, C_{occ}) is given, then we can obtain optimal fillings X^* and Y^* of 2FLCS on a standard input quadruple $(X, Y, \mathcal{M}_X, \mathcal{M}_Y)$ in polynomial time.*

Proof. Suppose that the input $(X, Y, \mathcal{M}_X, \mathcal{M}_Y)$ of 2FLCS is standard, $|\mathcal{M}_X| = p$, and $|\mathcal{M}_Y| = q$. Let $X = \langle x_1, \ldots, x_n\rangle$ and $Y = \langle y_1, \ldots, y_m\rangle$. Then, by applying the arguments of Lemma 17 and Corollary 18 to all the symbols recursively, we can obtain the sequences $\langle i_1, \ldots, i_q\rangle$ and $\langle j_1, \ldots, j_p\rangle$ of different indices that satisfy the following:

$$L(X, Y, \mathcal{M}_X, \mathcal{M}_Y) = L(X \setminus \langle i_1, \ldots, i_q\rangle, Y \setminus \langle j_1, \ldots, j_p\rangle, \emptyset, \emptyset) + p + q.$$

One can verify that for the input $(X \setminus \langle i_1, \ldots, i_q\rangle, Y \setminus \langle j_1, \ldots, j_p\rangle, \emptyset, \emptyset)$ of 2FLCS, the longest common subsequence of $X \setminus \langle i_1, \ldots, i_q\rangle$ and $Y \setminus \langle j_1, \ldots, j_p\rangle$ is clearly an optimal solution of the classical LCS. Let Z' be such a sequence. Here, note that for every $\sigma \in \Sigma$, we can obtain:

$$occ(X \setminus \langle i_1, \ldots, i_q\rangle, \sigma) = occ(X,\sigma) - occ(\mathcal{M}_Y,\sigma), \text{ and}$$
$$occ(Y \setminus \langle j_1, \ldots, j_p\rangle, \sigma) = occ(Y,\sigma) - occ(\mathcal{M}_X,\sigma).$$

Therefore, we have:

$$occ(Z',\sigma) \le \min\{occ(X,\sigma) - occ(\mathcal{M}_Y,\sigma), occ(Y,\sigma) - occ(\mathcal{M}_X,\sigma)\}.$$

Since Z' is a common subsequence of X and Y, Z' is a feasible solution of RBLCS on (X, Y, C_{occ}). Therefore, $|Z_R| \ge |Z'|$ holds. It follows that $|Z_R| + |\mathcal{M}_X| + |\mathcal{M}_Y| \ge |Z'| + |\mathcal{M}_X| + |\mathcal{M}_Y| = L(X, Y, \mathcal{M}_X, \mathcal{M}_Y)$.

As for Z_R, $occ(Z_R,\sigma) \le C_{occ}(\sigma) = \min\{occ(X,\sigma) - occ(\mathcal{M}_Y,\sigma), occ(Y,\sigma) - occ(\mathcal{M}_X,\sigma)\}$ holds for every σ. Therefore, by applying Lemma 19 $occ(\mathcal{M}_X,\sigma)$-times for every symbol $\sigma \in \Sigma$, we can construct in polynomial time the filling X^+ of X and $\mathcal{M}_X$, and a common subsequence Z_1 of X^+ and Y such that $|Z_1| = |Z_R| + |\mathcal{M}_X|$ and $occ(Z_1,\sigma) \le C_{occ}(\sigma) + |\mathcal{M}_X|$.

Note that for every σ, $occ(X^+,\sigma) = occ(X,\sigma) + occ(\mathcal{M}_X,\sigma)$ and $occ(Z_1,\sigma) \le occ(X,\sigma) - occ(\mathcal{M}_Y,\sigma) + occ(\mathcal{M}_X,\sigma)$ hold. Hence, by applying Corollary 20 $occ(\mathcal{M}_Y,\sigma)$-times for every symbol σ, we can construct in polynomial time the filling Y^+ of Y and $\mathcal{M}_Y$, and a common subsequence Z_2 of X^+ and Y^+ such that $|Z_2| = |Z_1| + |\mathcal{M}_Y| = |Z_R| + |\mathcal{M}_X| + |\mathcal{M}_Y|$. Recall that $|Z_R| + |\mathcal{M}_X| + |\mathcal{M}_Y| \ge L(X, Y, \mathcal{M}_X, \mathcal{M}_Y)$. Therefore, $|Z_2| \ge L(X, Y, \mathcal{M}_X, \mathcal{M}_Y)$ holds.

As a result, X^+ and Y^+ are optimal fillings of 2FLCS on $(X, Y, \mathcal{M}_X, \mathcal{M}_Y)$ and those can be obtained in polynomial time if Z_R is given. This completes the proof. ◀

5 $O(1.41422^n)$-time exact algorithm for RBLCS

In [2], a dynamic programming (DP) based algorithm for RBLCS was provided and it was explicitly proved that its running time is $O(1.44255^n)$. In this section we improve the running time from $O(1.44255^n)$ to $O(1.41422^n)$, but give only the basic ideas here. Further details can be found in the journal version of this paper.

Now, let us consider the original LCS and its typical DP-based algorithm. Let $L(i, j)$ denote the length of a longest common subsequence of the ith prefix $X_{1..i}$ of X and the jth prefix $Y_{1..j}$ of Y. In the process of execution, each value of $L(i, j)$ is computed and is stored into a two-dimensional DP-table L_0 of size $(n + 1) \times (m + 1)$. For more details, e.g., see [7].

For RBLCS, the previous DP-based algorithm proposed in [2] has to store not only the length of the subsequence Z, but also the occurrence $occ(Z,\sigma)$ of every σ in Z not to break the occurrence constraint $C_{occ}(\sigma)$. To store the occurrences, the algorithm introduces an

occurrence vector $\boldsymbol{v}$. Let $L(i, j, \boldsymbol{v})$ be the length of a repetition-bounded longest common subsequence of $X_{1..i}$ and $Y_{1..j}$ satisfying the occurrence vector $\boldsymbol{v}$, i.e., the length of the longest subsequence which does not break the occurrence constraint. Then, each value of $L(i, j, \boldsymbol{v})$ is stored into a three-dimensional DP-table L_1 of size $(n+1) \times (m+1) \times \prod_\sigma (C_{occ}(\sigma)+1)$. In [2], the authors showed that the table size of L_1 is bounded above by $O(1.44255^n)$.

Our new DP-based algorithm prepares a smaller DP-table of size $(n+1) \times (m+1) \times \prod_\sigma (\min\{C_{occ}(\sigma), occ(X,\sigma) - C_{occ}(\sigma)\}+1)$. One can show that the DP-table size is reduced to $O(1.41422^n)$:

▶ **Theorem 29.** *There is an $O(1.41422^n)$-time DP-based algorithm to solve RBLCS for two input sequences X and Y, where $|X| = n$, $|Y| = m = O(poly(n))$, and $|X| \leq |Y|$.*

Recall that all reductions in the previous sections preserve X and Y. By our polynomial-time equivalences, we obtain the following corollary:

▶ **Corollary 30.** *MRCS, 1FLCS, and 2FLCS can be solved in $O(1.41422^n)$ time.*

6 A polynomial-time 2-approximation algorithm for 2FLCS

In this section, we give a polynomial-time algorithm for 2FLCS and show that its approximation ratio is bounded above by two by using the proof tools introduced in Section 4.1.

Algorithm. Suppose that a standard input quadruple $(X, Y, \mathcal{M}_X, \mathcal{M}_Y)$ is given, i.e., $occ(X,\sigma) \geq occ(\mathcal{M}_Y,\sigma)$ and $occ(Y,\sigma) \geq occ(\mathcal{M}_X,\sigma)$ are satisfied. Let $X = \{x_1, \ldots, x_n\}$ and $Y = \{y_1, \ldots, y_m\}$. Here is an outline of our algorithm `ALG`: (Step 1) Let $X_b = \varepsilon$ and $Y_f = \varepsilon$ be two empty sequences. (1-1) While scanning from x_1 to x_n of X, if the ith symbol x_i in X matches a symbol, say, σ_y, in $\mathcal{M}_Y$, then x_i $(= \sigma_y)$ is concatenated to Y_f, i.e., $Y_f = Y_f \oplus \langle \sigma_y \rangle$ and removed from $\mathcal{M}_Y$. Then, obtain a filling $Y_2 = Y_f \oplus Y$ of Y and $\mathcal{M}_Y$. Similarly, (1-2) while scanning from y_1 to y_m of Y, if the ith symbol y_i in Y matches a symbol, say, σ_x, in $\mathcal{M}_X$, then y_i $(= \sigma_x)$ is concatenated to X_b, i.e., $X_b = X_b \oplus \langle \sigma_x \rangle$ and removed from $\mathcal{M}_X$. Then, obtain a filling $X_2 = X \oplus X_b$ of X and $\mathcal{M}_X$ (n.b., not $X_b \oplus X$). (Step 2) Obtain a longest common subsequence Z of two fillings X^+ and Y^+. (Step 3) Output a solution triple (X^+, Y^+, Z). See Algorithm 1 for the detailed description of `ALG`.

▶ **Theorem 31.** *Algorithm `ALG` is a polynomial-time 2-approximation algorithm for 2FLCS on a standard input quadruple $(X, Y, \mathcal{M}_X, \mathcal{M}_Y)$.*

Proof. Suppose that the input $(X, Y, \mathcal{M}_X, \mathcal{M}_Y)$ of 2FLCS is standard. Let $X = \langle x_1, \ldots, x_n \rangle$ and $Y = \langle y_1, \ldots, y_m \rangle$. Then, by applying the arguments of Lemma 17 and Corollary 18 to all the symbols recursively, we can obtain the sequences $\langle i_1, \ldots, i_{|\mathcal{M}_Y|} \rangle$ and $\langle j_1, \ldots, j_{|\mathcal{M}_X|} \rangle$ of different indices that satisfy the following:

$$L(X, Y, \mathcal{M}_X, \mathcal{M}_Y) = L(X \setminus \langle i_1, \ldots, i_{|\mathcal{M}_Y|} \rangle, Y \setminus \langle j_1, \ldots, j_{|\mathcal{M}_X|} \rangle, \emptyset, \emptyset) + |\mathcal{M}_Y| + |\mathcal{M}_X|.$$

Clearly, the first term $L(X \setminus \langle i_1, \ldots, i_{|\mathcal{M}_Y|} \rangle, Y \setminus \langle j_1, \ldots, j_{|\mathcal{M}_X|} \rangle, \emptyset, \emptyset)$ of the right-hand side is at most $L(X, Y)$ since $X \setminus \langle i_1, \ldots, i_{|\mathcal{M}_Y|} \rangle$ and $Y \setminus \langle j_1, \ldots, j_{|\mathcal{M}_X|} \rangle$ are subsequences of X and Y, respectively. Therefore, the longest length OPT of 2FLCS is at most $L(X, Y) + |\mathcal{M}_X| + |\mathcal{M}_Y|$.

Algorithm 1 `ALG`.

Input: Two sequences $X = \langle x_1, \ldots, x_n \rangle$ and $Y = \langle y_1, \ldots, y_m \rangle$; and two multisets $\mathcal{M}_X$ and $\mathcal{M}_Y$
Output: Two fillings X^+ of X and $\mathcal{M}_X$ and Y^+ of Y and $\mathcal{M}_Y$; and a common subsequence Z of X^+ and Y^+

$X_b := \varepsilon$, $Y_f := \varepsilon$;
for $i = 1$ to n **do**
  if $x_i = \sigma_y$ for $\sigma_y \in \mathcal{M}_Y$ **then**
    $Y_f := Y_f \oplus \langle \sigma_y \rangle$, $\mathcal{M}_Y := \mathcal{M}_Y \setminus \{\sigma_y\}$;
$Y^+ := Y_f \oplus Y$;
for $i = 1$ to m **do**
  if $y_i = \sigma_x$ for $\sigma_x \in \mathcal{M}_X$ **then**
    $X_b := X_b \oplus \langle \sigma_x \rangle$, $\mathcal{M}_X := \mathcal{M}_X \setminus \{\sigma_x\}$;
$X^+ := X \oplus X_b$;
Find a longest common subsequence Z of the two sequences X^+ and Y^+;
return (X^+, Y^+, Z);

Let $ALG = |Z|$ be the length obtained by our algorithm `ALG` for the input $(X, Y, \mathcal{M}_X, \mathcal{M}_Y)$, i.e., $ALG = L(X^+, Y^+)$. Since a longest common subsequence of X and Y is a common subsequence of X^+ and Y^+, $ALG \geq L(X, Y)$ holds. Furthermore, since $LCS(X, Y_f) \oplus LCS(X_b, Y)$ is another common subsequence of X^+ and Y^+, $ALG \geq L(X, Y_f) + L(X_b, Y) = |\mathcal{M}_Y| + |\mathcal{M}_X|$ holds. As a result, the approximation ratio of `ALG` is bounded as follows:

$$\begin{aligned}
\frac{OPT}{ALG} &\leq \frac{L(X,Y) + |\mathcal{M}_X| + |\mathcal{M}_Y|}{\max\{L(X,Y), |\mathcal{M}_X| + |\mathcal{M}_Y|\}} \\
&= \frac{2(L(X,Y) + |\mathcal{M}_X| + |\mathcal{M}_Y|)}{2(\max\{L(X,Y), |\mathcal{M}_X| + |\mathcal{M}_Y|\})} \\
&\leq \frac{2(L(X,Y) + |\mathcal{M}_X| + |\mathcal{M}_Y|)}{L(X,Y) + |\mathcal{M}_X| + |\mathcal{M}_Y|} \\
&= 2.
\end{aligned}$$

Clearly, `ALG` runs in polynomial time. This completes the proof. ◀

For non-standard inputs, we can also obtain a 2-approximation algorithm by slightly modifying `ALG`. All we have to do is to add $\mathcal{M}_X\mathcal{M}_Y$-matches of redundant symbols. If the sequence of length ℓ is concatenated, then we get $\frac{OPT+\ell}{ALG+\ell} \leq 2$. Further details can be found in the journal version of this paper.

References

1 Said Sadique Adi, Marília D. V. Braga, Cristina G. Fernandes, Carlos Eduardo Ferreira, Fábio Viduani Martinez, Marie-France Sagot, Marco Aurelio Stefanes, Christian Tjandraatmadja, and Yoshiko Wakabayashi. Repetition-free longest common subsequence. *Electron. Notes Discret. Math.*, 30:243–248, 2008. doi:10.1016/j.endm.2008.01.042.

2 Yuichi Asahiro, Jesper Jansson, Guohui Lin, Eiji Miyano, Hirotaka Ono, and Tadatoshi Utashima. Exact algorithms for the repetition-bounded longest common subsequence problem. *Theor. Comput. Sci.*, 838:238–249, 2020. doi:10.1016/j.tcs.2020.07.042.

3 Lasse Bergroth, Harri Hakonen, and Timo Raita. A survey of longest common subsequence algorithms. In Pablo de la Fuente, editor, *Seventh International Symposium on String Processing and Information Retrieval, SPIRE 2000, A Coruña, Spain, September 27-29, 2000*, pages 39–48. IEEE Computer Society, 2000. doi:10.1109/SPIRE.2000.878178.

4 Laurent Bulteau, Falk Hüffner, Christian Komusiewicz, and Rolf Niedermeier. Multivariate algorithmics for np-hard string problems: The algorithmics column by Gerhard J. Woeginger. *Bull. EATCS*, 114, 2014. URL: http://eatcs.org/beatcs/index.php/beatcs/article/view/310.

5 Mauro Castelli, Riccardo Dondi, Giancarlo Mauri, and Italo Zoppis. The longest filled common subsequence problem. In Juha Kärkkäinen, Jakub Radoszewski, and Wojciech Rytter, editors, *28th Annual Symposium on Combinatorial Pattern Matching, CPM 2017, July 4-6, 2017, Warsaw, Poland*, volume 78 of *LIPIcs*, pages 14:1–14:13. Schloss Dagstuhl - Leibniz-Zentrum für Informatik, 2017. doi:10.4230/LIPIcs.CPM.2017.14.

6 Mauro Castelli, Riccardo Dondi, Giancarlo Mauri, and Italo Zoppis. Comparing incomplete sequences via longest common subsequence. *Theor. Comput. Sci.*, 796:272–285, 2019. doi:10.1016/j.tcs.2019.09.022.

7 Thomas H. Cormen, Charles E. Leiserson, Ronald L. Rivest, and Clifford Stein. *Introduction to Algorithms, 4th Edition*. MIT Press, 2022. URL: http://mitpress.mit.edu/books/introduction-algorithms-fourth-edition.

8 Daniel S. Hirschberg. A linear space algorithm for computing maximal common subsequences. *Commun. ACM*, 18(6):341–343, 1975. doi:10.1145/360825.360861.

9 Daniel S. Hirschberg. Algorithms for the longest common subsequence problem. *J. ACM*, 24(4):664–675, 1977. doi:10.1145/322033.322044.

10 Haitao Jiang, Chunfang Zheng, David Sankoff, and Binhai Zhu. Scaffold filling under the breakpoint and related distances. *IEEE ACM Trans. Comput. Biol. Bioinform.*, 9(4):1220–1229, 2012. doi:10.1109/TCBB.2012.57.

11 Radu Stefan Mincu and Alexandru Popa. Better heuristic algorithms for the repetition free LCS and other variants. In Travis Gagie, Alistair Moffat, Gonzalo Navarro, and Ernesto Cuadros-Vargas, editors, *String Processing and Information Retrieval - 25th International Symposium, SPIRE 2018, Lima, Peru, October 9-11, 2018, Proceedings*, volume 11147 of *Lecture Notes in Computer Science*, pages 297–310. Springer, 2018. doi:10.1007/978-3-030-00479-8_24.

12 Radu Stefan Mincu and Alexandru Popa. Heuristic algorithms for the longest filled common subsequence problem. In *20th International Symposium on Symbolic and Numeric Algorithms for Scientific Computing, SYNASC 2018, Timisoara, Romania, September 20-23, 2018*, pages 449–453. IEEE, 2018. doi:10.1109/SYNASC.2018.00075.

13 Adriana Muñoz, Chunfang Zheng, Qian Zhu, Victor A. Albert, Steve Rounsley, and David Sankoff. Scaffold filling, contig fusion and comparative gene order inference. *BMC Bioinform.*, 11:304, 2010. doi:10.1186/1471-2105-11-304.

14 Saul B. Needleman and Christian D. Wunsch. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of Molecular Biology*, 48(3):443–453, 1970.

15 David Sankoff. Matching sequences under deletion/insertion constraints. In *Proc. National Academy of Science USA*, volume 69, pages 4–6, 1972. doi:10.1073/pnas.69.1.4.

16 Robert A. Wagner and Michael J. Fischer. The string-to-string correction problem. *J. ACM*, 21(1):168–173, 1974. doi:10.1145/321796.321811.