Theoretical Computer Science 646 (2016) 16–25

Contents lists available at ScienceDirect

Theoretical Computer Science

www.elsevier.com/locate/tcs

Parameterized tractability of the maximum-duo preservation string mapping problem

Stefano Beretta [a,b,*], Mauro Castelli [c], Riccardo Dondi [d]

[a] Istituto di Tecnologie Biomediche, Consiglio Nazionale delle Ricerche, Segrate, Italy
[b] Dipartimento di Informatica, Sistemistica e Comunicazione, Università degli Studi di Milano – Bicocca, Milano, Italy
[c] NOVA IMS, Universidade Nova de Lisboa, Lisboa, Portugal
[d] Dipartimento di Lettere, Filosofia e Comunicazione, Università degli Studi di Bergamo, Bergamo, Italy

ARTICLE INFO

Article history:
Received 17 December 2015
Received in revised form 10 June 2016
Accepted 6 July 2016
Available online 15 July 2016
Communicated by M. Golin

Keywords:
Computational biology
Common string partition
Parameterized algorithms
Kernelization

ABSTRACT

In this paper we investigate the parameterized complexity of the Maximum-Duo Preservation String Mapping Problem, the complementary of the Minimum Common String Partition Problem. We show that this problem is fixed-parameter tractable when parameterized by the number k of conserved duos, by first giving a parameterized algorithm based on the color-coding technique and then presenting a reduction to a kernel of size $O(k^6)$.

1. Introduction

Minimum Common String Partition (MCSP) is a problem that emerged in the field of comparative genomics [9] and, in particular, in the context of ortholog gene assignments [9]. Given two strings (genomes) A and B of length n, MCSP asks for a partition of the two strings into a minimum cardinality multiset of identical substrings. The complexity of this problem has been previously studied in the literature. More precisely, the MCSP problem is known to be APX-hard, even when each symbol has at most 2 occurrences in each input string [15], while it admits a polynomial-time algorithm when each symbol occurs exactly once in each input string. Approximation algorithms for this problem have been proposed in [10,11,15,18]. More precisely, an $O(\log n \log^* n)$-approximation algorithm has been given in [11], while an $O(k)$-approximation algorithm, when the number of occurrences of each symbol is bounded by k, has been given in [18]. When $k = 2$ and $k = 3$ respectively, approximation algorithms of factor 1.1037 and 4, respectively, have been given in [15].

The parameterized complexity [13,19] of MCSP has also been investigated. First, fixed-parameter algorithms have been given when the problem is parameterized by two parameters. In [12] this problem has been shown to be fixed-parameter tractable, when parameterized by the number of substrings in the solution and by the repetition number of the input strings. Then, the MCSP problem has been shown to be fixed-parameter tractable when parameterized by the number of substrings in the solution and the maximum number of occurrences of a symbol in an input string [6,17]. Recently, MCSP has been shown to be fixed-parameter tractable when parameterized by the single parameter number of substrings of the partition [7].

* Corresponding author at: Dipartimento di Informatica, Sistemistica e Comunicazione, Università degli Studi di Milano – Bicocca, Milano, Italy.
E-mail addresses: stefano.beretta@disco.unimib.it (S. Beretta), mcastelli@novaims.unl.pt (M. Castelli), riccardo.dondi@unibg.it (R. Dondi).

http://dx.doi.org/10.1016/j.tcs.2016.07.011

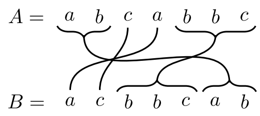

Fig. 1. An example of two related strings A and B. The mapping of their positions is represented by connecting positions/substrings. Position 1 and 2 of A are mapped into positions 6 and 7 of B, hence duo $(A[1], A[2])$ of A is preserved; position 1 in A induces duo $(A[1], A[2])$. Similarly, the sequence $d_{A(5,7)}$ of consecutive duos is mapped into the sequence $d_{B(3,5)}$ of consecutive duos; hence duos $(A[5], A[6])$, $(A[6], A[7])$ of A are preserved. The number of preserved duos induced by the mapping is 3.

Here, we consider the complementary of the MCSP problem, called Maximum-Duo Preservation String Mapping Problem, where instead of minimizing the number of identical substrings in the partition, we aim to maximize the number of preserved duos. Informally, the idea is to define a mapping between the positions (having a same symbol) of the input strings such that the number of adjacent positions of one string mapped to adjacent positions of the other one (called preserved duos) is maximized. This problem has been proposed in [8], has been shown to be APX-hard when each symbol has at most 2 occurrences in each input string [4], and can be approximated within factor $\frac{1}{4}$ [4].

In this work, we study the parameterized complexity of the Maximum-Duo Preservation String Mapping Problem, where the parameter is the number of preserved duos. More precisely, after introducing definitions and properties of Maximum-Duo Preservation String Mapping in Section 2, we describe in Section 3 a fixed-parameter algorithm for the problem, based on the color-coding technique. Then, in Section 4, we present a reduction to a polynomial kernel of size $O(k^6)$ obtained by appropriately selecting subsets of duos of the input strings and by defining two new strings containing these subsets on an extended alphabet.

The results described in this paper are mainly of theoretical interest, since a solution of the Maximum-Duo Preservation String Mapping Problem is expected to preserve many adjacencies. Indeed, the complexity of the first algorithm and size of the kernel both depend (exponentially and polynomially, respectively) on the number of preserved adjacencies. However, the fixed-parameter algorithms we propose can be of interest for describing the parameterized complexity status of the combinatorial problems related to string partitioning. For example, while it is still unknown whether Minimum Common String Partition Problem admits a polynomial kernel, the result in Section 4 shows that such a kernel exists for the complementary problem.

2. Preliminaries

In this section, we introduce some concepts that will be used in the rest of the paper and we give the formal definition of the Maximum-Duo Preservation String Mapping Problem. Fig. 1 illustrates some of the definitions we give in this section.

Let Σ be a non-empty finite set of symbols. Given a string A over Σ, we denote by $|A|$ the length of A and by $A[i]$, with $1 \leq i \leq |A|$, the symbol of A at position i. Moreover, we denote by $A[i, j]$, with $1 \leq i < j \leq |A|$, the substring of A starting at position i and ending at position j. Given a string A, a *duo* is an ordered pair of consecutive elements $(A[i], A[i+1])$. Consider a duo $(A[i], A[i+1])$ in a string A; it is *preservable* if there exists a duo $(B[j], B[j+1])$ in a string B such that $A[i] = B[j]$ and $A[i+1] = B[j+1]$. Symmetrically, the property holds for a duo of string B.

Given two strings A and B, such that B is a permutation of A, we say that A and B are *related*. In the rest of the paper we assume that $|A| = |B| = n$.

Given two related strings A and B, a *mapping* m of A into B is a bijective function from the positions of A to the positions of B such that $m(i) = j$ implies that $A[i] = B[j]$, i.e. the two positions i, j of the two strings contain the same symbol. A *partial mapping* m of A into B is a bijective function from a subset of positions of A to a subset of positions of B such that $m(i) = j$ implies that $A[i] = B[j]$. The definition of mapping and partial mapping can be extended to two sets of duos of related strings A and B, that is if positions i and $i+1$, with $1 \leq i \leq n-1$, are mapped into positions j and $j+1$, with $1 \leq j \leq n-1$, we say that duo $(A[i], A[i+1])$ is mapped into duo $(B[j], B[j+1])$.

Given two related strings A and B, and a mapping m of the positions of A into the positions of B, a duo $(A[i], A[i+1])$ is preserved if $m(i) = j$ and $m(i+1) = j+1$ (see Fig. 1 for an example).

Now, we give the definition of the Maximum-Duo Preservation String Mapping Problem (in its decision version).

Maximum-Duo Preservation String Mapping Problem (Max-Duo PSM)
Input: two related strings A and B, an integer k.
Output: is there a mapping m of A into B such that the number of preserved duos is at least k?

In this paper, we focus on the parameterized complexity of Max-Duo PSM, when parameterized by the number k of preserved duos.

Consider a string S, with $S \in \{A, B\}$, and a string $\bar{S} \in \{A, B\} \setminus \{S\}$. Given two positions $1 \leq i < j \leq n$, we denote by $d_{S(i,j)}$ the sequence of *consecutive duos* $(S[i], S[i+1]), \ldots, (S[j-1], S[j])$; the length of $d_{S(i,j)}$ is the number $j-i$ of consecutive duos in it. Given the sequence $d_{S(i,j)}$ of consecutive duos, the *string corresponding* to $d_{S(i,j)}$ is $S[i, j]$. Given a string $S[i, j]$ the sequence of consecutive duos *induced* by $S[i, j]$ is $d_{S(i,j)}$. We say that position i, with $1 \leq i \leq n-1$, of a string S, induces duo $(S[i], S[i+1])$.