What makes the ARC-PRESERVING SUBSEQUENCE problem hard?*

Guillaume Blin[1], Guillaume Fertin[1], Romeo Rizzi[2], and Stéphane Vialette[3]

[1] LINA - FRE CNRS 2729 Université de Nantes,
2 rue de la Houssinière BP 92208 44322 Nantes Cedex 3 - FRANCE
{blin,fertin}@lina.univ-nantes.fr

[2] Universit degli Studi di Trento Facolt di Scienze - Dipartimento di Informatica e Telecomunicazioni
Via Sommarive, 14 - I38050 Povo - Trento (TN) - ITALY
Romeo.Rizzi@unitn.it

[3] LRI - UMR CNRS 8623 Faculté des Sciences d'Orsay, Université Paris-Sud
Bât 490, 91405 Orsay Cedex - FRANCE
vialette@lri.fr

Abstract. In molecular biology, RNA structure comparison and motif search are of great interest for solving major problems such as phylogeny reconstruction, prediction of molecule folding and identification of common functions. RNA structures can be represented by arc-annotated sequences (primary sequence along with arc annotations), and this paper mainly focuses on the so-called *arc-preserving subsequence* (APS) problem where, given two arc-annotated sequences (S, P) and (T, Q), we are asking whether (T, Q) can be obtained from (S, P) by deleting some of its bases (together with their incident arcs, if any). In previous studies, this problem has been naturally divided into subproblems reflecting the intrinsic complexity of the arc structures. We show that APS(CROSSING, PLAIN) is **NP**-complete, thereby answering an open problem posed in [11]. Furthermore, to get more insight into where the actual border between the polynomial and the **NP**-complete cases lies, we refine the classical subproblems of the APS problem in much the same way as in [19] and prove that both APS($\{\sqsubset, \between\}, \emptyset$) and APS($\{<, \between\}, \emptyset$) are **NP**-complete. We end this paper by giving some new positive results, namely showing that APS($\{\between\}$, $\emptyset$) and APS($\{\between\}$,$\{\between\}$) are polynomial time solvable.

Keywords: RNA structures, Arc-Preserving Subsequence problem, Computational complexity.

1 Introduction

At a molecular state, the understanding of biological mechanisms is subordinated to the discovery and the study of RNA functions. Indeed, it is established that the

* This work was partially supported by the French-Italian PAI Galileo project number 08484VH and by the CNRS project ACI Masse de Données "NavGraphe". A preliminary version of this paper appeared in the Proc. of IWBRA'05, Springer, V.S. Sunderam et al. (Eds.): ICCS 2005, LNCS 3515, pp. 860-868, 2005.

conformation of a single-stranded RNA molecule (a linear sequence composed of ribonucleotides A, U, C and G, also called primary structure) partly determines the function of the molecule. This conformation results from the folding process due to local pairings between complementary bases ($A{-}U$ and $C{-}G$, connected by a hydrogen bond). The secondary structure of an RNA (a simplification of the complex 3-dimensional folding of the sequence) is the collection of folding patterns (stem, hairpin loop, bulge loop, internal loop, branch loop and pseudoknot) that occur in it.

RNA secondary structure comparison is important in many contexts, such as:

- identification of highly conserved structures during evolution, non detectable in the primary sequence which is often slightly preserved. These structures suggest a significant common function for the studied RNA molecules [16, 18, 13, 8],
- RNA classification of various species (phylogeny)[4, 3, 21],
- RNA folding prediction by considering a set of already known secondary structures [24, 14],
- identification of a consensus structure and consequently of a common role for molecules [22, 5].

Structure comparison for RNA has thus become a central computational problem bearing many challenging computer science questions. At a theoretical level, the RNA structure is often modeled as an *arc-annotated sequence*, that is a pair (S, P) where S is the sequence of ribonucleotides and P represents the hydrogen bonds between pairs of elements of S. Different pattern matching and motif search problems have been investigated in the context of arc-annotated sequences among which we can mention the *arc-preserving subsequence* (APS) problem, the Edit Distance problem, the *arc-substructure* (AST) problem and the *longest arc-preserving subsequence* (LAPCS) problem (see for instance [6, 15, 12, 11, 2]). For other related studies concerning algorithmic aspects of (protein) structure comparison using *contact maps*, refer to [10, 17].

In this paper, we focus on the *arc-preserving subsequence* (APS) problem: given two arc-annotated sequences (S, P) and (T, Q), this problem asks whether (T, Q) can be exactly obtained from (S, P) by deleting some of its bases together with their incident arcs, if any. This problem is commonly encountered when one is searching for a given RNA pattern in an RNA database [12]. Moreover, from a theoretical point of view, the APS problem can be seen as a restricted version of the LAPCS problem, and hence has applications in the structural comparison of RNA and protein sequences [6, 10, 23]. The APS problem has been extensively studied in the past few years [11, 12, 6]. Of course, different restrictions on arc-annotation alter the computational complexity of the APS problem, and hence this problem has been naturally divided into subproblems reflecting the complexity of the arc structure of both (S, P) and (T, Q): plain, chain, nested, crossing or unlimited (see Section 2 for details). All of them but one have been classified as to whether they are polynomial time solvable or **NP**-complete. The problem of the existence of a polynomial time algorithm for the APS(Crossing,Plain) problem was mentioned in [11] as the last open problem in the context of arc-preserving subsequences (cf. Table 1). Unfortunately, as we shall prove in Section 4, the APS(Crossing,Plain) problem is **NP**-complete even for restricted special cases.

In analyzing the computational complexity of a problem, we are often trying to define the precise boundary between the polynomial and the **NP**-complete cases. Therefore, as another step towards establishing the precise complexity landscape of the APS problem, it is of great interest to subdivide the existing cases into more precise ones, that is to refine the classical complexity levels of the APS problem, for determining more precisely what makes the problem hard. For that purpose, we use the framework introduced by Vialette [19] in the context of 2-intervals (a simple abstract structure for modelling RNA secondary structures). As a consequence, the number of complexity levels rises from 4 (not taking into account the unlimited case) to 8, and all the entries of this new complexity table need to be filled. Previous known results concerning the APS problem, along with two **NP**-completeness and two polynomiality proofs, allow us to fill all the entries of this new table, therefore determining what exactly makes the APS problem hard.

The paper is organized as follows. In Section 2, we give notations and definitions concerning the APS problem. In Section 3 we introduce and explain the new refinements of the complexity levels we are going to study. In Section 4, we show that the APS($\{\sqsubset, \between\}, \emptyset$) problem is **NP**-complete thereby proving that the (classical) APS(Crossing, Plain) problem is **NP**-complete as well. As another refinement to that result, we prove that the APS($\{<, \between\}, \emptyset$) problem is **NP**-complete. Finally, in Section 5, we give new polynomial time solvable algorithms for restricted instances of the APS(Crossing, Plain) problem.

2 Preliminaries

An RNA structure is commonly represented as an arc-annotated sequence (S, P) where S is the sequence of ribonucleotides (or bases) and P is the set of arcs connecting pairs of bases in S. Let (S, P) and (T, Q) be two arc-annotated sequences such that $|S| \geq |T|$ (in the following, $n = |S|$ and $m = |T|$). The APS problem asks whether (T, Q) can be exactly obtained from (S, P) by deleting some of its bases together with their incident arcs, if any.

Since the general problem is easily seen to be intractable [6], the arc structure must be restricted. Evans [6] proposed four possible restrictions on P (resp. Q) which were largely reused in the subsequent literature:

1. there is no base incident to more than one arc,
2. there are no arcs crossing,
3. there is no arc contained in another,
4. there is no arc.

These restrictions are used progressively and inclusively to produce five different levels of allowed arc structure:

- Unlimited - the general problem with no restrictions
- Crossing - restriction 1
- Nested - restrictions 1 and 2
- Chain - restrictions 1, 2 and 3
- Plain - restriction 4

Guo proved in [12] that the APS(Crossing, Chain) problem is **NP**-complete. Guo et al. observed in [11] that the **NP**-completeness of the APS(Crossing, Crossing) and APS(Unlimited, Plain) easily follows from results of Evans [6] concerning the LAPCS problem. Furthermore, they gave a $O(nm)$ time for the APS(Nested, Nested) problem. This algorithm can be applied to easier problems such as APS(Nested, Chain), APS(Nested, Plain), APS(Chain, Chain) and APS(Chain,Plain). Finally, Guo et al. mentioned in [11] that APS(Chain, Plain) can be solved in $O(n+m)$ time. Until now, the question of the existence of an exact polynomial algorithm for the problem APS(Crossing, Plain) remained open. We will first show in the present paper that the problem APS(Crossing,Plain) is **NP**-complete. Table 1 surveys known and new results for various types of APS. Observe that the Unlimited level has no restrictions, and hence is of limited interest in our study. Consequently, from now on we will not be concerned anymore with that level.

APS				
	Crossing	Nested	Chain	Plain
Crossing	**NP**-complete [6]	**NP**-complete [12]		**NP**-complete $\star$
Nested		$O(nm)$ [11]		
Chain			$O(nm)$ [11]	$O(n+m)$ [11]

Table 1. APS problem complexity where $n = |S|$ and $m = |T|$. $\star$ result from this paper.

3 Refinement of the APS problem

In this section, we propose a refinement of the APS problem. We first state formally our approach and explain why such a refinement is relevant for both theoretical and experimental studies. We end the section by giving easy properties of the proposed refinement that will prove extremely useful in Section 5.

3.1 Splitting the levels

As we will show in Section 4, the APS(Crossing, Plain) problem is **NP**-complete. That result answers the last open problem concerning the computational complexity of the APS problem with respect to classical complexity levels, *i.e.*, Plain, Chain, Nested and Crossing (cf. Table 1). However, we are mainly interested in the elaboration of the precise border between **NP**-complete and polynomially solvable cases. Indeed, both theorists and practitioners might naturally ask for more information concerning the hard cases of the APS problem in order to get valuable insight into what makes the problem difficult.

As a next step towards a better understanding of what makes the APS problem hard, we propose to refine the models which are classically used for classifying arc-annotated sequences. Our refinement consists in splitting those models of arc-annotated sequences into more precise relations between arcs. For example, such a refinement provides a general framework for investigating polynomial time solvable and hard restricted instances of APS(Crossing, Plain), thereby refining in many ways Theorem 1 (see Section 5).

We use the three relations first introduced by Vialette [19, 20] in the context of 2-*intervals* (a simple abstract structure for modelling RNA secondary structures). Actually, his definition of 2-intervals could almost apply in this paper (the main difference lies in the fact that Vialette used 2-intervals for representing sets of contiguous arcs). Vialette defined three possible relations between 2-intervals that can be used for arc-annotated sequences as well. They are the following: for any two arcs $p_1 = (i, j)$ and $p_2 = (k, l)$ in P, we will write $p_1 < p_2$ if $i < j < k < l$ (*precedence* relation), $p_1 \sqsubset p_2$ if $k < i < j < l$ (*nested* relation) and $p_1 \between p_2$ if $i < k < j < l$ (*crossing* relation). Two arcs p_1 and p_2 are τ-comparable for some $\tau \in \{<, \sqsubset, \between\}$ if $p_1 \tau p_2$ or $p_2 \tau p_1$. Let $\mathcal{P}$ be a set of arcs and R be a non-empty subset of $\{<, \sqsubset, \between\}$. The set $\mathcal{P}$ is said to be R-*comparable* if any two distinct arcs of $\mathcal{P}$ are τ-comparable for some $\tau \in R$. An arc-annotated sequence (S, P) is said to be an R-arc-annotated sequence for some non-empty subset R of $\{<, \sqsubset, \between\}$ if P is R-comparable. We will write $R = \emptyset$ in case $P = \emptyset$. Observe that our model cannot deal with arc-annotated sequences which contain only one arc. However, having only one arc or none can not really affect the computational complexity of the problem. Just one guess reduces from one case to the other. Details are omitted here.

As a straightforward illustration of the above definitions, classical complexity levels for the APS problem can be expressed in terms of combinations of our new relations: Plain is fully described by $R = \emptyset$, Chain is fully described by $R = \{<\}$, Nested is fully described by $R = \{<, \sqsubset\}$ and Crossing is fully described by $R = \{<, \sqsubset, \between\}$. The key point is to observe that our refinement allows us to consider new structures for arc-annotated sequences, namely $R = \{\sqsubset\}$, $R = \{\between\}$, $R = \{<, \between\}$ and $R = \{\sqsubset, \between\}$, which could not be considered using the classical complexity levels. Although other refinements may be possible (in particular well-suited for parameterized complexity analysis), we do believe that such an approach allows a more precise analysis of the complexity of the APS problem.

Of course one might object that some of these subdivisions are unlikely to appear in RNA secondary structures. While this is true, it is also true that it is of great interest to answer, at least partly, the following question: Where is the precise boundary between the polynomial and the **NP**-complete cases? Indeed, such a question is relevant for both theoretical and experimental studies.

For one, many important optimization problems are known to be **NP**-complete. That is, unless $\mathbf{P} = \mathbf{NP}$, there is no polynomial time algorithm that optimally solves these on every input instance, and hence proving a problem to be **NP**-complete is generally accepted as a proof of its difficulty. However the problem to be solved may be much more specialized than the general one that was proved to be **NP**-complete. Therefore, during the past three decades, many studies have been devoted to proving **NP**-completeness results for highly restricted instances in order to precisely define the border between tractable and intractable problems. Our refinements have thus to be seen as another step towards establishing the precise complexity landscape of the APS problem.

For another, it is worthwhile keeping in mind that intractability must be coped with and problems must be solved in practical applications. Computer science theory has articulated a few general programs for systematically coping with the ubiquitous phenomena of computational intractability: average case analysis, approximation algorithm, randomized algorithm and fixed parameter complexity. Fully understanding where the boundary lies between efficiently solvable formulations and intractable ones is another important approach. Indeed, from an engineering point of view for which the emphasis is on efficiency, that precise boundary might be a good starting point for designing efficient heuristics or for exploring fixed-parameter tractability. The better our understanding of the problem, the better our ability in defining efficient algorithms for practical applications.

3.2 Immediate results

First, observe that, as in Table 1, we only have to consider cases of $\text{APS}(R_1,R_2)$ where R_1 and R_2 are compatible, i.e. $R_2 \subseteq R_1$. Indeed, if this is not the case, we can immediately answer negatively since there exists two arcs in T which satisfy a relation in R_2 which is not in R_1, and hence T simply cannot be obtained from S by deleting bases of S. Those incompatible cases are simply denoted by hatched areas in Table 2.

Some known results allow us to fill many entries of the new complexity table derived from our refinement. The remainder of this subsection is devoted to detailing these first easy statements. We begin with an observation concerning complexity propagation properties of the APS problems in our refined model.

Observation 1 *Let R_1, R_2, R'_1 and R'_2 be four subsets of $\{<, \sqsubset, \between\}$ such that $R'_2 \subseteq R_2 \subseteq R_1$ and $R'_2 \subseteq R'_1 \subseteq R_1$. If $\text{APS}(R'_1,\ R'_2)$ is **NP**-complete (resp. $\text{APS}(R_1,\ R_2)$ is polynomial time solvable) then so is $\text{APS}(R_1,\ R_2)$ (resp. $\text{APS}(R'_1,\ R'_2)$).*

On the positive side, Gramm *et al.* have shown that APS(Nested, Nested) is solvable in $O(nm)$ time [11]. Another way of stating this is to say that $\text{APS}(\{<, \sqsubset\}, \{<, \sqsubset\})$ is solvable in $O(mn)$ time. That result together with Observation 1 may be summarized by saying that $\text{APS}(R_1,\ R_2)$ for any compatible R_1 and R_2 such that $\between \notin R_1$ and $\between \notin R_2$ is polynomial time solvable.

Conversely, the **NP**-completeness of APS(Crossing,Crossing) has been proved by Evans [6]. A simple reading shows that her proof is concerned with $\{<,\sqsubset,\between\}$-arc-annotated sequences, and hence she actually proved that APS($\{<,\sqsubset,\between\}$, $\{<,\sqsubset,\between\}$) is **NP**-complete. Similarly, in proving that APS(Crossing, Chain) is **NP**-complete [12], Guo actually proved that APS($\{<,\sqsubset,\between\}$, $\{<\}$) is **NP**-complete. Note that according to Observation 1, this latter result implies that APS($\{<,\sqsubset,\between\}$, $\{<,\sqsubset\}$) and APS($\{<,\sqsubset,\between\}$,$\{<,\between\}$) are **NP**-complete.

Table 2 surveys known and new results for various types of our refined APS problem. Observe that this paper answers all questions concerning the APS problem with respect to the new complexity levels.

APS								
R_2 R_1	$\{<,\sqsubset,\between\}$	$\{\sqsubset,\between\}$	$\{<,\between\}$	$\{\between\}$	$\{<,\sqsubset\}$	$\{\sqsubset\}$	$\{<\}$	$\emptyset$
$\{<,\sqsubset,\between\}$	**NP**-C [6]	?	**NP**-C [12]	?	**NP**-C [12]	?	**NP**-C [12]	?
$\{\sqsubset,\between\}$		?	////	?	////	?	////	?
$\{<,\between\}$			?	?	////	////	?	?
$\{\between\}$				?	////	////	////	?
$\{<,\sqsubset\}$					$O(nm)$ [11]	$O(nm)$ [11]	$O(nm)$ [11]	$O(nm)$ [11]
$\{\sqsubset\}$						$O(nm)$ [11]	////	$O(nm)$ [11]
$\{<\}$							$O(nm)$ [11]	$O(n+m)$ [11]
$\emptyset$								$O(n+m)$ [11]

Table 2. Complexity results after refinement of the complexity levels. ////: incompatible cases. ?: open problems.

4 Hardness results

We show in this section that APS($\{\sqsubset,\between\},\emptyset$) is **NP**-complete thereby proving that the (classical) APS(Crossing, Plain) problem is **NP**-complete. That result answers an open problem posed in [11], which was also the last open problem concerning the computational complexity of the APS problem with respect to classical complexity levels, *i.e.*, Plain, Chain, Nested and Crossing (cf. Table 1). Furthermore, we prove that the APS($\{<,\between\},\emptyset$) is **NP**-complete as well.

We provide a polynomial time reduction from the 3-Sat problem: Given a set $\mathcal{V}_n$ of n variables and a set $\mathcal{C}_q$ of q clauses (each composed of three literals) over $\mathcal{V}_n$, the problem asks to find a truth assignment for $\mathcal{V}_n$ that satisfies all clauses of $\mathcal{C}_q$. It is well-known that the 3-Sat problem is **NP**-complete [9].

It is easily seen that the APS($\{\sqsubset,\between\},\emptyset$) problem is in **NP**. The remainder of the section is devoted to proving that it is also **NP**-hard. Let $\mathcal{V}_n = \{x_1, x_2, ...x_n\}$ be a finite set of n variables and $\mathcal{C}_q = \{c_1, c_2, \dots, c_q\}$ a collection of q clauses. Observe that there is no loss of generality in assuming that, in each clause, the literals are ordered from left to right, *i.e.*, if $c_i = (x_j \vee x_k \vee x_l)$ then $j < k < l$.

Let us first detail the construction of the sequences S and T:

$$S = S^s_{x_1} A\ S^s_{\overline{x_1}}\ S^s_{x_2} A\ S^s_{\overline{x_2}} \dots S^s_{x_n} A\ S^s_{\overline{x_n}}\ S_{c_1}\ S_{c_2} \dots S_{c_q}\ S^e_{x_1}\ S^e_{x_2} \dots S^e_{x_n}$$
$$T = T^s_{x_1}\ T^s_{x_2} \dots T^s_{x_n}\ T_{c_1}\ T_{c_2} \dots T_{c_q}\ T^e_{x_1}\ T^e_{x_2} \dots T^e_{x_n}$$

We now detail the subsequences that compose S and T. Let γ_m (resp. $\gamma_{\overline{m}}$) be the number of occurrences of literal x_m (resp. $\overline{x_m}$) in $\mathcal{C}_q$ and let $k_m = \max(\gamma_m, \gamma_{\overline{m}})$. For each variable $x_m \in \mathcal{V}_n$, $1 \leq m \leq n$, we construct words $S^s_{x_m} = AC^{k_m}$, $S^s_{\overline{x_m}} = C^{k_m}A$ and $T^s_{x_m} = AC^{k_m}A$ where C^{k_m} represents a word of k_m consecutive bases C. For each clause c_i of $\mathcal{C}_q$, $1 \leq i \leq q$, we construct words $S_{c_i} = UGGGA$ and $T_{c_i} = UGA$. Finally, for each variable $x_m \in \mathcal{V}_n$, $1 \leq m \leq n$, we construct words $S^e_{x_m} = UUA$ and $T^e_{x_m} = UA$.

Having disposed of the two sequences, we now turn to defining the corresponding two arc structures (see Figure 1). In the following, $\mathcal{S}eq[i]$ will denote the i^{th} base of a sequence $\mathcal{S}eq$ and, for any $1 \leq m \leq n$, $l_{\overline{m}} = |S^s_{\overline{x_m}}|$. For all $1 \leq m \leq n$, we create the two following arcs: ($S^s_{x_m}[1]$,$S^e_{x_m}[1]$) and ($S^s_{\overline{x_m}}[l_{\overline{m}}]$,$S^e_{x_m}[2]$). For each clause c_i of $\mathcal{C}_q$, $1 \leq i \leq q$, and for each $1 \leq m \leq n$, if the k^{th} (*i.e.* 1^{st}, 2^{nd} or 3^{rd}) literal of c_i is x_m (resp. $\overline{x_m}$) then we create an arc between any free (i.e. not already incident to an arc) base C of $S^s_{\overline{x_m}}$ (resp. $S^s_{x_m}$) and the k^{th} base G of S_{c_i} (note that this is possible by definition of $S^s_{\overline{x_m}}$, $S^s_{x_m}$ and S_{c_i}). On the whole, the instance we have constructed is composed of $3q + 2n$ arcs. We denote by APS-CP-construction any construction of this type. In the following, we will distinguish arcs between bases A and U, denoted by AU-arcs, from arcs between bases C and G, denoted by CG-arcs. An illustration of an APS-CP-construction is given in Figure 1. Clearly, our construction can be carried out in polynomial time. Moreover, the result of such a construction is indeed an instance of APS($\{\sqsubset, \between\}, \emptyset$), since $Q = \emptyset$ (no arc is added to T) and P is a $\{\sqsubset, \between\}$-comparable set (since there are no arcs $\{<\}$-comparable.

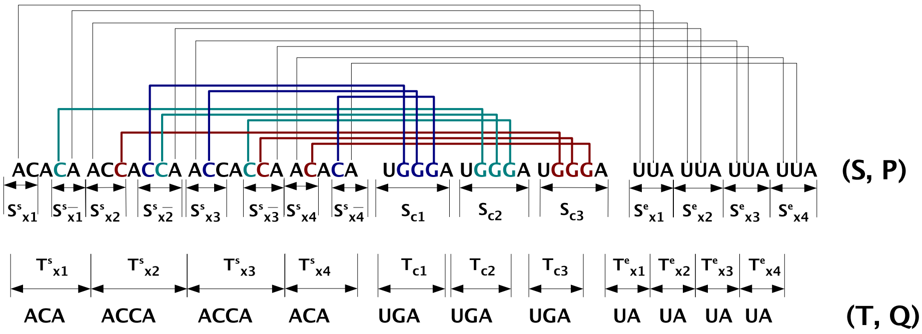


Fig. 1. Example of an APS-CP-construction with $\mathcal{C}_q = (x_2 \vee \overline{x_3} \vee x_4) \wedge (x_1 \vee x_2 \vee x_3) \wedge (\overline{x_2} \vee x_3 \vee \overline{x_4})$.

We begin by proving a canonicity lemma of an APS-CP-construction.

Lemma 1. *Let (S, P) and (T, Q) be any two arc-annotated sequences obtained from an* APS-CP*-construction. If (T, Q) can be obtained from (S, P) by deleting some of its bases together with their incident arcs, if any, then for each $1 \leq i \leq q$ and $1 \leq m \leq n$:*

1. *T_{c_i} is obtained from S_{c_i} by deleting two of its three bases G,*
2. *$T^e_{x_m}$ is obtained from $S^e_{x_m}$ by deleting one of its two bases U,*
3. *$T^s_{x_m}$ is obtained from $S^s_{x_m} A S^s_{\overline{x_m}}$ by deleting either $S^s_{x_m}$ or $S^s_{\overline{x_m}}$.*

Proof. Let (S, P) and (T, Q) be two arc-annotated sequences resulting from an APS-CP-construction.
(1) By construction, the first base U appearing in S (resp. T) is $S_{c_1}[1]$ (resp. $T_{c_1}[1]$). Thus, $T_{c_1}[1]$ is obtained from a base U of S at, or after, $S_{c_1}[1]$. Moreover, the number of bases A appearing after $S_{c_1}[1]$ in S is equal to the number of bases A appearing after $T_{c_1}[1]$ in T. Therefore, every base A appearing after $S_{c_1}[1]$ and $T_{c_1}[1]$ must be matched. That is, for each $1 \leq i \leq q$, $T_{c_i}[3]$ is matched to $S_{c_i}[5]$. In particular, $T_{c_q}[3]$ is matched to $S_{c_q}[5]$. But since there are as many bases U between $S_{c_1}[1]$ and $S_{c_q}[5]$ as there are between $T_{c_1}[1]$ and $T_{c_q}[3]$, any base U in this interval in S must be matched to any base U in this interval in T; that is, for any $1 \leq i \leq q$, $T_{c_i}[1]$ is matched to $S_{c_i}[1]$. Thus, we conclude that for any $1 \leq i \leq q$, T_{c_i} is obtained by deleting two of the three bases G of S_{c_i}.
(2) By the above argument concerning the bases A appearing after $S_{c_1}[1]$ and $T_{c_1}[1]$, we know that if (T, Q) can be obtained from (S, P), then $T^e_{x_m}[2]$ is matched to $S^e_{x_m}[3]$ for any $1 \leq m \leq n$. Thus, for any $1 \leq m \leq n$, $T^e_{x_m}$ is obtained from $S^e_{x_m}$, and in particular $T^e_{x_m}[1]$ is matched to either $S^e_{x_m}[1]$ or $S^e_{x_m}[2]$.
(3) By definition, as there is no arc incident to bases of T, at least one base incident to every arc of P has to be deleted. We just mentioned that $T^e_{x_m}[1]$ is matched to either $S^e_{x_m}[1]$ or $S^e_{x_m}[2]$ for any $1 \leq m \leq n$. Thus, since by construction there is an arc between $S^e_{x_m}[1]$ and $S^s_{x_m}[1]$ (resp. $S^e_{x_m}[2]$ and $S^s_{\overline{x_m}}[l_{\overline{m}}]$), for any $1 \leq m \leq n$ either $S^s_{x_m}[1]$ or $S^s_{\overline{x_m}}[l_{\overline{m}}]$ has to be deleted; and all these arcs connect a base A appearing before $S_{c_1}[1]$ to a base U appearing after $S_{c_q}[5]$. Therefore, for any $1 \leq m \leq n$ a base A appearing before $S_{c_1}[1]$ in S is deleted. Originally, there are $3n$ bases A appearing before $S_{c_1}[1]$ in S and $2n$ appearing before the first base of $T_{c_1}[1]$ in T. Thus, the number of bases A matched in S and appearing before $S_{c_1}[1]$ is equal to the number of bases A appearing before $T_{c_1}[1]$ in T. But since, for each $1 \leq m \leq n$, a base A of either $S^s_{x_m}$ or $S^s_{\overline{x_m}}$ is deleted, we conclude that for each $1 \leq m \leq n$, $T^s_{x_m}$ is obtained from $S^s_{x_m} A S^s_{\overline{x_m}}$, by deleting either $S^s_{x_m}$ or $S^s_{\overline{x_m}}$. □

We now turn to proving that our construction is a polynomial time reduction from 3-SAT to APS(CROSSING, PLAIN).

Lemma 2. *Let I be an instance of the problem* 3-SAT *with n variables and q clauses, and I' an instance $((S, P); (T, Q))$ of* APS$(\{\sqsubset, \between\}, \emptyset)$ *obtained by an* APS-CP*-construction from I. An assignment of the variables that satisfies the boolean formula of I exists iff T is an Arc-Preserving Subsequence of S.*

Proof. ($\Rightarrow$) Suppose we have an assignment AS of the n variables that satisfies the boolean formula of I. By definition, for each clause there is at least one literal that satisfies it. In the following, j_i will define, for any $1 \leq i \leq q$, the smallest index of the literal of c_i (i.e. 1, 2 or 3) which, by its assignment, satisfies c_i. Let (S, P) and (T, Q) be two sequences obtained from an APS-cp-construction from I. We look for a set $\mathcal{B}$ of bases to delete from S in order to obtain T. For each variable $x_m \in AS$ with $1 \leq m \leq n$, we define $\mathcal{B}$ as follows:

- if $x_m = True$ then $\mathcal{B}$ contains each base of $S^s_{\overline{x_m}}$ and $S^e_{x_m}[1]$,
- if $x_m = False$ then $\mathcal{B}$ contains each base of $S^s_{x_m}$ and $S^e_{x_m}[2]$,
- if $j_i = 1$ then $\mathcal{B}$ contains $S_{c_i}[3]$ and $S_{c_i}[4]$,
- if $j_i = 2$ then $\mathcal{B}$ contains $S_{c_i}[2]$ and $S_{c_i}[4]$,
- if $j_i = 3$ then $\mathcal{B}$ contains $S_{c_i}[2]$ and $S_{c_i}[3]$.

Since a variable has a unique value (i.e. $True$ or $False$), either each base of $S^s_{\overline{x_m}}$ and $S^e_{x_m}[1]$ or each base of $S^s_{x_m}$ and $S^e_{x_m}[2]$ are in $\mathcal{B}$ for all $1 \leq m \leq n$. Thus, $\mathcal{B}$ contains at least one base in S of any AU-arc of P.

For any $1 \leq i \leq q$, two of the three bases G of S_{c_i} are in $\mathcal{B}$. Thus, $\mathcal{B}$ contains at least one base in S of two thirds of the CG-arcs of P. Moreover, $S_{c_i}[j_i + 1]$ is the base G that is not in $\mathcal{B}$. We suppose in the following that the j_i^{th} literal of the clause c_i is x_m, with $1 \leq m \leq n$. Thus, by the way we build the APS-cp-construction, there is an arc between a base C of $S^s_{\overline{x_m}}$ and $S_{c_i}[j_i + 1]$ in P. By definition, if AS is an assignment of the n variables that satisfies the boolean formula, AS satisfies c_i and thus $x_m = True$. We mentioned, in the definition of $\mathcal{B}$ that if $x_m = True$ then each base of $S^s_{\overline{x_m}}$ is in $\mathcal{B}$. Thus, the base C of $S^s_{\overline{x_m}}$ incident to the CG-arc in P with $S_{c_i}[j_i+1]$ is in $\mathcal{B}$. A similar result can be found if the j_i^{th} literal of the clause c_i is $\overline{x_m}$. Thus, $\mathcal{B}$ contains at least one base in S of any CG-arc of P.

If S' is the sequence obtained from S by deleting all the bases of $\mathcal{B}$ together with their incident arcs, then there is no arc in S' (i.e. neither AU-arcs or CG-arcs). By the way we define $\mathcal{B}$, S' is obtained from S by deleting all the bases of either $S^s_{x_m}$ or $S^s_{\overline{x_m}}$, two bases G of S_{c_i} and either $S^e_{x_m}[1]$ or $S^e_{x_m}[2]$, for $1 \leq i \leq q$ and $1 \leq m \leq n$. According to Lemma 1, it is easily seen that sequence S' obtained is similar to T.

($\Leftarrow$) Let I be an instance of the problem 3-Sat with n variables and q clauses. Let I' be an instance $((S, P); (T, Q))$ of APS$(\{\sqsubset, \between\}, \emptyset)$ obtained by an APS-cp-construction from I such that (T, Q) can be obtained from (S, P) by deleting some of its bases (i.e. a set of bases $\mathcal{B}$) together with their incident arcs, if any. By Lemma 1, either all bases of $S^s_{\overline{x_m}}$ or all bases of $S^s_{x_m}$ are in $\mathcal{B}$. Consequently, for $1 \leq m \leq n$, we define an assignment AS of the n variables of I as follows:

- if all bases of $S^s_{\overline{x_m}}$ are in $\mathcal{B}$ then $x_m = True$,
- if all bases of $S^s_{x_m}$ are in $\mathcal{B}$ then $x_m = False$.

Now, let us prove that for any $1 \leq i \leq q$ the clause c_i is satisfied by AS. By Lemma 1, for any $1 \leq i \leq q$ there is a base G of substring S_{c_i} (say the $j_i + 1^{th}$) that is not in $\mathcal{B}$. By the the way we build the APS-cp-construction, there is a

CG-arc in P between $S_{c_i}[j_i+1]$ and a base C of $S^s_{\overline{x_m}}$ (resp. $S^s_{x_m}$) if the j_i^{th} literal of c_i is x_m (resp. $\overline{x_m}$).

Suppose, *w.l.o.g.*, that the j_i^{th} literal of c_i is x_m. Since Q is an empty set, at least one base of any arc of P is in $\mathcal{B}$. Thus, the base C of $S^s_{\overline{x_m}}$ incident to the CG-arc in P with $S_{c_i}[j_i+1]$ is in $\mathcal{B}$ (since $S_{c_i}[j_i+1] \notin \mathcal{B}$). Therefore, by Lemma 1, all the bases of $S^s_{\overline{x_m}}$ are in $\mathcal{B}$. By the way we define AS, $x_m = True$ and thus c_i is satisfied. The same conclusion can be similarly derived if the j_i^{th} literal of c_i is $\overline{x_m}$. □

We have thus proved the following theorem.

Theorem 1. *The* APS$(\{\sqsubset, \between\}, \emptyset)$ *problem is* **NP***-complete.*

It follows immediately from Theorem 1 that the APS$(\{<, \sqsubset, \between\}, \emptyset)$ problem, and hence the classical APS(Crossing, Plain) problem, is **NP**-complete.

One might naturally ask for more information concerning the hard cases of the APS problem in order to get valuable insight into what makes the problem difficult. Another refinement of Theorem 1 is given by the following theorem.

Theorem 2. *The* APS$(\{<, \between\}, \emptyset)$ *problem is* **NP***-complete.*

As for Theorem 1, the proof is by reduction from the 3-Sat problem. It is easily seen that the APS$(\{<, \between\}, \emptyset)$ problem is in **NP**. The remainder of this section is devoted to proving that it is also **NP**-hard. Let $\mathcal{V}_n = \{x_1, x_2, ...x_n\}$ be a finite set of n variables and $\mathcal{C}_q = \{c_1, c_2, \dots, c_q\}$ a collection of q clauses. The instance of the APS$(\{<, \between\}, \emptyset)$ problem we will build is decomposed in two parts: a *Truth Setting part* and a *Checking part*. For readability, we denote by APS2-cp-construction any construction of the type described hereafter. Moreover, we will present separately the *Truth Setting part* and the *Checking part*: first, we will describe the *Truth Setting part*, then the *Checking part* and end by the description of the set of arcs connecting those two parts. Indeed, the instance of the APS$(\{<, \between\}, \emptyset)$ problem will be the concatenation of those two parts.

Truth Setting part

Let us first detail the construction of sequences S' and T' of the *Truth Setting part*:

$$\begin{aligned} S' &= \overbrace{S^e_{x_1}\ S^e_{x_2} \dots S^e_{x_n}}^{S_\alpha}\ \mathbf{GGG}\ \ \overbrace{S^s_{x_1} A\ S^s_{\overline{x_1}}\ S^s_{x_2} A\ S^s_{\overline{x_2}} \dots S^s_{x_n} A\ S^s_{\overline{x_n}}}^{S_\beta} \\ T' &= \underbrace{T^e_{x_1}\ T^e_{x_2} \dots T^e_{x_n}}_{T_{\alpha'}}\ \mathbf{GGG}\ \underbrace{T^s_{x_1}\ T^s_{x_2} \dots T^s_{x_n}}_{T_{\beta'}} \end{aligned}$$

We now detail subsequences that compose S' and T'. Let γ_m (resp. $\gamma_{\overline{m}}$) be the number of occurrences of literal x_m (resp. $\overline{x_m}$) in $\mathcal{C}_q$ and let $k_m = \max(\gamma_m, \gamma_{\overline{m}})$. For each variable $x_m \in \mathcal{V}_n$, we construct substrings $S^e_{x_m} = UUA$, $T^e_{x_m} = UA$, $S^s_{x_m} = AC^{k_m}$, $S^s_{\overline{x_m}} = C^{k_m}A$ and $T^s_{x_m} = AC^{k_m}A$, where C^{k_m} represents a substring of k_m consecutive bases C. Having disposed of the two sequences, we now turn to defining the corresponding arc structure (see Figure 2). For all $1 \leq$

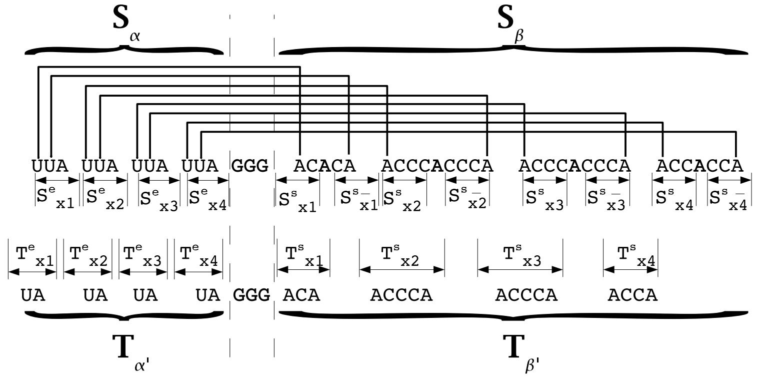


Fig. 2. The truth setting part of an APS2-CP-construction with $\mathcal{C}_q = (x_2 \vee \overline{x_3} \vee x_4) \wedge (x_1 \vee x_2 \vee x_3) \wedge (\overline{x_2} \vee x_3 \vee \overline{x_4})$.

$m \leq n$, we create the two following arcs: ($S^e_{x_m}[1]$,$S^s_{x_m}[1]$) and ($S^e_{x_m}[2]$,$S^s_{\overline{x_m}}[k_m + 1]$). Remark that, by now, all the arcs defined are $\{\between\}$-comparable.

Checking part

Let us now detail the construction of sequences S_ζ and $T_{\zeta'}$ of the *Checking part*:

$$S_\zeta = U \ \overbrace{S^1_{x_1}\ S^1_{x_2}...S^1_{x_n}}^{S^1}\ U\ \overbrace{S^{\overline{1}}_{x_1}\ S^{\overline{1}}_{x_2}...S^{\overline{1}}_{x_n}}^{S^{\overline{1}}}\ U...U\ \overbrace{S^q_{x_1}\ S^q_{x_2}...S^q_{x_n}}^{S^q}\ U\ \overbrace{S^{\overline{q}}_{x_1}\ S^{\overline{q}}_{x_2}...S^{\overline{q}}_{x_n}}^{S^{\overline{q}}}\ U$$
$$T_{\zeta'} = U\ T^1\ U\ T^{\overline{1}}\ U...U\ T^q\ U\ T^{\overline{q}}\ U$$

We now detail subsequences that compose S_ζ and $T_{\zeta'}$. For any $1 \leq m \leq n$ and any $1 \leq i \leq q$, let γ^i_m (resp. $\gamma^i_{\overline{m}}$) be the number of occurrences of literal x_m (resp. $\overline{x_m}$) in the set of clauses c_j with $i < j \leq q$ and let $\lambda^i_m = \gamma^i_m + \gamma^i_{\overline{m}}$. For any $1 \leq m \leq n$ and for any $1 \leq i \leq q$, let $y^i_m = 1$ if $x_m \in c_i$, $y^i_m = 0$ otherwise. For any $1 \leq m \leq n$ and for any $1 \leq i \leq q$, let $y^i_{\overline{m}} = 1$ if $\overline{x_m} \in c_i$, $y^i_{\overline{m}} = 0$ otherwise. For any $1 \leq m \leq n$ and $1 \leq i \leq q$, we construct substrings:

$$\begin{aligned} S^i_{x_m} &= (GGA)^{\lambda^i_m + y^i_{\overline{m}}}(GA)^{y^i_m}(GGA)^{\lambda^i_m + y^i_m}(GA)^{y^i_{\overline{m}}} \\ S^{\overline{i}}_{x_m} &= (CCA)^{\lambda^i_m}(CA)^{y^i_{\overline{m}}}(CCA)^{\lambda^i_m}(CA)^{y^i_m} \\ T^i &= (GA)^{4+6q-6i} \\ T^{\overline{i}} &= (CA)^{2+6q-6i} \end{aligned}$$

For example, assuming that $\mathcal{C}_q = (x_2 \vee \overline{x_3} \vee x_4) \wedge (x_1 \vee x_2 \vee x_3) \wedge (\overline{x_2} \vee x_3 \vee \overline{x_4})$ we have, among others, the following segments:

$$\begin{aligned}
S^1_{x_1} &= (GGA)^1(GA)^0(GGA)^1(GA)^0 = GGA\ GGA\\
S^1_{x_2} &= (GGA)^2(GA)^1(GGA)^3 = GGA\ GGA\ GA\ GGA\ GGA\ GGA\\
S^{\overline{2}}_{x_3} &= (CCA)^1(CA)^0(CCA)^1(CA)^1 = CCA\ CCA\ CA\\
T^2 &= (GA)^{4+6*3-6*2} = GA\ GA\ GA\ GA\ GA\ GA\ GA\ GA\ GA\ GA\\
T^{\overline{3}} &= (CA)^{2+6*3-6*3} = CA\ CA
\end{aligned}$$

Having disposed of the two sequences, we now turn to defining the corresponding arc structure (see Figure 3). By construction, $S^i_{x_m}$ (resp. $S^{\overline{i}}_{x_m}$) is composed of substrings GA and GGA (resp. CA and CCA). We denote by *repeater* any substring GGA or CCA. We denote by *terminal* any substring GA or CA which is not part of a repeater. Let $term(i, m, j)$ (resp. $rep(i, m, j)$) be the j^{th} terminal (resp. repeater) of $S^i_{x_m}$, and let $term(\overline{i}, m, j)$ (resp. $rep(\overline{i}, m, j)$) be the j^{th} terminal (resp. repeater) of $S^{\overline{i}}_{x_m}$.

For all $1 \leq m \leq n$, $1 \leq j \leq 2\lambda^i_m + 1$ and $1 \leq i < q$, we create the following arcs:

- an arc between the second base G of $rep(i, m, j)$ and the first base C of the j^{th} element (i.e. either a terminal or a repeater) of $S^{\overline{i}}_{x_m}$;
- an arc between the second base C of $rep(\overline{i}, m, j)$ and the first base G of the j^{th} element of $S^{i+1}_{x_m}$.

Final Construction

Final sequences S and T are respectively obtained by concatenating S' with S_ζ and T' with $T_{\zeta'}$. Moreover, we create, for all $1 \leq m \leq n$ and all $1 \leq j \leq \gamma_m + \gamma_{\overline{m}}$, an arc between the j^{th} base C of substring $S^s_{x_m} A S^s_{\overline{x_m}}$ in S' and the first base G of the j^{th} element of $S^1_{x_m}$ in S_ζ. In the rest of the paper, S^i will refer to $S^i_{x_1}\ S^i_{x_2}\ \ldots\ S^i_{x_n}$ and $S^{\overline{i}}$ will refer to $S^{\overline{i}}_{x_1}\ S^{\overline{i}}_{x_2} \ldots S^{\overline{i}}_{x_n}$.

In the following, we will show that P is $\{<, \between\}$-comparable. Let a_1 and a_2 be any two arcs connecting a base of S_β to a base of S_ζ. As all the arcs connecting a base of S_β to a base of S_ζ are of the same form, we consider, w.l.o.g. that:

- for a given j and a given $1 \leq m \leq n$, a_1 is the arc which connects the j^{th} base C of substring $S^s_{x_m} A S^s_{\overline{x_m}}$ to the first base G of the j^{th} element of $S^1_{x_m}$;
- for a given k and a given $1 \leq m' \leq n$, a_2 is the arc which connects the k^{th} base C of substring $S^s_{x_{m'}} A S^s_{\overline{x_{m'}}}$ to the first base G of the k^{th} element of $S^1_{x_{m'}}$;
- $j < k$.

We now consider the three following cases: (i) $m = m'$, (ii) $m < m'$ and (iii) $m > m'$. Suppose $m = m'$. As $j < k$, the j^{th} base C precedes the k^{th} base C of substring $S^s_{x_m} A S^s_{\overline{x_m}}$. Moreover, the first base G of the j^{th} element of

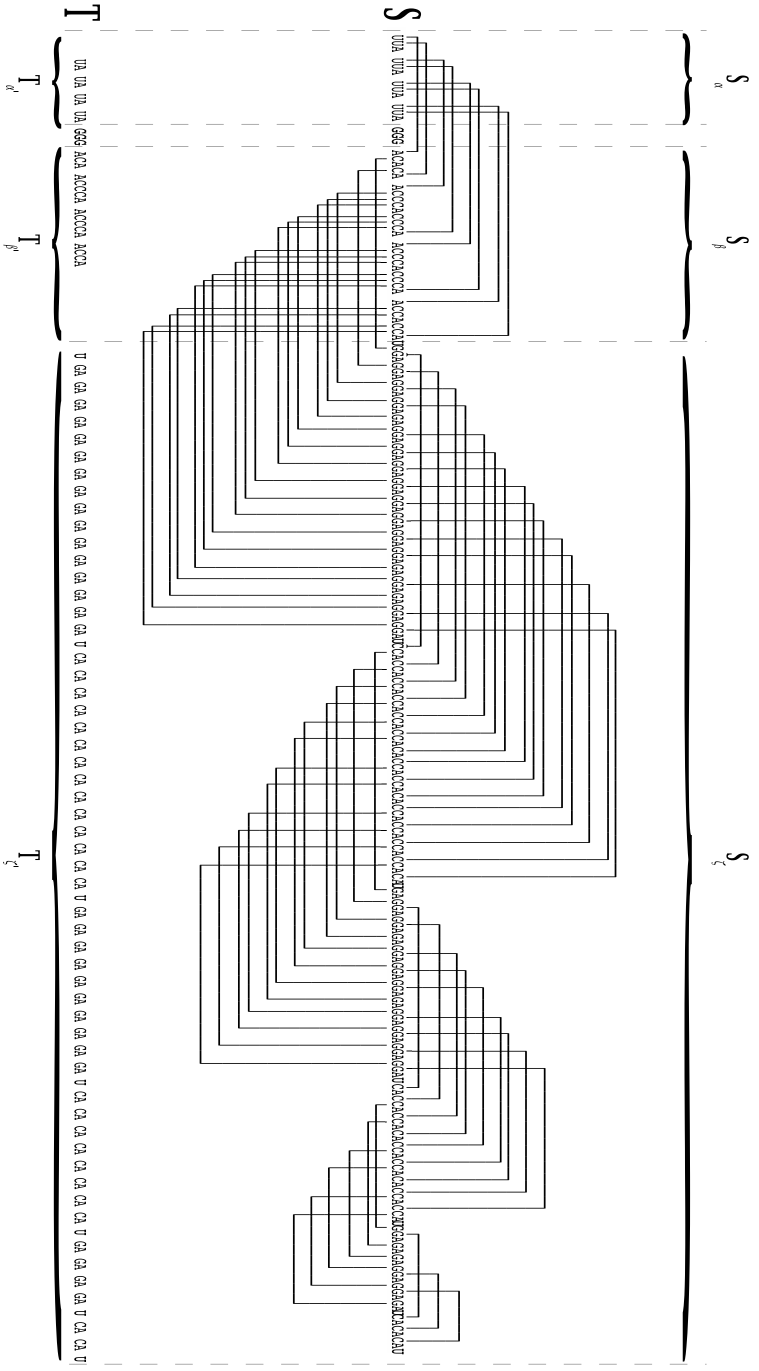


Fig. 3. Example of an APS2-cp-construction with $\mathcal{C}_q = (x_2 \vee \overline{x_3} \vee x_4) \wedge (x_1 \vee x_2 \vee x_3) \wedge (\overline{x_2} \vee x_3 \vee \overline{x_4})$.

$S^1_{x_m}$ precedes the first base G of the k^{th} element of $S^1_{x_m}$. Thus, a_1 and a_2 are $\{\between\}$-comparable.

Suppose now $m < m'$. Then, the j^{th} base C of substring $S^s_{x_m} A S^s_{\overline{x_m}}$ precedes the k^{th} base C of substring $S^s_{x_{m'}} A S^s_{\overline{x_{m'}}}$. Moreover, the first base G of the j^{th} element of $S^1_{x_m}$ precedes the first base G of the k^{th} element of $S^1_{x_{m'}}$. Thus, a_1 and a_2 are $\{\between\}$-comparable. The case where $m > m'$ is fully similar. Therefore, given two arcs a_1 and a_2 connecting a base of S_β and a base of S_ζ, a_1 and a_2 are $\{\between\}$-comparable, and thus, $\{<, \between\}$-comparable.

Let a_1 and a_2 be any two arcs connecting two bases of S_ζ. There are two types of arcs connecting two bases of S_ζ:

1. arcs connecting, for a given $1 \leq i \leq q$ and a given j, a base of the j^{th} repeater of S^i to a base of the j^{th} element of $S^{\overline{i}}$;
2. arcs connecting, for a given $1 \leq i < q$ and a given j, a base of the j^{th} repeater of $S^{\overline{i}}$ to a base of the j^{th} element of S^{i+1}.

By definition, a_1 and a_2 can be either of type 1 or type 2. Since the cases where a_1 and a_2 are of different types are fully similar, we detail hereafter three cases: (a) a_1 and a_2 are of type 1, (b) a_1 is of type 1 and a_2 is of type 2, and (c) a_1 and a_2 are of type 2.

(a) Suppose that a_1 and a_2 are of type 1. Since a_2 is of type 1, a_2 connects, for a given $1 \leq i' \leq q$ and a given k, a base of the k^{th} repeater of $S^{i'}$ to a base of the k^{th} element of $S^{\overline{i'}}$. Suppose, w.l.o.g., that $j < k$. By construction, if $i \neq i'$ then either a_1 precedes a_2 or a_2 precedes a_1. Therefore, if $i \neq i'$ then a_1 and a_2 are $\{<\}$-comparable. Moreover, if $i = i'$ then a_1 and a_2 are $\{\between\}$-comparable.

(b) Suppose that a_1 is of type 1 and a_2 is of type 2. Since a_2 is of type 2, a_2 connects, for a given $1 \leq i' \leq q$ and a given k, a base of the k^{th} repeater of $S^{\overline{i'}}$ to a base of the k^{th} element of $S^{i'+1}$. By construction, if $i \neq i'$ then either a_1 precedes a_2 or a_2 precedes a_1. Therefore, if $i \neq i'$ then a_1 and a_2 are $\{<\}$-comparable. Consider now the case where $i = i'$. Suppose first that $j < k$. If $i = i'$ then, as $S^{\overline{i}}$ precedes S^{i+1} and $j < k$, a_1 and a_2 are $\{<\}$-comparable. Suppose now that $j > k$. If $i = i'$ then, as $S^{\overline{i}}$ precedes S^{i+1} and $k < j$, a_1 and a_2 are $\{\between\}$-comparable.

(c) Suppose that a_1 and a_2 are of type 2. Since a_2 is of type 2, a_2 connects, for a given $1 \leq i' \leq q$ and a given k, a base of the k^{th} repeater of $S^{\overline{i'}}$ to a base of the k^{th} element of $S^{i'+1}$. Suppose, w.l.o.g., that $j < k$. By construction, if $i \neq i'$ then either a_1 precedes a_2 or a_2 precedes a_1. Therefore, if $i \neq i'$ then a_1 and a_2 are $\{<\}$-comparable. Moreover, if $i = i'$ then a_1 and a_2 are $\{\between\}$-comparable.

Therefore, given two arcs a_1 and a_2 connecting two bases of S_ζ, a_1 and a_2 are $\{<, \between\}$-comparable. We now turn to proving that the set P is $\{<, \between\}$-comparable. Notice, first, that there is no arc connecting two bases of S_β (resp. S_α). We proved previously that given two arcs a_1 and a_2 connecting a base of S_β

and a base of S_ζ, a_1 and a_2 are $\{<, \between\}$-comparable. Finally, we proved that given two arcs a_1 and a_2 connecting a base of S_α and a base of S_β, a_1 and a_2 are $\{\between\}$-comparable.Therefore, the set of arcs starting in $S_\alpha \bigcup S_\beta$ is $\{<, \between\}$-comparable.

Let $a_\zeta = (u', v')$, where u' and v' are bases, denote the arc connecting a base of S_β to a base of S_ζ and which ends the last. By construction, all the arcs connecting two bases of S_ζ are ending after v'. Therefore, the set of arcs in S (*i.e.* the set P) is $\{<, \between\}$-comparable.

A full illustration of an APS2-CP-construction is given in Figure 3. Clearly, our construction can be carried out in polynomial time. Moreover, the result of such a construction is indeed an instance of APS$(\{<, \between\}, \emptyset)$, since $Q = \emptyset$ (no arc is added to T) and P is a $\{<, \between\}$-comparable set of arcs.

Let (S, P) and (T, Q) be two sequences obtained from an APS2-CP-construction. In the following, we will give some technical lemmas that will be useful for the comprehension of proof of Theorem 2.

Definition 1. *A canonical alignment of two sequences (S, P) and (T, Q) obtained from an* APS2-CP*-construction is an alignment where, for any $1 \leq i \leq q$ and $1 \leq m \leq n$:*

- *any base of $S^e_{x_m}$ is either matched with a base of $T^e_{x_m}$ or deleted,*
- *either each base of $S^s_{x_m}A$ is matched with a base of $T^s_{x_m}$ and all bases of $S^s_{\overline{x_m}}$ are deleted, or each base of $AS^s_{\overline{x_m}}$ is matched with a base of $T^s_{x_m}$ and all bases of $S^s_{x_m}$ are deleted,*
- *any base of S^i is either matched with a base of T^i or deleted,*
- *any base of $S^{\bar{i}}$ is either matched with a base of $T^{\bar{i}}$ or deleted.*

Lemma 3. *Let (S, P) and (T, Q) be two sequences obtained from an* APS2-CP*-construction. If (T, Q) is an arc-preserving subsequence of (S, P) then any corresponding alignment is canonical.*

Proof. Suppose (T, Q) is an arc-preserving subsequence of (S, P). Let $\mathcal{A}$ denote any corresponding alignment. In T, there is a substring GGG between $T_{\alpha'}$ and $T_{\beta'}$. In S, bases G are present either between S_α and S_β, or in S_ζ. The number of bases U in S_ζ and in $T_{\zeta'}$ is equal. Moreover, in both S_ζ and $T_{\zeta'}$ the first (*i.e.* leftmost) base is a base U. Therefore, in $\mathcal{A}$, none of the bases of the substring GGG in T between $T_{\alpha'}$ and $T_{\beta'}$ can be matched to a base G of S_ζ since, in that case, at least one base U of $T_{\zeta'}$ would not be matched. Thus, in $\mathcal{A}$, substring GGG of S has to be matched with substring GGG of T and $T_{\alpha'}$ must be matched with substrings of S_α.

Moreover, the number of bases U in S_ζ and in $T_{\zeta'}$ is equal; besides, in S_β and $T_{\beta'}$ there is no base U. Thus, $T_{\beta'}$ (resp. $T_{\zeta'}$) must be matched with substrings of S_β (resp. S_ζ). Therefore, we will consider the three cases ($S_\alpha/T_{\alpha'}$, $S_\beta/T_{\beta'}$, $S_\zeta/T_{\zeta'}$) separately.

Consider S_α and $T_{\alpha'}$. There are exactly n bases A both in S_α and $T_{\alpha'}$. Consequently, in $\mathcal{A}$, for all $1 \leq m \leq n$, $S^e_{x_m}$ has to be matched with $T^e_{x_m}$. More precisely, $T^e_{x_m}[1]$ has to be matched to either $S^e_{x_m}[1]$ or $S^e_{x_m}[2]$ for all $1 \leq m \leq n$.

Consider S_β and $T_{\beta'}$. By definition, as $Q = \emptyset$, at least one base incident to every arc of P has to be deleted. We just mentioned that $T^e_{x_m}[1]$ has to be matched to either $S^e_{x_m}[1]$ or $S^e_{x_m}[2]$ for any $1 \leq m \leq n$. Thus, since by construction there is an arc between $S^e_{x_m}[1]$ and $S^s_{x_m}[1]$ (resp. $S^e_{x_m}[2]$ and $S^s_{\overline{x_m}}[k_m + 1]$), for any $1 \leq m \leq n$, either $S^s_{x_m}[1]$ or $S^s_{\overline{x_m}}[k_m + 1]$ is deleted. Therefore, n bases A appearing in S_β are deleted. Note that there are $3n$ bases A in S_β and $2n$ in $T_{\beta'}$. Thus, the number of bases A not deleted in S_β is equal to the number of bases A in $T_{\beta'}$. Since, for each $1 \leq m \leq n$, a base A of either $S^s_{x_m}$ or $S^s_{\overline{x_m}}$ is deleted, we conclude that for each $1 \leq m \leq n$, $T^s_{x_m}$ is obtained from $S^s_{x_m} A S^s_{\overline{x_m}}$, by deleting all bases of either $S^s_{x_m}$ or $S^s_{\overline{x_m}}$.

Consider S_ζ and $T_{\zeta'}$. By construction, there are $2q + 1$ bases U in S_ζ and in $T_{\zeta'}$. Thus, in $\mathcal{A}$, the $2q + 1$ bases U of S_ζ have to be matched with the $2q + 1$ bases U of $T_{\zeta'}$. Therefore, in $\mathcal{A}$, for any $1 \leq i \leq q$, any base of S^i is either matched with a base of T^i or deleted, and any base of $S^{\overline{i}}$ is either matched with a base of $T^{\overline{i}}$ or deleted. □

In the following, given an alignment $\mathcal{A}$ of S and T, if the first base of a terminal is matched (resp. deleted) in $\mathcal{A}$ then the corresponding terminal will be denoted as *active* (resp. *inactive*). Similarly, a repeater is said to be inactive (resp. active) when its two first bases (resp. exactly one out of its two first bases) are deleted in $\mathcal{A}$. Notice that the case where none of the two first bases of a repeater is deleted in $\mathcal{A}$ is not considered.

Notice that, by construction, for any $1 \leq i \leq q$, there are no two consecutive bases G in $T_{\zeta'}$, and there are no two consecutive bases C in $T_{\zeta'}$. Thus, at least one out of any two consecutive bases C or G of S_ζ is deleted in $\mathcal{A}$. Therefore, given a canonical alignment, for any repeater of S, either the repeater is active or all its bases C or G are deleted.

Lemma 4. *Let (S, P) and (T, Q) be two sequences obtained from an* APS2-CP-*construction. If (T, Q) is an arc-preserving subsequence of (S, P), then for any corresponding alignment $\mathcal{A}$ and for any $1 \leq i \leq q$, one of the three following cases must occur:*

- *all the repeaters and one terminal of S^i are active,*
- *all the repeaters but one and two terminals of S^i are active,*
- *all the repeaters but two and three terminals of S^i are active.*

Proof. By Lemma 3, $\mathcal{A}$ is canonical. Moreover, by definition, in any canonical alignment, for all $1 \leq i \leq q$, any base of S^i is either matched with a base of T^i or deleted. Let ω_j (resp. ω'_j) denote the j^{th} element of S^i (resp. T^i).

By construction, in T^i, there are two bases A less than in S^i. Therefore, we know that in $\mathcal{A}$, all the bases A of S^i but two will be matched. Let ω_k and ω_l, with $k < l$, denote the two elements of S^i which contain the deleted bases A. There are two cases, as illustrated in Figure 4: either (a) $l = k + 1$ or (b) $l > k + 1$. Let us consider those two cases separately.

(a) Suppose $l = k + 1$ (*i.e.* ω_k and ω_l are consecutive). In that case, since all the bases A but two will be matched in S^i, the base A of ω_{k-1} (resp. ω_{l+1})

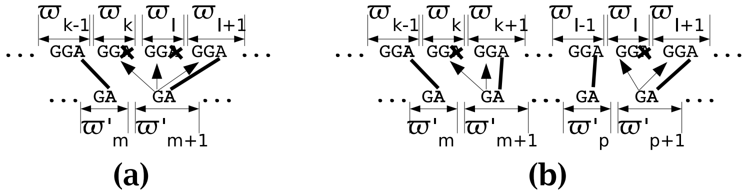


Fig. 4. Illustration of Lemma 4. (a) $l = k + 1$ or (b) $l > k + 1$.

is matched with a base A of an element of T^i, say ω'_m (resp. ω'_{m+1}). Therefore, the base G of ω'_{m+1} is either matched with a base of ω_k, ω_l or ω_{l+1}. In each of those cases, all the elements but two of S^i are active.

(b) Suppose $l > k+1$ (*i.e.* ω_k and ω_l are not consecutive). In that case, since all the bases A but two will be matched in S^i, the base A of ω_{k-1} (resp. ω_{k+1}) is matched with a base A of an element of T^i, say ω'_m (resp. ω'_{m+1}). Similarly, the base A of ω_{l-1} (resp. ω_{l+1}) is matched with a base A of an element of T^i, say ω'_p (resp. ω'_{p+1}). Therefore, the base G of ω'_{m+1} (resp. ω'_{p+1}) is either matched with a base of ω_k or ω_{k+1} (resp. ω_l or ω_{l+1}). In each of those cases, all the elements but two of S^i are active.

Therefore, either two terminals, or one repeater and one terminal, or two repeaters of S^i are inactive. □

Lemma 5. *Let (S, P) and (T, Q) be two sequences obtained from an* APS2-CP-*construction. If (T, Q) is an arc-preserving subsequence of (S, P), then for any corresponding alignment $\mathcal{A}$, all the repeaters and two terminals of $S^{\overline{1}}$ are active.*

Proof. Note that in this lemma, we focus on the first clause (*i.e.* c_1). c_1 is defined by three literals (say x_i, x_j and x_k). Since c_1 is equal to the disjunction of variables built with x_i, x_j and x_k, c_1 can have eight different forms, because each literal can appear in either its positive (x_i) or negative ($\overline{x_i}$) form. In the following, we suppose, to illustrate the proof, that $c_1 = (x_i \vee x_j \vee \overline{x_k})$ as illustrated in Figure 5. The other cases will not be considered here, but can be treated similarly.

By Lemma 3, $\mathcal{A}$ is canonical. Moreover, by definition, in any canonical alignment, for all $1 \le i \le q$, any base of $S^{\overline{i}}$ is either matched with a base of $T^{\overline{i}}$ or deleted. We recall that ω_j (resp. ω'_j) denotes the j^{th} element of $S^{\overline{i}}$ (resp. $T^{\overline{i}}$).

By construction, in $T^{\overline{1}}$, there is one base A less than in $S^{\overline{1}}$. Therefore, we know that in $\mathcal{A}$, all the bases A of $S^{\overline{1}}$ but one will be matched. Let ω_k denote the element of $S^{\overline{1}}$ which contains the deleted base A. Since all the bases A of $S^{\overline{1}}$ but two will be matched, the base A of ω_{k-1} (resp. ω_{k+1}) is matched with a base A of an element of $T^{\overline{1}}$, say ω'_m (resp. ω'_{m+1}). Therefore, the base C of ω'_{m+1} is either matched with a base of ω_k or ω_{k+1}. Consequently, all the elements but one of $S^{\overline{1}}$ are active.

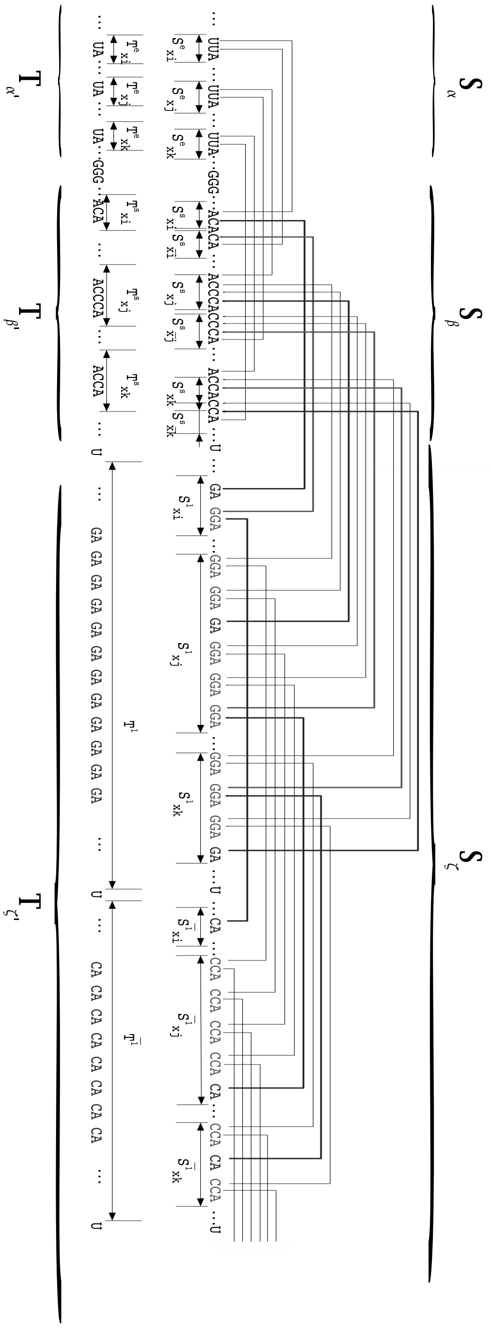

S_α
S_β
S_ζ
T_α'
T_β'
T_ζ'
UUA … UUA … UUA … GGG … ACACA … ACCCACCCA … ACCACCA … U … GA GGA … GGA GGA GA GGA GGA GGA … GGA GGA GGA GA … U … CA … CCA CCA CCA CCA CA … CCA CA CCA … U
UA … UA … UA … GGG … ACA … ACCCA … ACCA … U … GA GA GA GA GA GA GA GA GA GA GA GA GA … U … CA CA CA CA CA CA CA CA CA … U
S^e_xi
S^e_xj
S^e_xk
S^s_xi
S^s_x̄i
S^s_xj
S^s_x̄j
S^s_xk
S^s_x̄k
S^l_xi
S^l_xj
S^l_xk
S^l̄_xi
S^l̄_xj
S^l̄_xk
T^e_xi
T^e_xj
T^e_xk
T^s_xi
T^s_xj
T^s_xk
T^l
T^l̄

Fig. 5. Part of an APS2-CP-construction corresponding to a clause $c_1 = (x_i \vee x_j \vee \overline{x_k})$. Bold arcs correspond to the different cases studied in Lemma 5.

To prove that the inactive element is a terminal, we suppose, by contradiction, that one repeater of $S^{\overline{1}}$ is inactive. Therefore, the three terminals of $\{S^{\overline{1}}_{x_i}, S^{\overline{1}}_{x_j}, S^{\overline{1}}_{x_k}\}$ are active. Moreover, by Lemma 4, either:

1. all the repeaters of S^1 and one terminal of $\{S^1_{x_i}, S^1_{x_j}, S^1_{x_k}\}$ are active,
2. all the repeaters but one of S^1 and two terminals of $\{S^1_{x_i}, S^1_{x_j}, S^1_{x_k}\}$ are active,
3. all the repeaters but two of S^1 and three terminals of $\{S^1_{x_i}, S^1_{x_j}, S^1_{x_k}\}$ are active.

Let us consider those three cases separately:

(1) Suppose that all the repeaters of S^1 and one terminal of $\{S^1_{x_i}, S^1_{x_j}, S^1_{x_k}\}$ are active. The active terminal can be in either $S^1_{x_i}$, $S^1_{x_j}$ or $S^1_{x_k}$. We recall that the clause considered is $c_1 = (x_i \vee x_j \vee \overline{x_k})$. Since the cases where the active terminal is either in $S^1_{x_i}$ or $S^1_{x_j}$ are fully similar, we detail hereafter only two cases: (a) the active terminal is in $S^1_{x_i}$ and (b) the active terminal is in $S^1_{x_k}$.

(a) Suppose that the active terminal is in $S^1_{x_i}$. By construction, there is a repeater *rep* of $S^1_{x_i}$ such that $(\delta, rep[1]) \in P$, $(rep[2], \theta) \in P$ where δ (resp. θ) is a base C of $S^s_{\overline{x_i}}$ (resp. the first base of the terminal in $S^{\overline{1}}_{x_i}$), as illustrated in Figure 5. Since, by hypothesis, the three terminals of $\{S^{\overline{1}}_{x_i}, S^{\overline{1}}_{x_j}, S^{\overline{1}}_{x_k}\}$ are active, then θ is matched. By definition, as $Q = \emptyset$, at least one base incident to every arc of P has to be deleted. Therefore, $rep[2]$ is deleted. Since *rep* is an active repeater, $rep[1]$ is matched. Thus, δ is deleted. Moreover, by construction, there is an arc between a base C of $S^s_{x_i}$ and the first base of the terminal in $S^1_{x_i}$ (cf. Figure 5). Therefore, since the first base of terminal in $S^1_{x_i}$ is matched (because we supposed that the active terminal is in $S^1_{x_i}$), a base C of $S^s_{x_i}$ is deleted. Thus, a base of both $S^s_{x_i}$ and $S^s_{\overline{x_i}}$ is deleted. Therefore, by Definition 1, the alignment is not canonical, a contradiction.

(b) Suppose now that the active terminal is in $S^1_{x_k}$. By construction, there is a repeater *rep* of $S^1_{x_k}$ such that $(\delta, rep[1]) \in P$, $(rep[2], \theta) \in P$ where δ (resp. θ) is a base C of $S^s_{x_k}$ (resp. the first base of the terminal in $S^{\overline{1}}_{x_k}$), as illustrated in Figure 5. Since, by hypothesis, the three terminals of $\{S^{\overline{1}}_{x_i}, S^{\overline{1}}_{x_j}, S^{\overline{1}}_{x_k}\}$ are active, then θ is matched. By definition, as $Q = \emptyset$, at least one base incident to every arc of P has to be deleted. Therefore, $rep[2]$ is deleted. Since *rep* is an active repeater, $rep[1]$ is matched. Thus, δ is deleted. Moreover, by construction, there is an arc between a base C of $S^s_{\overline{x_k}}$ and the first base of the terminal in $S^1_{x_k}$ (cf. Figure 5). Therefore, since the first base of terminal in $S^1_{x_k}$ is matched (because we supposed that the active terminal is in $S^1_{x_k}$), a base C of $S^s_{\overline{x_k}}$ is deleted. Thus, a base of both $S^s_{x_k}$ and $S^s_{\overline{x_k}}$ is deleted. Therefore, by Definition 1, the alignment is not canonical, a contradiction.

(2) Suppose that all the repeaters but one of S^1 and two terminals of $\{S^1_{x_i}$, $S^1_{x_j}$, $S^1_{x_k}\}$ are active. The active terminals can be in either ($S^1_{x_i}$, $S^1_{x_j}$), ($S^1_{x_i}$, $S^1_{x_k}$) or ($S^1_{x_j}$, $S^1_{x_k}$). Since the cases where the active terminals are either in ($S^1_{x_i}$, $S^1_{x_k}$) or ($S^1_{x_j}$, $S^1_{x_k}$) are fully similar, we detail hereafter only two cases: (a) the active terminals are in ($S^1_{x_i}$, $S^1_{x_j}$) and (b) the active terminals are in ($S^1_{x_i}$, $S^1_{x_k}$).

(a) Suppose that the active terminals are in ($S^1_{x_i}$, $S^1_{x_j}$). By construction, there is a repeater *rep* of $S^1_{x_i}$ such that $(\delta, rep[1]) \in P$, $(rep[2], \theta) \in P$ where δ (resp. θ) is a base C of $S^s_{\overline{x_i}}$ (resp. the first base of the terminal in $S^{\overline{1}}_{x_i}$), as illustrated in Figure 5. Similarly, by construction, there is a repeater *rep*$'$ of $S^1_{x_j}$ such that $(\delta', rep'[1]) \in P$, $(rep'[2], \theta') \in P$ where δ' (resp. θ') is a base C of $S^s_{\overline{x_j}}$ (resp. the first base of the terminal in $S^{\overline{1}}_{x_j}$). Since, by hypothesis, the three terminals of $\{S^{\overline{1}}_{x_i}, S^{\overline{1}}_{x_j}, S^{\overline{1}}_{x_k}\}$ are active, then θ and θ' are matched. Therefore, since both $\{\theta, \theta'\}$ are matched, *rep*[2] and *rep*$'$[2] are deleted. Since either *rep* or *rep*$'$ is active, either *rep*[1] or *rep*$'$[1] is matched. Thus, either δ or δ' is deleted.

Moreover, by construction, there is an arc between a base C of $S^s_{x_i}$ (resp. $S^s_{x_j}$) and the first base of the terminal in $S^1_{x_i}$ (resp. $S^1_{x_j}$). Therefore, since two terminals of $\{S^1_{x_i}, S^1_{x_j}, S^1_{x_k}\}$ are active, at least one base C of either $S^s_{x_i}$ or $S^s_{x_j}$ is deleted. Thus, a base of either both $S^s_{x_i}$ and $S^s_{\overline{x_i}}$ or both $S^s_{x_j}$ and $S^s_{\overline{x_j}}$ is deleted. Consequently, by Definition 1, the alignment is not canonical, a contradiction.

(b) Suppose now that the active terminals are in ($S^1_{x_i}$, $S^1_{x_k}$). By construction, there is a repeater *rep* of $S^1_{x_i}$ such that $(\delta, rep[1]) \in P$, $(rep[2], \theta) \in P$ where δ (resp. θ) is a base C of $S^s_{\overline{x_i}}$ (resp. the first base of the terminal in $S^{\overline{1}}_{x_i}$), as illustrated in Figure 5. Similarly, by construction, there is a repeater *rep*$'$ of $S^1_{x_k}$ such that $(\delta', rep'[1]) \in P$, $(rep'[2], \theta') \in P$ where δ' (resp. θ') is a base C of $S^s_{x_k}$ (resp. the first base of the terminal in $S^{\overline{1}}_{x_k}$). Since, by hypothesis, the three terminals of $\{S^{\overline{1}}_{x_i}, S^{\overline{1}}_{x_j}, S^{\overline{1}}_{x_k}\}$ are active, then θ and θ' are matched. Therefore, since both $\{\theta, \theta'\}$ are matched, *rep*[2] and *rep*$'$[2] are deleted. Since either *rep* or *rep*$'$ is active, either *rep*[1] or *rep*$'$[1] is matched. Thus, either δ or δ' is deleted. Moreover, by construction, there is an arc between a base C of $S^s_{x_i}$ (resp. $S^s_{\overline{x_k}}$) and the first base of the terminal in $S^1_{x_i}$ (resp. $S^1_{x_k}$). Therefore, since two terminals of $\{S^1_{x_i}, S^1_{x_j}, S^1_{x_k}\}$ are active, at least one base C of either $S^s_{x_i}$ or $S^s_{\overline{x_k}}$ is deleted. Thus, a base of either both $S^s_{x_i}$ and $S^s_{\overline{x_i}}$ or both $S^s_{x_k}$ and $S^s_{\overline{x_k}}$ is deleted. Consequently, by Definition 1, the alignment is not canonical, a contradiction.

(3) Suppose that all the repeaters but two of S^1 and three terminals of $\{S^1_{x_i}, S^1_{x_j}, S^1_{x_k}\}$ are active. By construction, there is a repeater *rep* such that $(\delta, rep[1]) \in P$, $(rep[2], \theta) \in P$ where δ (resp. θ) is a base C of $S^s_{\overline{x_i}}$ (resp. the first base of the terminal in $S^{\overline{1}}_{x_i}$). Similarly, by construction, there is a repeater *rep*$'$ such that $(\delta', rep'[1]) \in P$, $(rep'[2], \theta') \in P$ where δ' (resp. θ') is a base C of $S^s_{\overline{x_j}}$ (resp. the first base of the terminal in $S^{\overline{1}}_{x_j}$). By construction, there is a repeater *rep*$''$ such that $(\delta'', rep''[1]) \in P$, $(rep''[2], \theta'') \in P$ where δ'' (resp. θ'') is a base C of $S^s_{x_k}$ (resp. the first base of the terminal in $S^{\overline{1}}_{x_k}$). Since, by hypothesis, the three terminals of $\{S^{\overline{1}}_{x_i}, S^{\overline{1}}_{x_j}, S^{\overline{1}}_{x_k}\}$ are active, then θ, θ' and θ'' are matched. Therefore, since both $\{\theta, \theta', \theta''\}$ are matched, *rep*[2], *rep*$'$[2] and *rep*$''$[2] are deleted. Since either *rep*, *rep*$'$ or *rep*$''$ is active, either *rep*[1], *rep*$'$[1] or *rep*$''$[1] is matched. Thus, either δ, δ' or δ'' is deleted. Moreover, by construction, there is an arc between a base C of $S^s_{x_i}$ (resp. $S^s_{x_j}$ and $S^s_{\overline{x_k}}$) and the first base of the terminal in $S^1_{x_i}$ (resp. $S^1_{x_j}$ and $S^1_{x_k}$). Therefore, since three terminals of $\{S^1_{x_i}, S^1_{x_j}, S^1_{x_k}\}$ are

active, at least one base C of either $S^s_{x_i}$, $S^s_{x_j}$ or $S^s_{\overline{x_k}}$ is deleted. Thus, a base of either both $S^s_{x_i}$ and $S^s_{\overline{x_i}}$ or both $S^s_{x_j}$ and $S^s_{\overline{x_j}}$ or both $S^s_{x_k}$ and $S^s_{\overline{x_k}}$ is deleted. Therefore, by Definition 1, the alignment is not canonical, a contradiction.

Thus, the hypothesis that one repeater of $S^{\overline{1}}$ is inactive is wrong. Consequently, only a terminal of $S^{\overline{1}}$ can be inactive. We deduce that all the repeaters and two terminals of $S^{\overline{1}}$ are active. □

We now turn to proving that our construction is a polynomial time reduction from 3-SAT to APS$(\{<, \between\}, \emptyset)$.

Lemma 6. *Let I be an instance of the problem* 3-SAT *with n variables and q clauses, and I' an instance $((S, P); (T, Q))$ of* APS$(\{<, \between\}, \emptyset)$ *obtained by an* APS2-CP-*construction from I. An assignment of the variables that satisfies the boolean formula of I exists iff (T, Q) is an Arc-Preserving Subsequence of (S, P).*

Proof. (⇒) Suppose we have an assignment AS of the n variables that satisfies the boolean formula of I. By definition, for each clause there is at least one literal that satisfies it. Let (S, P) and (T, Q) be two sequences obtained from an APS2-CP-construction from I. We look for a set of bases to delete from S in order to obtain T. We define this set in three steps as follows.

(Step 1) For each variable $x_m \in AS$, $1 \leq m \leq n$:

– if $x_m = True$ then $S^e_{x_m}[2]$ and all the bases of $S^s_{x_m}$ are deleted,
– if $x_m = False$ then $S^e_{x_m}[1]$ and all the bases of $S^s_{\overline{x_m}}$ are deleted.

Notice that the sequence obtained from S_α (resp. S_β) by deleting the bases described above is similar to $T_{\alpha'}$ (resp. $T_{\beta'}$), when not considering arcs.

(Step 2) We recall that, for any $1 \leq m \leq n$ and any $1 \leq i \leq q$, γ^i_m (resp. $\gamma^i_{\overline{m}}$) denotes be the number of occurrences of literal x_m (resp. $\overline{x_m}$) in the set of clauses c_j with $i < j \leq q$ and $\lambda^i_m = \gamma^i_m + \gamma^i_{\overline{m}}$. For any $1 \leq m \leq n$ and for any $1 \leq i \leq q$, we also recall that $y^i_m = 1$ (resp. $y^i_{\overline{m}} = 1$) if $x_m \in c_i$ (resp. $\overline{x_m} \in c_i$), $y^i_m = 0$ (resp. $y^i_{\overline{m}} = 0$) otherwise. For each variable $x_m \in AS$, $1 \leq m \leq n$ and $1 \leq i \leq q$:

– if $x_m = True$ then the following bases are deleted:
  - $rep(i, m, j)[2]$ for all $1 \leq j \leq \lambda^i_m + y^i_{\overline{m}}$,
  - $rep(i, m, j)[1]$ for all $\lambda^i_m + y^i_{\overline{m}} < j \leq 2\lambda^i_m + y^i_{\overline{m}} + y^i_m$,
  - $rep(\overline{i}, m, j)[2]$ for all $1 \leq j \leq \lambda^i_m$,
  - $rep(\overline{i}, m, j)[1]$ for all $\lambda^i_m < j \leq 2\lambda^i_m$

– if $x_m = False$ then the following bases are deleted:
  - $rep(i, m, j)[1]$ with $1 \leq j \leq \lambda^i_m + y^i_{\overline{m}}$,
  - $rep(i, m, j)[2]$ with $\lambda^i_m + y^i_{\overline{m}} < j \leq 2\lambda^i_m + y^i_{\overline{m}} + y^i_m$,
  - $rep(\overline{i}, m, j)[1]$ with $1 \leq j \leq \lambda^i_m$,
  - $rep(\overline{i}, m, j)[2]$ with $\lambda^i_m < j \leq 2\lambda^i_m$

Let $j_i \in \{1,2,3\}$ denote the smallest position of the literal(s) satisfying c_i. For each $1 \leq i \leq q$, all the bases of the j_i^{th} terminal of $S^{\overline{i}}$ are deleted.

Notice that, for all $1 \leq m \leq n$ and all $1 \leq i \leq q$, a base G (resp. C) of each repeater of $S^i_{x_m}$ (resp. $S^{\overline{i}}_{x_m}$) is deleted. The sequence obtained from $S^{\overline{i}}$ by deleting the bases described in Step 2 is a sequence of $2+2\sum_{m=1}^{n}\lambda_m^i$ substrings CA (since, by construction, $S^{\overline{i}}$ is initially composed of $2\sum_{m=1}^{n}\lambda_m^i$ repeaters and 3 terminals).

By definition, $\sum_{m=1}^{n}\lambda_m^i$ represents the number of literals in all the clauses c_j with $i < j \leq q$. Since any clause is composed of three literals, we can deduce that $\sum_{m=1}^{n}\lambda_m^i = 3(q-i)$. Therefore, there are $2+2\sum_{m=1}^{n}\lambda_m^i$ (*i.e.* $2+6q-6i$) terminals (*i.e.* CA) in $T^{\overline{i}}$. Consequently, the sequence obtained from $S^{\overline{i}}$ by deleting the bases described in Step 2 is similar to $T^{\overline{i}}$ (when not considering arcs).

(Step 3) For each clause $c_i \in \mathcal{C}_q$ with $1 \leq i \leq q$, the following bases are deleted:

- if exactly one literal (i.e. the j_i^{th}) satisfies c_i then all the bases of the k^{th} and the l^{th} terminals of S^i with $k \neq l$ and $k,l \in \{1,2,3\}\backslash\{j_i\}$.
- if exactly two literals (say the j_i^{th} and k^{th}) satisfy c_i then:
  - all the bases of the l^{th} terminal of S^i with $l \neq k$, $l \neq j_i$ and $l \in \{1,2,3\}$,
  - all the bases of the repeater of S^i connected to the bases of the k^{th} terminal of $S^{\overline{i}}$.
- if exactly three literals (i.e. the j_i^{th}, k^{th} and l^{th}) satisfy c_i then:
  - all the bases of the repeater of S^i connected to the bases of the k^{th} terminal of $S^{\overline{i}}$
  - all the bases of the repeater of S^i connected to the bases of the l^{th} terminal of $S^{\overline{i}}$.

The sequence obtained from S^i by deleting the bases described in Step 2 is composed of a sequence of $6+2\sum_{m=1}^{n}\lambda_m^i$ substrings GA (since, by construction, S^i is initially composed of $3+2\sum_{m=1}^{n}\lambda_m^i$ repeaters and 3 terminals). Moreover, we know that $\sum_{m=1}^{n}\lambda_m^i = 3(q-i)$. Therefore, there are $4+2\sum_{m=1}^{n}\lambda_m^i$ (*i.e.* $4+6q-6i$) terminals (*i.e.* substrings GA) in T^i. As in each of the above cases, all the bases of two elements of S^i have been deleted, the sequence obtained from S^i by deleting the bases described in Step 2 and Step 3 is similar to T^i (when not considering arcs).

Thus, the sequence obtained from S by deleting the bases described in Step 1, Step 2 and Step 3 is similar to T (when not considering arcs). We now turn to demonstrating that at least one base of any arc of P has been deleted. In the following, we will distinguish arcs between bases A and U, denoted by AU-arcs, from arcs between bases C and G, denoted by CG-arcs. Let us consider those two types of arcs separately:

(1) By construction, for all $1 \leq m \leq n$, the following AU-arcs have been created: ($S^e_{x_m}[1]$,$S^s_{x_m}[1]$) and ($S^e_{x_m}[2]$,$S^s_{\overline{x_m}}[k_m+1]$).

By Step 1, since a variable x_m has a unique value, either each base of $S^s_{\overline{x_m}}$ and $S^e_{x_m}[1]$, or each base of $S^s_{x_m}$ and $S^e_{x_m}[2]$ is deleted for all $1 \leq m \leq n$. Thus, at least one base in S of any AU-arc of P is deleted.

(2) By construction, the following CG-arcs have been created:

- for all $1 \leq m \leq n$, $1 \leq j \leq 2\lambda^i_m$ and $1 \leq i < q$:
  - an arc between the second base G of $rep(i, m, j)$ and the first base C of the j^{th} element (i.e. either a terminal or a repeater) of $S^{\overline{i}}_{x_m}$;
  - an arc between the second base C of $rep(\overline{i}, m, j)$ and the first base G of the j^{th} element of $S^{i+1}_{x_m}$.
- for all $1 \leq j \leq \gamma_m + \gamma_{\overline{m}}$, an arc between the j^{th} base C of substring $S^s_{x_m} A S^s_{\overline{x_m}}$ in S_β and the first base G of the j^{th} element of $S^1_{x_m}$ in S_ζ.

In the following, we focus on the arcs of a clause c_i and the arcs between c_i and c_{i+1}, for any given $1 \leq i < q$ (cf. Figure 6). More precisely, we will demonstrate that, for any given $1 \leq m \leq n$, at least one base of any arc in $\{S^i_m, S^{\overline{i}}_m, S^{i+1}_m, S^{\overline{i+1}}_m\}$ is deleted. This will prove that at least one base of any arc connecting two bases of S_ζ is deleted. In a second step, we will focus on the first clause and prove that at least one base of any arc connecting a base of S_β and a base of S^1 is deleted.

We recall that by construction:

$$\begin{aligned} S^i_{x_m} &= (GGA)^{\lambda^i_m + y^i_{\overline{m}}}(GA)^{y^i_m}(GGA)^{\lambda^i_m + y^i_m}(GA)^{y^i_{\overline{m}}} \\ S^{\overline{i}}_{x_m} &= (CCA)^{\lambda^i_m}(CA)^{y^i_{\overline{m}}}(CCA)^{\lambda^i_m}(CA)^{y^i_m} \end{aligned}$$

Consider any variable x_m with $1 \leq m \leq n$. For any given $1 \leq m \leq n$ and $1 \leq i \leq q$, we define the following four subsets of arcs:

- (A^i_m) for each $1 \leq m \leq n$, the $\lambda^i_m + y^i_{\overline{m}}$ first arcs between a base of $S^i_{x_m}$ and a base of $S^{\overline{i}}_{x_m}$;
- (B^i_m) for each $1 \leq m \leq n$, the rest of the arcs between a base of $S^i_{x_m}$ and a base of $S^{\overline{i}}_{x_m}$;
- (C^i_m) for each $1 \leq m \leq n$, the λ^i_m first arcs between a base of $S^{\overline{i}}_{x_m}$ and a base of $S^{i+1}_{x_m}$;
- (D^i_m) for each $1 \leq m \leq n$, the rest of the arcs between a base of $S^{\overline{i}}_{x_m}$ and a base of $S^{i+1}_{x_m}$.

Suppose first that $x_m = True$. We now consider separately the nine following cases:

- (a_1) $x_m, x_{\overline{m}} \notin \{c_i, c_{i+1}\}$;
- (a_2) $x_m, x_{\overline{m}} \notin c_i$ and $x_m \in c_{i+1}$;
- (a_3) $x_m, x_{\overline{m}} \notin c_i$ and $x_{\overline{m}} \in c_{i+1}$;
- (b_1) $x_m \in c_i$ and $x_m, x_{\overline{m}} \notin c_{i+1}$;

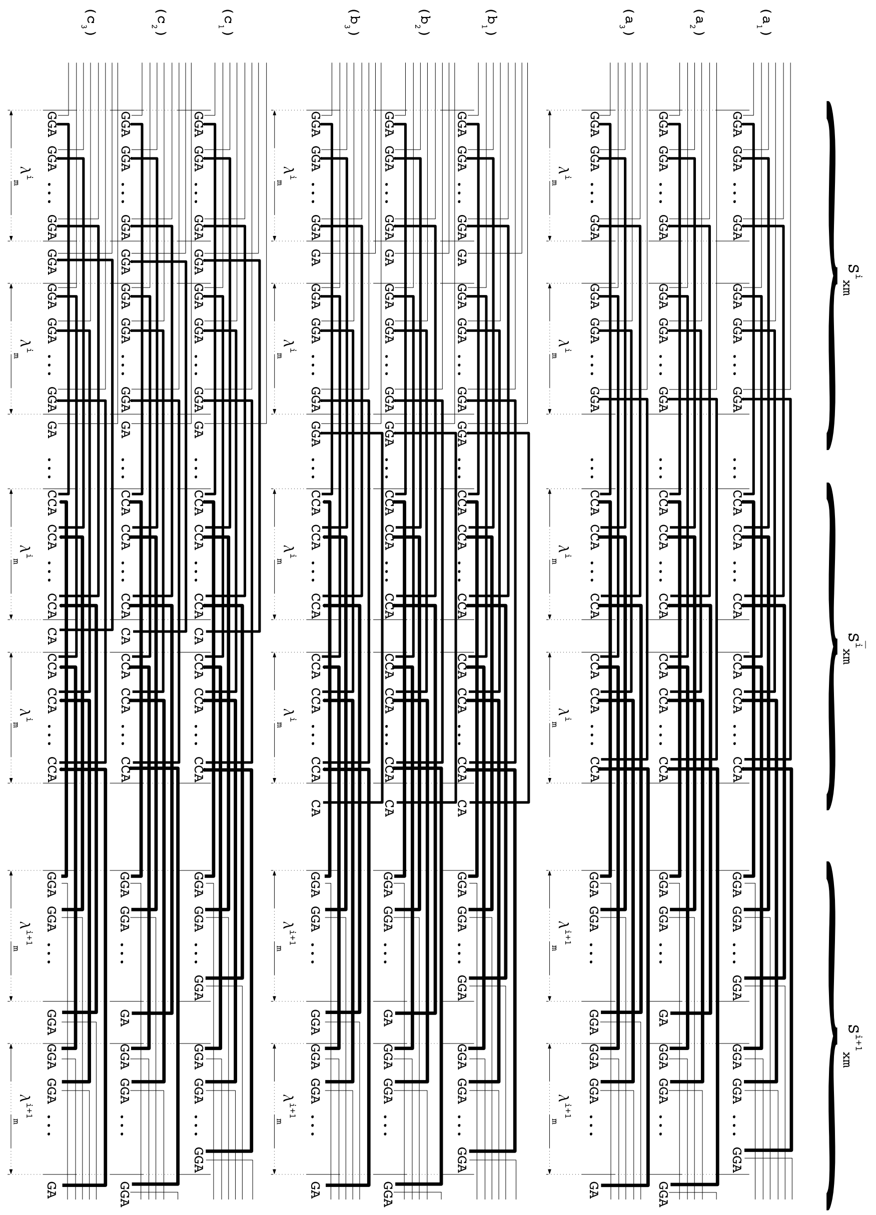


Fig. 6. Sketch of the arc-structure of a clause c_i, for any given $1 \leq m \leq n$ and $1 \leq i < q$. (a_1) when $x_m, x_{\overline{m}} \notin \{c_i, c_{i+1}\}$. (a_2) when $x_m, x_{\overline{m}} \notin c_i$ and $x_m \in c_{i+1}$. (a_3) when $x_m, x_{\overline{m}} \notin c_i$ and $x_{\overline{m}} \in c_{i+1}$. (b_1) when $x_m \in c_i$ and $x_m, x_{\overline{m}} \notin c_{i+1}$. (b_2) when $x_m \in c_i$ and $x_m \in c_{i+1}$. (b_3) when $x_m \in c_i$ and $x_{\overline{m}} \in c_{i+1}$. (c_1) when $x_{\overline{m}} \in c_i$ and $x_m, x_{\overline{m}} \notin c_{i+1}$. (c_2) when $x_{\overline{m}} \in c_i$ and $x_m \in c_{i+1}$. (c_3) when $x_{\overline{m}} \in c_i$ and $x_{\overline{m}} \in c_{i+1}$.

- (b_2) $x_m \in c_i$ and $x_m \in c_{i+1}$;
- (b_3) $x_m \in c_i$ and $x_{\overline{m}} \in c_{i+1}$;
- (c_1) $x_{\overline{m}} \in c_i$ and $x_m, x_{\overline{m}} \notin c_{i+1}$;
- (c_2) $x_{\overline{m}} \in c_i$ and $x_m \in c_{i+1}$;
- (c_3) $x_{\overline{m}} \in c_i$ and $x_{\overline{m}} \in c_{i+1}$.

(a_1). Since $x_m, x_{\overline{m}} \notin \{c_i, c_{i+1}\}$, by definition, $y^i_m = y^i_{\overline{m}} = y^{i+1}_m = y^{i+1}_{\overline{m}} = 0$. Moreover, as $x_m = True$, for all $1 \leq i \leq q$ and all $1 \leq j \leq \lambda^i_m + y^i_{\overline{m}}$, $rep(i, m, j)[2]$ is deleted. Thus, at least one base of any arc of the set (A^i_m) is deleted.

Since $x_m = True$, $rep(\overline{i}, m, j)[1]$ is deleted for all $1 \leq i \leq q$ and all $\lambda^i_m < j \leq 2\lambda^i_m$ (cf. Step 2). Therefore, at least one base of any arc of the set (B^i_m) is deleted.

Moreover, as $x_m = True$, for all $1 \leq i \leq q$ and all $1 \leq j \leq \lambda^i_m$, $rep(\overline{i}, m, j)[2]$ is deleted (cf. Step 2). Consequently, at least one base of any arc of the set (C^i_m) is deleted.

Finally, $x_m = True$ implies that $rep(i+1, m, j)[1]$ is deleted for all $1 \leq i < q$ and all $\lambda^{i+1}_m + y^{i+1}_{\overline{m}} < j \leq 2\lambda^{i+1}_m + y^{i+1}_{\overline{m}} + y^{i+1}_m$. Therefore, at least one base of any arc of the set (D^i_m) is deleted.

(a_2). The proof is fully similar to the one of (a_1).

(a_3). Since $x_m, x_{\overline{m}} \notin c_i$ and $x_{\overline{m}} \in c_{i+1}$, by definition, $y^i_m = y^i_{\overline{m}} = y^{i+1}_m = 0$ and $y^{i+1}_{\overline{m}} = 1$. Moreover, as $x_m = True$, for all $1 \leq i \leq q$ and all $1 \leq j \leq \lambda^i_m + y^i_{\overline{m}}$, $rep(i, m, j)[2]$ is deleted. Thus, at least one base of any arc of the set (A^i_m) is deleted.

Since $x_m = True$, $rep(\overline{i}, m, j)[1]$ is deleted for all $1 \leq i \leq q$ and all $\lambda^i_m < j \leq 2\lambda^i_m$ (cf. Step 2). Therefore, at least one base of any arc of the set (B^i_m) is deleted.

Moreover, as $x_m = True$, for all $1 \leq i \leq q$ and all $1 \leq j \leq \lambda^i_m$, $rep(\overline{i}, m, j)[2]$ is deleted (cf. Step 2). Consequently, at least one base of any arc of the set (C^i_m) is deleted.

Finally, $x_m = True$ implies that $rep(i+1, m, j)[1]$ is deleted for all $1 \leq i < q$ and all $\lambda^{i+1}_m + y^{i+1}_{\overline{m}} < j \leq 2\lambda^{i+1}_m + y^{i+1}_{\overline{m}} + y^{i+1}_m$. Moreover, by construction, if $y^{i+1}_{\overline{m}} = 1$ then there is an arc connecting the base $rep(\overline{i}, m, j)[2]$ to a base of the j^{th} element (which is a terminal) of $S^{i+1}_{x_m}$ where $j = 2\lambda^i_m$. By definition, as $\overline{x_m} \in c_{i+1}$, $\overline{x_m}$ does not satisfies c_{i+1} (since $x_m = True$). By definition, there exists at least a literal which, by it assignment, satisfies c_{i+1}. Therefore, all the bases of the terminal of $S^{i+1}_{\overline{x_m}}$ have been deleted (cf. Step 3). Therefore, at least one base of any arc of the set (D^i_m) is deleted.

(b_1). Since $x_m \in c_i$ and $x_m, x_{\overline{m}} \notin c_{i+1}$, by definition, $y^i_m = 1$ and $y^i_{\overline{m}} = y^{i+1}_m = y^{i+1}_{\overline{m}} = 0$. Moreover, as $x_m = True$, for all $1 \leq i \leq q$ and all $1 \leq j \leq \lambda^i_m + y^i_{\overline{m}}$, $rep(i, m, j)[2]$ is deleted. Thus, at least one base of any arc of the set (A^i_m) is deleted.

Since $x_m = True$, $rep(\overline{i}, m, j)[1]$ is deleted for all $1 \leq i \leq q$ and all $\lambda^i_m < j \leq 2\lambda^i_m$ (cf. Step 2). Moreover, by construction, if $y^i_m = 1$ then there is an arc connecting the base $rep(i,m,j)[2]$ to a base of the j^{th} element (which is a terminal) of $S^{\bar{i}}_{x_m}$ where $j = 2\lambda^i_m + y^i_{\overline{m}} + y^i_m$. By definition, since $y^i_m = 1$, $x_m \in c_i$ and thus x_m satisfies c_i. If x_m is the literal with the smallest position of the literal(s) satisfying c_i, then all the bases of the terminal of $S^{\bar{i}}_{x_m}$ have been deleted. Otherwise, all the bases of the repeater of $S^i_{x_m}$ connected to the bases of the terminal of $S^{\bar{i}}_{x_m}$ are deleted (cf. Step 3). Therefore, at least one base of any arc of the set (B^i_m) is deleted.

Moreover, as $x_m = True$, for all $1 \leq i \leq q$ and all $1 \leq j \leq \lambda^i_m$, $rep(\bar{i},m,j)[2]$ is deleted (cf. Step 2). Consequently, at least one base of any arc of the set (C^i_m) is deleted.

Finally, $x_m = True$ implies that $rep(i+1,m,j)[1]$ is deleted for all $1 \leq i < q$ and all $\lambda^{i+1}_m + y^{i+1}_{\overline{m}} < j \leq 2\lambda^{i+1}_m + y^{i+1}_{\overline{m}} + y^{i+1}_m$. Therefore, at least one base of any arc of the set (D^i_m) is deleted.

(b_2). The proof is fully similar to the one of (b_1).

(b_3). Since $x_m \in c_i$ and $x_{\overline{m}} \in c_{i+1}$, by definition, $y^i_m = y^{i+1}_{\overline{m}} = 1$ and $y^{i+1}_m = y^i_{\overline{m}} = 0$. Moreover, as $x_m = True$, for all $1 \leq i \leq q$ and all $1 \leq j \leq \lambda^i_m + y^i_{\overline{m}}$, $rep(i,m,j)[2]$ is deleted. Thus, at least one base of any arc of the set (A^i_m) is deleted.

Since $x_m = True$, $rep(\bar{i},m,j)[1]$ is deleted for all $1 \leq i \leq q$ and all $\lambda^i_m < j \leq 2\lambda^i_m$ (cf. Step 2). Moreover, by construction, if $y^i_m = 1$ then there is an arc connecting the base $rep(i,m,j)[2]$ to a base of the j^{th} element (which is a terminal) of $S^{\bar{i}}_{x_m}$ where $j = 2\lambda^i_m + y^i_{\overline{m}} + y^i_m$. By definition, since $y^i_m = 1$, $x_m \in c_i$ and thus x_m satisfies c_i. If x_m is the literal with the smallest position of the literal(s) satisfying c_i then all the bases of the terminal of $S^{\bar{i}}_{x_m}$ have been deleted. Otherwise, all the bases of the repeater of $S^i_{x_m}$ connected to the bases of the terminal of $S^{\bar{i}}_{x_m}$ are deleted (cf. Step 3). Therefore, at least a base of any arc of the set (B^i_m) is deleted.

Moreover, as $x_m = True$, for all $1 \leq i \leq q$ and all $1 \leq j \leq \lambda^i_m$, $rep(\bar{i},m,j)[2]$ is deleted (cf. Step 2). Consequently, at least one base of any arc of the set (C^i_m) is deleted.

Finally, $x_m = True$ implies that $rep(i+1,m,j)[1]$ is deleted for all $1 \leq i < q$ and all $\lambda^{i+1}_m + y^{i+1}_{\overline{m}} < j \leq 2\lambda^{i+1}_m + y^{i+1}_{\overline{m}} + y^{i+1}_m$. Moreover, by construction, if $y^{i+1}_{\overline{m}} = 1$ then there is an arc connecting the base $rep(\bar{i},m,j)[2]$ to a base of the j^{th} element (which is a terminal) of $S^{i+1}_{x_m}$ where $j = 2\lambda^i_m$. By definition, as $\overline{x_m} \in c_{i+1}$, $\overline{x_m}$ does not satisfies c_{i+1} (since $x_m = True$). By definition, there exists at least a literal which, by it assignment, satisfies c_{i+1}. Therefore, all the bases of the terminal of $S^{i+1}_{\overline{x_m}}$ have been deleted (cf. Step 3). Therefore, at least one base of any arc of the set (D^i_m) is deleted.

(c_1). Since $x_{\overline{m}} \in c_i$ and $x_m, x_{\overline{m}} \notin c_{i+1}$, by definition, $y^i_{\overline{m}} = 1$ and $y^i_m = y^{i+1}_m = y^{i+1}_{\overline{m}} = 0$. Moreover, as $x_m = True$, for all $1 \leq i \leq q$ and all $1 \leq j \leq \lambda^i_m + y^i_{\overline{m}}$, $rep(i,m,j)[2]$ is deleted. Thus, at least one base of any arc of the set (A^i_m) is deleted.

Since $x_m = True$, $rep(\overline{i}, m, j)[1]$ is deleted for all $1 \leq i \leq q$ and all $\lambda_m^i < j \leq 2\lambda_m^i$ (cf. Step 2). Therefore, at least a base of any arc of the set (B_m^i) is deleted.

Moreover, as $x_m = True$, for all $1 \leq i \leq q$ and all $1 \leq j \leq \lambda_m^i$, $rep(\overline{i}, m, j)[2]$ is deleted (cf. Step 2). Consequently, at least one base of any arc of the set (C_m^i) is deleted.

Finally, $x_m = True$ implies that $rep(i+1, m, j)[1]$ is deleted for all $1 \leq i < q$ and all $\lambda_m^{i+1} + y_{\overline{m}}^{i+1} < j \leq 2\lambda_m^{i+1} + y_{\overline{m}}^{i+1} + y_m^{i+1}$. Moreover, by construction, if $y_{\overline{m}}^{i+1} = 1$ then there is an arc connecting the base $rep(\overline{i}, m, j)[2]$ to a base of the j^{th} element (which is a terminal) of $S_{x_m}^{i+1}$ where $j = 2\lambda_m^i$. By definition, as $\overline{x_m} \in c_{i+1}$, $\overline{x_m}$ does not satisfies c_{i+1} (since $x_m = True$). By definition, there exists at least a literal which, by it assignment, satisfies c_{i+1}. Therefore, all the bases of the terminal of $S_{\overline{x_m}}^{i+1}$ have been deleted (cf. Step 3). Therefore, at least one base of any arc of the set (D_m^i) is deleted.

(c_2). The proof is fully similar to the one of (c_1).

(c_3). The proof is fully similar to the one of (a_1).

Therefore, when $x_m = True$, at least one base of any CG-arc has been deleted. If $x_m = False$ then a similar reasoning leads to the same conclusion, *i.e.* at least one base of any CG-arc has been deleted. Thus, for any $1 < i \leq q$, any CG-arc between a base of an element of the representation of the clause c_{i-1} (*i.e.* $S^{i-1}\ U\ S^{\overline{i-1}}$) and a base of an element of the representation of the clause c_i (*i.e.* $S^i\ U\ S^{\overline{i}}$) has been deleted.

Moreover, for any $1 \leq i \leq q$, any CG-arc between two bases of the representation of the clause c_i has been deleted. Remains us to consider the special case of the first clause (*i.e.* c_1). Indeed, there is, for all $1 \leq j \leq \gamma_m + \gamma_{\overline{m}}$, an arc between the j^{th} base C of substring $S_{x_m}^s A S_{\overline{x_m}}^s$ in S_β and the first base G of the j^{th} element of $S_{x_m}^1$ in S_ζ.

For each $1 \leq m \leq n$, if $x_m = True$ then each base of $S_{x_m}^s$ and $S_{x_m}^e[2]$ is deleted and $rep(1, m, j)[2]$ is deleted with $1 \leq j \leq \lambda_m^i + y_{\overline{m}}^i$. Moreover, for each $1 \leq m \leq n$, if $x_m = False$ then each base of $S_{\overline{x_m}}^s$ and $S_{x_m}^e[1]$ is deleted and $rep(1, m, j)[1]$ is deleted with $1 \leq j \leq \lambda_m^i + y_{\overline{m}}^i$. Thus, at least one base in S of any CG-arc of P is deleted.

We just proved that if S' is the sequence obtained from S by deleting all the bases described in Step 1, Step 2 and Step 3 together with their incident arcs, then there is no arc in S' (i.e. neither AU-arcs or CG-arcs). Moreover, we demonstrated previously that the sequence S' is similar to T. Therefore, if an assignment of the variables that satisfies the boolean formula of I exists, then (T, Q) is an Arc-Preserving Subsequence of (S, P).

$(\Leftarrow)$ Let I be an instance of the problem 3-Sat with n variables and q clauses. Let I' be an instance $((S, P); (T, Q))$ of APS$(\{<, \between\}, \emptyset)$ obtained by an APS2-cp-construction from I such that (T, Q) can be obtained from (S, P) by deleting some of its bases together with their incident arcs, if any. By Lemma 3, any corresponding alignment of (S, P) and (T, Q) is canonical. Therefore, $T^s_{x_m}$ is matched with either $S^s_{x_m} A$ or $A\ S^s_{\overline{x_m}}$. Consequently, for any $1 \leq m \leq n$, we define an assignment AS of the variables of I as follows:

- if $T^s_{x_m}$ is matched with $S^s_{x_m} A$ then $x_m = False$,
- otherwise, $x_m = True$.

Now, let us prove that for any $1 \leq i \leq q$ the clause c_i is satisfied by AS. Let us first focus on the first clause (*i.e.* c_1). c_1 is defined by three literals (say x_i, x_j and x_k). Since, c_1 is equal to the disjunction of variables built with x_i, x_j and x_k, c_1 can have eight different forms, because each literal can appear in either its positive (x_i) or negative ($\overline{x_i}$) form. In the following, we suppose, to illustrate the proof, that $c_1 = (x_i \vee x_j \vee \overline{x_k})$ as illustrated in Figure 5, since the other cases can be treated similarly.

By Lemmas 4 and 5, the two following properties must be satisfied:

- all the repeaters and two terminals of $S^{\overline{1}}$ are active,
- and either:
  - all the repeaters and one terminal of S^1 are active,
  - all the repeaters but one and two terminals of S^1 are active,
  - all the repeaters but two and three terminals of S^1 are active.

(1) Suppose that all the repeaters of S^1 and one terminal of $\{S^1_{x_i}, S^1_{x_j}, S^1_{x_k}\}$ are active. The active terminal can be in either $S^1_{x_i}$, $S^1_{x_j}$ or $S^1_{x_k}$. Since the cases where the active terminal is either in $S^1_{x_i}$ or $S^1_{x_j}$ are fully similar, we detail hereafter only two cases: (a) the active terminal is in $S^1_{x_i}$ and (b) the active terminal is in $S^1_{x_k}$.

(a) Suppose that the active terminal is in $S^1_{x_i}$. By construction, there is an arc between a base C of $S^s_{x_i}$ and the first base of the terminal in $S^1_{x_i}$. Thus, a base C of $S^s_{x_i}$ is deleted. Therefore, by the way we defined AS, $x_i = True$ and thus c_1 is satisfied.

(b) Suppose that the active terminal is in $S^1_{x_k}$. By construction, there is an arc between a base C of $S^s_{\overline{x_k}}$ and the first base of the terminal in $S^1_{x_k}$. Thus, a base C of $S^s_{\overline{x_k}}$ is deleted. Therefore, by the way we defined AS, $x_k = False$ and thus c_1 is satisfied.

(2) Suppose that all the repeaters but one of S^1 and two terminals of $\{S^1_{x_i}$, $S^1_{x_j}$, $S^1_{x_k}\}$ are active. The active terminals can be in either ($S^1_{x_i}$, $S^1_{x_j}$), ($S^1_{x_i}$, $S^1_{x_k}$) or ($S^1_{x_j}$, $S^1_{x_k}$). Since the cases where the active terminals are either in ($S^1_{x_i}$, $S^1_{x_k}$) or ($S^1_{x_j}$, $S^1_{x_k}$) are fully similar, we detail hereafter only two cases: (a) the active terminals are in ($S^1_{x_i}$, $S^1_{x_j}$) and (b) the active terminals are in ($S^1_{x_i}$, $S^1_{x_k}$).

(a) Suppose that the active terminals are in ($S^1_{x_i}$, $S^1_{x_j}$). By construction, there is an arc between a base C of $S^s_{x_i}$ and the first base of the terminal in $S^1_{x_i}$. Thus, a base C of $S^s_{x_i}$ is deleted. Moreover, by construction, there is an arc between a base C of $S^s_{x_j}$ and the first base of the terminal in $S^1_{x_j}$. Thus, a base C of $S^s_{x_j}$

is deleted. Therefore, by the way we defined AS, $x_i = x_j = True$ and thus c_1 is satisfied.

For the sake of the proof, we now detail the alignment of the elements of c_1 in case (a). Since all the repeaters and two terminals of $S^{\overline{1}}$ are active, at least a terminal of either $S^{\overline{1}}_{x_i}$ or $S^{\overline{1}}_{x_j}$ is active. By construction, there is a repeater rep of $S^1_{x_i}$ such that $(\delta, rep[1]) \in P$ and $(rep[2], \theta) \in P$, where δ (resp. θ) is a base C of $S^s_{\overline{x_i}}$ (resp. the first base of the terminal in $S^{\overline{1}}_{x_i}$), as illustrated in Figure 5. Moreover, by construction, there is a repeater rep' of $S^1_{x_j}$ such that $(\delta', rep'[1]) \in P$ and $(rep'[2], \theta') \in P$, where δ' (resp. θ') is a base C of $S^s_{\overline{x_j}}$ (resp. the first base of the terminal in $S^{\overline{1}}_{x_j}$), as illustrated in Figure 5. Since at least a terminal of either $S^{\overline{1}}_{x_i}$ or $S^{\overline{1}}_{x_j}$ is active, then either θ or θ' is matched. By definition, as $Q = \emptyset$, at least one base incident to every arc of P has to be deleted. Therefore, either $rep[2]$ or $rep'[2]$ is deleted. Since either rep or rep' is an active repeater, either $rep[1]$ or $rep'[1]$ is matched. Thus, either δ or δ' is deleted. Since the alignment is canonical, for all $1 \leq m \leq n$, a base of both $S^s_{x_m}$ and $S^s_{\overline{x_m}}$ cannot be deleted. Therefore, the only two solutions are: either the terminal of $S^{\overline{1}}_{x_i}$ and rep' are inactive, or the terminal of $S^{\overline{1}}_{x_j}$ and rep are inactive.

(b) Suppose that the active terminals are in ($S^1_{x_i}$, $S^1_{x_k}$). By construction, there is an arc between a base C of $S^s_{x_i}$ and the first base of the terminal in $S^1_{x_i}$. Thus, a base C of $S^s_{x_i}$ is deleted. Moreover, by construction, there is an arc between a base C of $S^s_{\overline{x_k}}$ and the first base of the terminal in $S^1_{x_k}$. Thus, a base C of $S^s_{\overline{x_k}}$ is deleted. Therefore, by the way we defined AS, $x_i = True$, $x_k = False$ and thus c_1 is satisfied.

For the sake of the proof, we now detail the alignment of the elements of c_1 in case (b). Since all the repeaters and two terminals of $S^{\overline{1}}$ are active, at least a terminal of either $S^{\overline{1}}_{x_i}$ or $S^{\overline{1}}_{x_k}$ is active. By construction, there is a repeater rep of $S^1_{x_i}$ such that $(\delta, rep[1]) \in P$ and $(rep[2], \theta) \in P$, where δ (resp. θ) is a base C of $S^s_{\overline{x_i}}$ (resp. the first base of the terminal in $S^{\overline{1}}_{x_i}$), as illustrated in Figure 5. Moreover, by construction, there is a repeater rep' of $S^1_{x_k}$ such that $(\delta', rep'[1]) \in P$ and $(rep'[2], \theta') \in P$, where δ' (resp. θ') is a base C of $S^s_{x_k}$ (resp. the first base of the terminal in $S^{\overline{1}}_{x_k}$), as illustrated in Figure 5. Since at least a terminal of either $S^{\overline{1}}_{x_i}$ or $S^{\overline{1}}_{x_k}$ is active, then either θ or θ' is matched. By definition, as $Q = \emptyset$, at least one base incident to every arc of P has to be deleted. Therefore, either $rep[2]$ or $rep'[2]$ is deleted. Since either rep or rep' is an active repeater, either $rep[1]$ or $rep'[1]$ is matched. Thus, either δ or δ' is deleted. Since the alignment is canonical, for all $1 \leq m \leq n$, a base of both $S^s_{x_m}$ and $S^s_{\overline{x_m}}$ cannot be deleted. Therefore, the only two solutions are: either the terminal of $S^{\overline{1}}_{x_i}$ and rep' are inactive, or the terminal of $S^{\overline{1}}_{x_k}$ and rep are inactive.

(3) Suppose that all the repeaters but two of S^1 and three terminals of $\{S^1_{x_i}, S^1_{x_j}, S^1_{x_k}\}$ are active. By construction, there is an arc between a base C of $S^s_{x_i}$ and the first base of the terminal in $S^1_{x_i}$. Thus, a base C of $S^s_{x_i}$ is deleted. Moreover, there is an arc between a base C of $S^s_{x_j}$ and the first base of the

terminal in $S^1_{x_j}$. Thus, a base C of $S^s_{x_j}$ is deleted. Finally, by construction, there is an arc between a base C of $S^s_{\overline{x_k}}$ and the first base of the terminal in $S^1_{x_k}$. Thus, a base C of $S^s_{\overline{x_k}}$ is deleted. Therefore, by the way we defined AS, $x_i = x_j = True$, $x_k = False$ and thus c_1 is satisfied.

For the sake of the proof, we now detail the alignment of the elements of c_1 in case (3). Since all the repeaters and two terminals of $S^{\overline{1}}$ are active, at least two terminals of $S^{\overline{1}}_{x_i}, S^{\overline{1}}_{x_j}, S^{\overline{1}}_{x_k}$ are active. By construction, there is a repeater rep of $S^1_{x_i}$ such that $(\delta, rep[1]) \in P$ and $(rep[2], \theta) \in P$, where δ (resp. θ) is a base C of $S^s_{\overline{x_i}}$ (resp. the first base of the terminal in $S^{\overline{1}}_{x_i}$), as illustrated in Figure 5. Moreover, by construction, there is a repeater rep' of $S^1_{x_j}$ such that $(\delta', rep'[1]) \in P$ and $(rep'[2], \theta') \in P$, where δ' (resp. θ') is a base C of $S^s_{\overline{x_j}}$ (resp. the first base of the terminal in $S^{\overline{1}}_{x_j}$), as illustrated in Figure 5. Finally, by construction, there is a repeater rep'' of $S^1_{x_k}$ such that $(\delta'', rep''[1]) \in P$ and $(rep''[2], \theta'') \in P$, where δ'' (resp. θ'') is a base C of $S^s_{x_k}$ (resp. the first base of the terminal in $S^{\overline{1}}_{x_k}$), as illustrated in Figure 5. Since at least two terminals of $S^{\overline{1}}_{x_i}, S^{\overline{1}}_{x_j}, S^{\overline{1}}_{x_k}$ are active, then at least two of $(\theta, \theta', \theta'')$ are matched. By definition, as $Q = \emptyset$, at least one base incident to every arc of P has to be deleted. Therefore, two of $(rep[2], rep'[2], rep''[2])$ are deleted. Since rep, rep' or rep'' is an active repeater, either $rep[1]$, $rep'[1]$ or $rep''[1]$ is matched. Thus, either δ, δ' or δ'' is deleted. Since the alignment is canonical, for all $1 \leq m \leq n$ a base of both $S^s_{x_m}$ and $S^s_{\overline{x_m}}$ cannot be deleted. Therefore, the only three solutions are: either the terminal of $S^{\overline{1}}_{x_i}$ and rep' and rep'' are inactive, or the terminal of $S^{\overline{1}}_{x_j}$ and rep and rep'' are inactive, or the terminal of $S^{\overline{1}}_{x_k}$ and rep and rep' are inactive.

We just proved that if I' is a solution then the truth assignment we defined above satisfies clause c_1. Moreover, we proved that any inactive repeater of S^1 is linked to a terminal of $S^{\overline{1}}$ (*i.e.* its second base is connected to a base of a terminal of $S^{\overline{1}}$). Let rep be a repeater in S such that $rep[1]$ and $rep[2]$ are respectively connected to bases u and v. The particular design of the repeaters ensues that if rep is active then the situation is equivalent to the one where u and v are connected with an arc. Indeed, if (S, P) is an arc-preserving subsequence of (T, Q) and rep is active, then exactly one out of $\{rep[1], rep[2]\}$ is matched. Therefore, if v is matched then $rep[2]$ is deleted and $rep[1]$ is matched. Consequently, u is deleted. Similarly, if u is matched then v is deleted. More generally we can prove the following claim (illustrated in Figure 7):

Claim. Let u and v be two bases and $\{rep_1, rep_2 \ldots rep_k\}$ be a set of repeaters such that $(u, rep_1[1]) \in P$, $(rep_k[2], v) \in P$ and $(rep_i[2], rep_{i+1}[1]) \in P$ for all $1 \leq i < k$.

Let $\mathcal{A}$ be an alignment. If for each $1 \leq i \leq k$, rep_i is active in $\mathcal{A}$, then:

- if u is matched then v is deleted;
- if v is matched then u is deleted.

Therefore, since all the repeaters of $S^{\overline{1}}$ are active and the inactive repeaters of S^1 are linked to terminals of $S^{\overline{1}}$, by the above claim, considering clause c_2

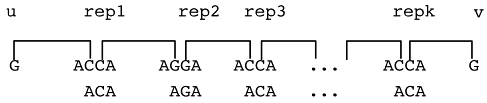


Fig. 7. Illustration of Claim 4.

is equivalent to considering c_1. Therefore, c_2 is satisfied and all the repeaters of $\overline{S^2}$ are active and the inactive repeaters of S^2 are linked to terminals of $\overline{S^2}$. Consequently, a similar reasoning can be done recursively for any clause c_i with $1 \leq i \leq q$. Thus, we just proved that if I' is a solution then the truth assignment we defined above satisfies all the clauses. □

5 Two polynomial time solvable APS problems

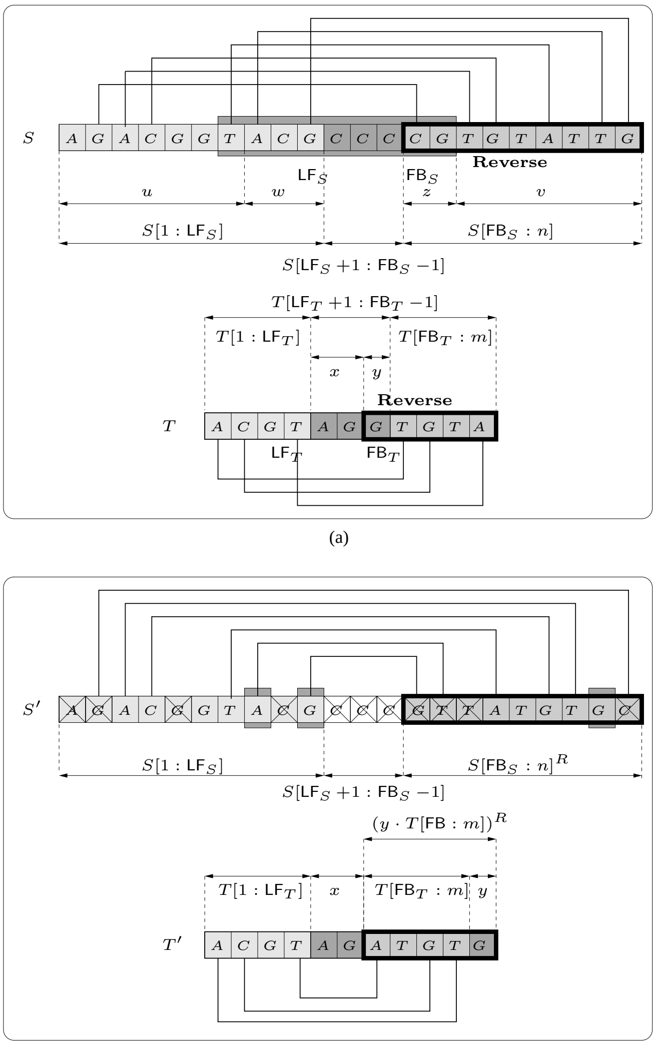
We prove in this section that APS($\{\between\}$,$\emptyset$) and APS($\{\between\}$,$\{\between\}$) are polynomial time solvable. In other words, the relation $\between$ alone does not imply **NP**-completeness.

We need the following notations. Sequences are the concatenation of zero or more elements from an alphabet. We use the period "." as the concatenation operator, but frequently the two operands are simply put side by side. Let $S = S[1]\,S[2]\ldots S[m]$ be a sequence of length m. For all $1 \leq i \leq j \leq m$, we write $S[i:j]$ to denote $S[i]\,S[i+1]\ldots S[j]$. The *reverse* of S is the sequence $S^R = S[m]\ldots S[2]\,S[1]$. A *factorization* of S is any decomposition $S = x_1\,x_2\ldots x_q$ where $x_1, x_2, \ldots x_q$ are (possibly empty) sequences. Let (S,P) be a $\{\between\}$-arc-annotated sequence and $(i,j) \in P$, $i < j$, be an arc. We call $S[i]$ a *forward base* and $S[j]$ a *backward base*. We will denote by LF_S the position of the last forward base in (S,P) and by FB_S the position of the first backward base in (S,P), *i.e.*, $\mathsf{LF}_S = \max\{i : (i,j) \in P\}$ and $\mathsf{FB}_S = \min\{j : (i,j) \in P\}$. By convention, we let $\mathsf{LF}_S = 0$ and $\mathsf{FB}_S = |S|+1$ if $P = \emptyset$. Observe that $\mathsf{LF}_S < \mathsf{FB}_S$.

We begin by proving a factorization result on $\{\between\}$-arc-annotated sequences.

Lemma 7. *Let S and T be two $\{\between\}$-arc-annotated sequences of length n and m, respectively. If T occurs as an arc preserving subsequence in S, then there exists a possibly trivial factorization $T[\mathsf{LF}_T+1 : \mathsf{FB}_T-1] = xy$ such that $T[1 : \mathsf{LF}_T] \cdot x \cdot (y \cdot T[\mathsf{FB}_T : m])^R$ occurs as an arc preserving subsequence in $S[1 : \mathsf{FB}_S-1] \cdot S[\mathsf{FB}_S : n]^R$.*

Proof. Suppose that T occurs as an arc preserving subsequence in S. Since both S and T are $\{\between\}$-arc-annotated sequences, then there exist two factorizations $S[1 : \mathsf{LF}_S] = uw$ and $S[\mathsf{FB}_S : n] = zv$ such that: (i) $T[1 : \mathsf{LF}_T]$ occurs in u, (ii) $T[\mathsf{LF}_T+1 : \mathsf{FB}_T-1]$ occurs in $w \cdot S[\mathsf{LF}_S+1 : \mathsf{FB}_S-1] \cdot z$ and (iii) $T[\mathsf{FB}_T : m]$ occurs in v. Then it follows that there exists a factorization $T[\mathsf{LF}_T+1 : \mathsf{FB}_T-1] = xy$ such that x occurs in $w \cdot S[\mathsf{LF}_S+1 : \mathsf{FB}_S-1]$ and y occurs in z, and hence $T' = T[1 : \mathsf{LF}_T] \cdot x \cdot (y \cdot T[\mathsf{FB}_T : m])^R$ occurs as an arc preserving subsequence in $S' = S[1 : \mathsf{FB}_S-1] \cdot S[\mathsf{FB}_S : n]^R$ (see Figure 8). □



(a)

(b)

Fig. 8. Illustration of Lemma 7.

Theorem 3. *The* APS($\{\between\}$,$\{\between\}$) *problem is solvable in* $O(nm^2)$ *time.*

Proof. The algorithm we propose is Algorithm 1.

Correctness of the algorithm follows from Lemma 7. What is left is to prove the time complexity. Clearly, $S' = S[1 : \mathsf{FB}_S - 1] \cdot S[\mathsf{FB}_S : n]^R$ is a $\{\sqsubset\}$-arc-annotated sequence. The key point is to note that, for any factorization $T[\mathsf{LF}_T + 1 : \mathsf{FB}_T - 1] = xy$, the obtained $T' = T[1 : \mathsf{LF}_T] \cdot x \cdot (y \cdot T[\mathsf{FB}_T : m])^R$ is a $\{\sqsubset\}$-arc-annotated sequence as well. Now let k be the number of arcs in T.

Algorithm 1: An $O(nm^2)$ time algorithm solving the APS($\{\between\}$,$\{\between\}$) problem

```
   Data   : Two $\{\between\}$-arc-annotated sequences $S$ and $T$ of length $n$ and $m$, respectively
   Result : true iff $T$ occurs as an arc-preserving subsequence in $S$
   begin
1      $S' = S[1 : \mathsf{FB}_S - 1] \cdot S[\mathsf{FB}_S : n]^R$
2      foreach factorization $T[\mathsf{LF}_T + 1 : \mathsf{FB}_T - 1]| = xy$ do
3          $T' = T[1 : \mathsf{LF}_T] \cdot x \cdot (y \cdot T[\mathsf{FB}_T : m])^R$
4          if $T'$ occurs as an arc preserving subsequence in $S'$ then
5              return true
6      return false
   end
```

So there are at most $m - 2k$ iterations to go before eventually returning false. According to the above, Line 4 constitutes an instance of APS($\{\sqsubset\}$,$\{\sqsubset\}$). But APS($\{\sqsubset\}$,$\{\sqsubset\}$) is a special case of APS($\{<, \sqsubset\}$,$\{<, \sqsubset\}$), and hence is solvable in $O(nm)$ time [11]. Then it follows that the algorithm as a whole runs in $O(nm(m - 2k)) = O(nm^2)$ time. □

Clearly, proof of Theorem 3 relies on an efficient algorithm for solving APS($\{\sqsubset$ $\}$,$\{\sqsubset\}$): the better the complexity for APS($\{\sqsubset\}$,$\{\sqsubset\}$), the better the complexity for APS($\{\between\}$,$\{\between\}$). We have used only the fact that APS($\{\sqsubset\}$,$\{\sqsubset\}$) is a special case of APS($\{<, \sqsubset\}$,$\{<, \sqsubset\}$). It remains open, however, wether a better complexity can be achieved for APS($\{\sqsubset\}$,$\{\sqsubset\}$).

Theorem 3 carries out easily to restricted versions (Observation 1).

Corollary 1. APS($\{\between\}$,$\emptyset$) *is solvable in* $O(nm^2)$ *time.*

6 Conclusion

In this paper, we investigated the APS problem time complexity and gave a precise characterization of what makes the APS problem hard. We proved that APS(Crossing,Plain) is **NP**-complete thereby answering an open problem posed in [11] (see Table 3). Note that this result answers the last open problem concerning APS computational complexity with respect to classical complexity levels, *i.e.*, Plain, Chain, Nested and Crossing. Also, we refined the four above mentioned levels for exploring the border between polynomial time solvable and **NP**-complete problems. We proved that both APS($\{\sqsubset, \between\}, \emptyset$) and APS($\{<, \between\}, \emptyset$) are **NP**-complete and gave positive results by showing that APS($\{\between\}$, $\emptyset$) and APS($\{\between\}$,$\{\between\}$) are polynomial time solvable. Hence, the refinement we suggest shows that APS problem becomes hard when one considers sequences containing $\{\between, \alpha\}$-comparable arcs with $\alpha \neq \emptyset$. Therefore, crossing arcs alone do not imply APS hardness. It is of course a challenging problem to further explore the complexity of the APS problem, and especially the parameterized views, by considering additional parameters such as the cutwidth or the depth of the arc structures.

APS								
R_1 \ R_2	$\{<, \sqsubset, \between\}$	$\{\sqsubset, \between\}$	$\{<, \between\}$	$\{\between\}$	$\{<, \sqsubset\}$	$\{\sqsubset\}$	$\{<\}$	$\emptyset$
$\{<, \sqsubset, \between\}$	**NP**-C [6]	**NP**-C $\star$	**NP**-C [12]	**NP**-C $\star$	**NP**-C [12]	**NP**-C $\star$	**NP**-C [12]	**NP**-C $\star$
$\{\sqsubset, \between\}$		**NP**-C $\star$	////	**NP**-C $\star$	////	**NP**-C $\star$	////	**NP**-C $\star$
$\{<, \between\}$			**NP**-C $\star$	**NP**-C $\star$	////	////	**NP**-C $\star$	**NP**-C $\star$
$\{\between\}$				$O(nm^2)$ $\star$	////	////	////	$O(nm^2)$ $\star$
$\{<, \sqsubset\}$					$O(nm)$ [11]	$O(nm)$ [11]	$O(nm)$ [11]	$O(nm)$ [11]
$\{\sqsubset\}$						$O(nm)$ [11]	////	$O(nm)$ [11]
$\{<\}$							$O(nm)$ [11]	$O(n+m)$ [11]
$\emptyset$								$O(n+m)$ [11]

Table 3. Complexity results after refinement of the complexity levels. $\star$: results from this paper.

References

1. J. Alber, J. Gramm, J. Guo, and R. Niedermeier. Towards optimally solving the longest common subsequence problem for sequences with nested arc annotations in linear time. In *Proc. of the 13th Symposium on Combinatorial Pattern Matching (CPM02)*, volume 2373 of *LNCS*, pages 99–114. Springer-Verlag, 2002.
2. J. Alber, J. Gramm, J. Guo, and R. Niedermeier. Computing the similarity of two sequences with nested arc annotations. *Theoretical Computer Science*, 312(2-3):337–358, 2004.
3. B. Billoud, M.-A. Guerrucci, M. Masselot, and J.S. Deutsch. Cirripede phylogeny using a novel approach: Molecular morphometrics. *Molecular Biology and Evolution*, 19:138–148, 2000.
4. G. Caetano-Anolls. Tracing the evolution of RNA structure in ribosomes. *Nucl. Acids. Res.*, 30:2575–2587, 2002.
5. W. Chaia and V. Stewart. RNA Sequence Requirements for NasR-mediated, Nitrate-responsive Transcription Antitermination of the Klebsiella oxytoca M5al nasF Operon Leader. *Journal of Molecular Biology*, 292:203–216, 1999.
6. P. Evans. *Algorithms and Complexity for Annotated Sequence Analysis.* PhD thesis, U. Victoria, 1999.
7. P. Evans. Finding common subsequences with arcs and pseudoknots. In *Proc. of the 10th Symposium Combinatorial Pattern Matching (CPM99)*, volume 1645 of *LNCS*, pages 270–280. Springer-Verlag, 1999.
8. A.D. Farris, G. Koelsch, G.J. Pruijn, W.J. van Venrooij, and J.B. Harley. Conserved features of Y RNAs revealed by automated phylogenetic secondary structure analysis. *Nucl. Acids. Res.*, 27:1070–1078, 1999.
9. M. Garey and D. Johnson. *Computers and Intractability: A Guide to the Theory of NP-Completeness.* W. H. Freeman and Company, 1979.
10. D. Goldman, S. Istrail, and C.H. Papadimitriou. Algorithmic aspects of protein structure similarity. In *Proc. of the 40th Symposium of Foundations of Computer Science (FOCS99)*, pages 512–522, 1999.
11. J. Gramm, J. Guo, and R. Niedermeier. Pattern matching for arc-annotated sequences. In *Proc. of the 22nd Conference on Foundations of Software Technology and Theoretical Computer Science (FSTTCS02)*, volume 2556 of *LNCS*, pages 182–193, 2002.

12. J. Guo. Exact algorithms for the longest common subsequence problem for arc-annotated sequences. Master's Thesis, Universitat Tubingen, Fed. Rep. of Germany, 2002.
13. K. Hellendoorn, P.J. Michiels, R. Buitenhuis, and C.W. Pleij. Protonatable hairpins are conserved in the 5'-untranslated region of tymovirus RNAs. *Nucl. Acids. Res.*, 24:4910–4917, 1996.
14. L. Hofacker, M. Fekete, C. Flamm, M.A. Huynen, S. Rauscher, P.E. Stolorz, and P.F. Stadler. Automatic detection of conserved RNA structure elements in complete RNA virus genomes. *Nucl. Acids. Res.*, 26:3825–3836, 1998.
15. T. Jiang, G.-H. Lin, B. Ma, and K. Zhang. The longest common subsequence problem for arc-annotated sequences. In *Proc. 11th Symposium on Combinatorial Pattern Matching (CPM00)*, volume 1848 of *LNCS*, pages 154–165. Springer-Verlag, 2000.
16. V. Juan, C. Crain, and S. Wilson. Evidence for evolutionarily conserved secondary structure in the H19 tumor suppressor RNA. *Nucl. Acids. Res.*, 28:1221–1227, 2000.
17. G. Lancia, R. Carr, B. Walenz, and S. Istrail. 101 optimal PDB structure alignments: a branch-and-cut algorithm for the maximum contact map overlap problem. In *Proceedings of the 5th ACM International Conference on Computational Molecular Biology (RECOMB01)*, pages 193–202, 2001.
18. S.W.M. Teunissen, M.J.M. Kruithof, A.D. Farris, J.B. Harley, W.J. van Venrooij, and G.J.M. Pruijn. Conserved features of Y RNAs: a comparison of experimentally derived secondary structures. *Nucl. Acids. Res.*, 28:610–619, 2000.
19. S. Vialette. Pattern matching over 2-intervals sets. In *Proc. 13th Annual Symposium Combinatorial Pattern Matching (CPM02)*, volume 2373 of *LNCS*, pages 53–63. Springer-Verlag, 2002.
20. S. Vialette. On the computational complexity of 2-interval pattern matching. *Theoretical Computer Science*, 312(2-3):223–249, 2004.
21. H.-Y. Wang and S.-C. Lee. Secondary structure of mitochondrial 12S rRNA among fish and its phylogenetic applications. *Molecular Biology and Evolution*, 19:138–148, 2002.
22. J. Wuyts, P. De Rijk, Y. Van de Peer, G. Pison, P. Rousseeuw, and R. De Wachter. Comparative analysis of more than 3000 sequences reveals the existence of two pseudoknots in area V4 of eukaryotic small subunit ribosomal RNA. *Nucl. Acids. Res.*, 28:4698–4708, 2000.
23. K. Zhang, L. Wang, and B. Ma. Computing the similarity between RNA structures. In *Proc. 10th Symposium on Combinatorial Pattern Matching (CPM99)*, volume 1645 of *LNCS*, pages 281–293. Springer-Verlag, 1999.
24. M. Zuker. RNA folding. *Meth. Enzymology*, 180:262–288, 1989.