

# The complexity of multiple sequence alignment with SP-score that is a metric

Paola Bonizzoni<sup>a,1</sup> Gianluca Della Vedova<sup>a,1</sup>

<sup>a</sup>*Dipartimento di Informatica, Sistemistica e Comunicazione, Università degli Studi di Milano - Bicocca, via Bicocca degli Arcimboldi 8, 20126 Milano - Italy, e-mail: {bonizzoni, dellavedova}@disco.unimib.it*

---

## Abstract

This paper analyzes the computational complexity of computing the optimal alignment of a set of sequences under the SP (sum of all pairs) score scheme. We solve an open question by showing that the problem is *NP*-complete in the very restricted case in which the sequences are over a binary alphabet and the score is a metric. This result establishes the intractability of multiple sequence alignment under a score function of mathematical interest, which has indeed received much attention in biological sequence comparison.

*Key words:* multiple sequence alignment, *SP*-score, intractability.

---

## 1 Introduction

Multiple sequence alignment is one of the most popular and important problems in computational biology [7]. It finds different applications in molecular biology, mainly in two related areas: finding information about the structure and function of the molecules, and estimate the evolutionary distance between species from their associated sequences.

An *alignment* of  $k$  sequences is defined by a matrix  $k \times m$  in which each row contains a sequence interleaved by spaces. Then, the similarity of sequences in the alignment is measured by using a score or *distance* between elements of the matrix. More precisely, in DNA (or RNA) sequences, the alphabet contains four letters and the score assigned to the comparison between two letters (or

---

<sup>1</sup> Supported by grant Cofinanziato 98: “Modelli di calcolo innovativi: metodi sintattici e combinatori”.

nucleotides) may be zero if there is a match, i.e. the letters are identical, otherwise the score may be one. A popular assumption in biological alignment is that the score is a metric, that is the distance between identical letters is zero and it satisfies the triangle inequality. Among different score schemes, the *sum of all pairs* score, in short the *SP*-score, is the one that has received more attention, mainly for its mathematical elegance. By means of the *SP*-score a value is assigned to a multiple alignment; an *optimal alignment* is the one that minimizes the value over all possible alignments.

Several methods have been developed for multiple sequence alignment [3,2], but no efficient methods are known to find the optimal alignment. Recently, a polynomial time approximation algorithm for the problem has been proposed by Gusfield [6] who achieved a  $2 - 2/k$  approximation factor by assembling an alignment of  $k$  sequences from optimal alignment of pairs of sequences. The approximation ratio has been improved to a  $2 - l/k$  factor, for any fixed  $l$ , by Bafna, Lawler and Pevzner [1]. But, besides these results it was an open question whether the problem is *NP*-complete. The computational complexity of multiple sequence alignment has been investigated in [9] where is given a simple proof of *NP*-completeness of the alignment with score scheme over a fixed alphabet of four letters that satisfies the triangle inequality, and assigns a non zero distance between identical letters. But, this result leaves open the problem of analyzing the complexity of computing optimal *SP*-score multiple sequence alignments for instances of this problem which are of practical biological relevance. Mainly, the result in [9] does not consider an important requirement for score schemes ([4,11]) which is the property of *metricity*: this one implies a zero distance between identical letters.

Here, we prove the intractability of multiple sequence alignment in the very restricted case in which sequences are over a binary alphabet and the score is a metric. The significance of the intractability in this case is that it establishes the *NP*-completeness for less restricted cases encountered in practice, as well as for general instances of the alignment problem in which  $|\Sigma| > 2$ .

## 2 Preliminaries

A *DNA sequence* is a string over the alphabet  $\Sigma$  that contains four letters  $A, C, G$  and  $T$  representing four distinct nucleotides. Protein sequences are over an alphabet of 20 letters, each representing a unique amino acid. A *multiple alignment* of  $k$  sequences is obtained by inserting spaces in the sequences such that the sequences have the same length  $l$  and they can be arrayed in  $k$  rows of  $l$  columns each. A space is denoted by  $\Delta$  and is viewed as a new letter over alphabet  $\Gamma = \Sigma \cup \{\Delta\}$ . Given two sequences  $s_1$  and  $s_2$  in the alignment, then each letter  $\sigma$  of  $s_1$  is in the same column of a letter of  $s_2$ ; we say that  $\sigma$  is

*opposite* to a unique letter of  $s_2$ . A *match* occurs where two identical letters are opposite in the two sequences  $s_1$  and  $s_2$ , otherwise two non-identical opposing letters give a *mismatch* which is viewed as a replacement. The insertion of a space in a sequence opposing a letter  $\sigma$  of a second sequence, is viewed as the deletion in the first sequence of the letter  $\sigma$  or an insertion of  $\sigma$  into the second one. A score  $d$  is assigned to each pair of letters and it is generally described by means of a  $|\Gamma| \times |\Gamma|$  symmetric matrix. The following properties are considered a mathematical requirement for cost matrices: [3,10]

- (i)  $d(a, a) = 0$ , for every  $a \in \Gamma$ ,
- (ii)  $d(a, b) = 0$  implies that  $a = b$ , for every  $a, b \in \Gamma$ ,
- (iii)  $d(a, b) = d(b, a)$ , for every  $a, b \in \Gamma$ ,
- (iv)  $d(a, c) \leq d(a, b) + d(b, c)$ , for every  $a, b, c \in \Gamma$ ,
- (v)  $d(a, c) \leq \max\{d(a, b), d(b, c)\}$ , for every  $a, b, c \in \Gamma$ .

A score scheme that satisfies properties (i) – (iii) is a semi-metric, the score is a metric if property (iv) is also satisfied and is an ultrametric if all above specified properties hold.

By means of a score scheme a value is assigned to a multiple alignment. A very popular score scheme, called *SP-score* defines the value of a multiple alignment as the sum of the scores of all columns, where the score of each column is the sum of the scores of all distinct unordered pairs of letters in the column. Then, the value of the alignment of a column  $x$  of height  $l$  is  $\sum_{1 \leq i < j \leq l} d(x(i), x(j))$ , where  $x(i)$  is the letter in  $i$ -th row of column  $x$  and  $d(x(i), x(j))$  is the score between the two letters  $x(i)$  and  $x(j)$ .

Another way of viewing the *SP-score* value of an alignment is as sum of pairwise sequence alignment values: given  $\mathcal{A}$  an alignment with  $m$  rows and  $k$  columns, and  $s_i, s_j$  the  $i$ -th and  $j$ -th rows of  $\mathcal{A}$ , the value of the pairwise alignment in  $\mathcal{A}$  of  $s_i$  and  $s_j$ , denoted as  $d_{\mathcal{A}}(s_i, s_j)$  is  $\sum_{1 \leq l \leq k} d(s_i(l), s_j(l))$ , where  $s_i(l)$  ( $s_j(l)$ ) is the  $l$ -th symbols of  $s_i$  ( $s_j$  respectively). Then, the value of  $\mathcal{A}$  can be expressed as  $\sum_{1 \leq i < j \leq m} d_{\mathcal{A}}(s_i, s_j)$ . We assume that an alignment cannot contain a column of only  $\Delta$ 's. An *optimal multiple alignment* of a set of sequences is the one that minimizes the value over all possible alignments.

Let  $B$  be a subset of a set  $\mathcal{S}$  of sequences and  $\mathcal{A}$  an alignment of  $\mathcal{S}$ . Then, by  $\mathcal{A}_B$  we denote the array consisting of all rows of  $\mathcal{A}$  containing sequences in  $B$  (in this case in  $\mathcal{A}_B$  there may be some columns containing only  $\Delta$ 's).

By  $D(\mathcal{A})$  we will denote the value of an alignment  $\mathcal{A}$  of a set of sequences. By  $\mathcal{A}[i]$ , we denote the column of  $\mathcal{A}$  of index  $i$ . Let  $B$  and  $C$  be two disjoint subsets of sequences of  $\mathcal{S}$ , and let  $B(t)$  and  $C(t)$  be the  $t$ -th sequence in  $B$  and  $C$ , respectively, then by  $D(\mathcal{A}_{B,C})$  we denote  $\sum_{i,j} d_{\mathcal{A}}(B(i), C(j))$ .

**Lemma 1** *Let  $s_1, s_2$  be two sequences over  $\Sigma$  such that  $l_1 = |s_1|, l_2 = |s_2|$ ,*

$l_2 \geq l_1$  and there are  $m$  symbols of  $s_1$  that are not in  $s_2$ . Then every alignment of the set  $\{s_1, s_2\}$  has at least  $m + l_2 - l_1$  mismatches.

The following two properties hold for every alignment over a score which is a metric and has non null values greater or equal to 1.

**Corollary 2** Let  $s_1, s_2$  be two sequences over  $\Sigma$ , such that  $l_1 = |s_1|$ ,  $l_2 = |s_2|$ ,  $l_2 \geq l_1$  and there are  $m$  symbols of  $s_1$  that are not in  $s_2$ . Then for every alignment of the set  $\{s_1, s_2\}$ ,  $D(\mathcal{A}_{\{s_1, s_2\}}) \geq m + l_2 - l_1$ .

**PROOF.** It follows from Lemma 1.

**Lemma 3** Let  $U$  be a subset of a set  $S$  of sequences over  $\Sigma$  such that  $U$  contains only identical sequences, and let  $\mathcal{A}$  be an optimal alignment of  $S$ . Then  $D(\mathcal{A}_U) = 0$ .

**PROOF.** Assume to the contrary that  $\mathcal{A}$  is an optimal alignment of  $S$  and  $D(\mathcal{A}_U) > 0$ . Let  $u \in U$  be the sequence that minimizes the value  $D(\mathcal{A}_{\{u\}, S-U})$ . Then, let  $\mathcal{A}_1$  be the alignment obtained from  $\mathcal{A}$  by assuming that all sequences in  $S - U$  are aligned as in  $\mathcal{A}$  (i.e.  $\mathcal{A}_{S-U} = \mathcal{A}_{1S-U}$ ), while all sequences in  $U$  are aligned identically to the alignment of  $u$  in  $\mathcal{A}$ . Since  $D(\mathcal{A}_{S-U}) = D(\mathcal{A}_{1S-U})$  and  $D(\mathcal{A}_{1U}) < D(\mathcal{A}_U)$ , it follows that  $D(\mathcal{A}_1) < D(\mathcal{A})$ , which is a contradiction.

We will prove that multiple sequence alignment is *NP*-complete over a fixed score scheme that is a metric, by giving a reduction from the node cover problem (*NC*) which is *NP*-complete [5].

The *NC* and sequence alignment decision problems are defined in the following.

**NC** (*Node cover*)

**Instance:** A graph  $G = (V, E)$  and an integer  $k \leq |V|$ .

**Question:** Is there a node cover  $V_1$  of  $G$  of size  $k$  or less, i.e. a subset  $V_1$  of  $V$  such that for each edge  $e = (u, v) \in E$  at least one of  $u$  and  $v$  belongs to  $V_1$ ?

## Multiple Sequence Alignment

**Instance:** A set  $\mathcal{S} = \{s_1, \dots, s_n\}$  of finite sequences over a fixed alphabet  $\Sigma$  and a *SP*-score. An integer  $C$ .

**Question:** Is there a multiple alignment of the sequences in  $\mathcal{S}$  that is of value  $C$  or less?

### 3 Multiple alignment over alphabet of size 6

We first describe a reduction from the node cover problem on graphs ( $NC$ ) [5] to sequence alignment over an alphabet of size 6. The proof for the case of binary alphabet, that will be stated in section 4, is rather involved. The encoding used to obtain the result in the current section does not require the same level of complexity, while it allow us to point out the main ideas on which both reductions are based.

The  $SP$ -score for multiple alignment over alphabet  $\Sigma = \{a, b, 0, 1, c, d\}$  is the one described in the following Table 1.

	$a$	$b$	$0$	$1$	$c$	$d$	$\Delta$
$a$	0	1	1	1	1	2	2
$b$	1	0	1	1	2	1	2
$0$	1	1	0	2	2	2	2
$1$	1	1	2	0	2	1	2
$c$	1	2	2	2	0	2	1
$d$	2	1	2	1	2	0	2
$\Delta$	2	2	2	2	1	2	0

Table 1  
 $SP$ -score for alphabet of size 6

**The reduction.** The transformation from  $NC$  to alignment consists of constructing a set  $\mathcal{S}$  of sequences encoding the graph  $G$  and a value  $C$ , depending on  $k$  and on the number  $l$  of edges of  $G$ , such that  $C$  is an upper bound for the value of an optimal alignment of  $\mathcal{S}$  if and only if  $k$  is the size of a node cover for  $G$ .

Let  $G = (V, E)$  be a graph with  $V = \{v_1, \dots, v_n\}$  and  $E = \{e_1, \dots, e_l\}$ .

Now we construct an encoding for the edges of the graph that gives the set of sequences which is instance of the alignment problem. In the following, given a letter  $\sigma \in \Gamma$ , and an integer  $j > 0$ , by  $\sigma^j$  we denote the sequence of  $j$  symbols  $\sigma$ .

Given an edge  $e = (v_i, v_j)$ , where we assume that  $i < j$ , we construct an encoding of such an edge with a sequence, called *edge sequence* constructed as

follows:

$$s(i, j) = a^{3i} b a^{3(j-i)-2} b a^{3(n-j)+3}.$$

Note that for each edge  $(v_i, v_j)$  the edge sequence  $s(i, j)$  has length  $3n + 3$ . Moreover, we construct a *template sequence*  $t$  of length  $3n + 4$ :

$$t = c(001)^n 00c$$

We also construct the *test sequence*  $x(k)$  of the form:

$$x(k) = cd^k c$$

Note that the test sequence depends on  $k$ . The set of sequences that is instance of the alignment problem associated to the instance  $(G, k)$  of the *NC* problem, is the set  $\mathcal{S} = S \cup T \cup X$ , where  $S = \{s(i, j) : (v_i, v_j) \in E\}$ ,  $T$  contains  $C_2$  sequences  $t$  and  $X$  contains  $C_1$  sequences  $x(k)$ , where  $C_1$  and  $C_2$  will be determined later.

In Fig. 1 is represented an alignment of the encoding of a graph  $G$ .

The main idea on which the encoding of  $\mathcal{S}$  is based, is that an optimal alignment  $\mathcal{A}$  of  $\mathcal{S}$  is obtained when  $\mathcal{A}$  satisfies certain properties: such an alignment will be called *standard alignment*. More precisely we will show that the value of a standard alignment is bounded by a given threshold  $C$  only when  $G$  has a node cover of a given size  $k$ . This fact is obtained by forcing  $d$ 's of the test sequences to be opposite to  $b$ 's of the edge-sequences. By construction, only one  $b$  of each edge sequence can be opposite to a  $d$ , and the number of such  $b$ 's determines the value of the alignment. If the total number of  $b$ 's opposite to  $d$ 's is equal to the number of edges, which is possible only if there are  $k$  vertices which cover one end of each edge sequence, then  $D(\mathcal{A}) < C$ , otherwise  $D(\mathcal{A}) > C$ .

**Definition 4 (Standard alignment)** *Let  $\mathcal{A}$  be an alignment of  $\mathcal{S}$ . Then  $\mathcal{A}$  is a standard alignment if it satisfies the following properties:*

- (i) *there are no  $\Delta$ 's in  $\mathcal{A}_T$ ;*
- (ii) *all  $\Delta$ 's in  $\mathcal{A}_S$  are aligned with  $c$ 's of  $\mathcal{A}_T$ ;*
- (iii) *all  $d$ 's of  $\mathcal{A}_X$  are aligned with  $1$ 's of  $\mathcal{A}_T$ ;*
- (iv) *all  $c$ 's of  $\mathcal{A}_X$  are aligned with  $c$ 's of  $\mathcal{A}_T$ ;*
- (v) *no column of  $\mathcal{A}_X$  contains both  $\Delta$ 's and  $d$ 's.*

It follows easily that each standard alignment has exactly  $r = 3n + 4$  columns. Note that in Fig. 1 is represented a standard alignment of  $\mathcal{S}$  where, for simplicity, all  $\Delta$ 's are not shown.

Fig. 1. An example of the encoding of a graph

In the following we give some useful properties of standard alignments:

**Proposition 5** *Let  $\mathcal{A}$  be a standard alignment of  $\mathcal{S}$ . Then for each edge sequence encoding the edge  $(v_i, v_j)$ , the  $b$  encoding one end vertex  $v_h$ , for  $h \in \{i, j\}$ , is opposite to the  $h$ -th 1 of each template sequence, while the other  $b$  is opposite to 0's of the template sequences.*

**Lemma 6** *Let  $\mathcal{A}$  be an optimal alignment of  $\mathcal{S}$  and let  $\mathcal{A}_1$  be a standard alignment of  $\mathcal{S}$ . Then  $D(\mathcal{A}_{X,T}) = D(\mathcal{A}_{1X,T})$  or  $D(\mathcal{A}_{X,T}) \geq D(\mathcal{A}_{1X,T}) + C_2$ .*

**PROOF.** Assume that  $D(\mathcal{A}_{X,T})$  is minimum over all possible alignments of  $\mathcal{S}$ . Then,  $\mathcal{A}_{\{x,t\}}$  contains exactly  $r - 2$  mismatches of value 1, for every  $x \in X$  and  $t \in T$ . In fact, since the mismatches  $(d, 1)$  have a value 1, while the mismatches  $(d, 0)$ ,  $(d, \Delta)$ ,  $(d, c)$  all have value 2 it is advantageous to align all  $d$ 's of  $x$  with 1's of  $t$ . By the SP-score  $d(c, c) = 0$ ,  $d(c, \Delta) = 1$ ,  $d(c, 0) = d(c, 1) = d(\Delta, 0) = d(\Delta, 1) = 2$ , it follows that it is advantageous to align the  $c$ 's of  $x$  with the  $c$ 's of  $t$ . Note that any other alignment of  $\{x, t\}$  cannot be optimal. It is immediate to verify that  $D(\mathcal{A}_{X,T}) \geq D(\mathcal{A}_{1X,T})$ .

Now, assume that  $D(\mathcal{A}_{X,T}) \neq D(\mathcal{A}_{1X,T})$ . Since for every sequence  $x \in X$  and  $t \in T$ ,  $\mathcal{A}_{1\{x,t\}}$  contains exactly  $r - 2$  mismatches of value 1, then there is a sequence  $x_1$  such that  $\mathcal{A}_{\{x_1,t\}}$  must contain at least  $r - 1$  mismatches. Since, by Lemma 3,  $D(\mathcal{A}_X) = 0$ ,  $D(\mathcal{A}_T) = 0$ , and  $|T| = C_2$ , it follows that  $D(\mathcal{A}_{X,T}) = C_2|X|D(\mathcal{A}_{\{x_1,t\}})$ . Consequently, since  $D(\mathcal{A}_{1X,T}) = C_2|X|D(\mathcal{A}_{1\{x_1,t\}})$ , it follows

that  $D(\mathcal{A}_{X,T}) \geq D(\mathcal{A}_{1X,T}) + C_2$ , as required.

**Lemma 7** *Let  $\mathcal{A}$  be an optimal alignment of  $\mathcal{S}$  and let  $\mathcal{A}_1$  be a standard alignment of  $\mathcal{S}$ . Then  $D(\mathcal{A}_{S,T}) = D(\mathcal{A}_{1S,T})$ , or  $D(\mathcal{A}_{S,T}) \geq D(\mathcal{A}_{1S,T}) + C_2$ .*

**PROOF.** Assume that  $D(\mathcal{A}_{S,T})$  is minimum over all possible alignments of  $\mathcal{S}$ . Then,  $\mathcal{A}_{\{s,t\}}$  contains exactly  $r$  mismatches of value 1, for every  $s \in S$  and  $t \in T$ . In fact, by Corollary 2, since there is no symbol common to both sequences  $s$  and  $t$ , and  $|t| = r$ ,  $|s| = r - 1$  it follows that every alignment  $\mathcal{A}'$  for  $\mathcal{S}$  is such that  $D(\mathcal{A}'_{\{s,t\}}) \geq r$ . By construction of standard alignment and by the SP-score, it follows easily that  $D(\mathcal{A}_{1\{s,t\}}) = r$ .

Now, assume that  $D(\mathcal{A}_{S,T}) \neq D(\mathcal{A}_{1S,T})$ . Then, there is a sequence  $s_1 \in S$  such that  $\mathcal{A}_{\{s_1,t\}}$  must contain either at least  $r + 1$  mismatches or  $r$  mismatches one of which is of value 2. Since, by Lemma 3,  $D(\mathcal{A}_T) = 0$ , and  $|T| = C_2$ , it follows that  $D(\mathcal{A}_{S,T}) = C_2 D(\mathcal{A}_{\{s_1,t\}}) + D(\mathcal{A}_{S-\{s_1\},T})$ . But,  $D(\mathcal{A}_{1S,T}) = C_2 D(\mathcal{A}_{1\{s_1,t\}}) + D(\mathcal{A}_{1S-\{s_1\},T})$ . It follows that  $D(\mathcal{A}_{S,T}) \geq D(\mathcal{A}_{1S,T}) + C_2$ , as required.

**Corollary 8** *Let  $\mathcal{A}$  be a standard alignment of  $\mathcal{S}$ . Then  $D(\mathcal{A}_X)$ ,  $D(\mathcal{A}_T)$ ,  $D(\mathcal{A}_{X,T})$  and  $D(\mathcal{A}_{S,T})$  are fixed and minimum over all possible alignments.*

**PROOF.** By definition of standard alignment and by Lemmas 6, 7, the proof is immediate.

In the following by  $D_{SD}$  we denote the sum  $D(\mathcal{A}_X) + D(\mathcal{A}_T) + D(\mathcal{A}_{X,T}) + D(\mathcal{A}_{S,T})$  over a standard alignment  $\mathcal{A}$  of  $\mathcal{S}$ .

**Lemma 9** *Let  $\mathcal{A}$  be a standard alignment of  $\mathcal{S}$ . Then  $D(\mathcal{A}_S) < 8l^2r$  and  $D(\mathcal{A}_{S,X}) < 4C_1lr$ .*

**PROOF.** It follows easily from the SP-score and the fact that a standard alignment consists of  $r$  columns.

We will denote such upper bounds for  $D(\mathcal{A}_S)$  and  $D(\mathcal{A}_{S,X})$  with  $U_S$  and  $U_{S,X}$  respectively. By Corollary 8 and Lemma 9 it follows easily that each standard alignment (hence each optimal alignment) has value not greater than  $D_{SD} + U_S + U_{S,X}$ . By assuming that  $C_1 > U_S$  and  $C_2 > U_S + U_{S,X}$ , we can prove that an optimal alignment must be a standard one.

**Lemma 10** *Let  $\mathcal{A}$  be an optimal alignment of  $\mathcal{S}$ . Then  $\mathcal{A}$  must be a standard alignment.*

**PROOF.** Let  $\mathcal{A}_1$  be a standard alignment of  $\mathcal{S}$ . If  $\mathcal{A}$  does not satisfy one of the properties (i) – (iv) of standard alignment, it is immediate to verify that  $D(\mathcal{A}_{X,T}) \neq D(\mathcal{A}_{1X,T})$  or  $D(\mathcal{A}_{S,T}) \neq D(\mathcal{A}_{1S,T})$ . By Lemmas 6, 7, it follows that  $D(\mathcal{A}_{X,T}) + D(\mathcal{A}_{S,T}) \geq D(\mathcal{A}_{1X,T}) + D(\mathcal{A}_{1S,T}) + C_2$ . Since  $C_2 > U_S + U_{S,X}$ , by Corollary 8 it follows that  $\mathcal{A}$  is not optimal.

Assume now that  $\mathcal{A}$  does not satisfy property (v) of standard alignment. Then by Lemma 3,  $\mathcal{A}$  cannot be optimal. Consequently,  $\mathcal{A}$  must be a standard alignment.

We are now able to prove that the multiple alignment problem is NP-complete with a fixed SP-score that is a metric and with an alphabet of six symbols. In the following, if  $\mathcal{A}$  is an alignment of  $\mathcal{S}$ , by  $n_\sigma(i)$  we denote the number of  $\sigma$ 's occurring in the column of index  $i$  of  $\mathcal{A}$ .

**Theorem 11** *Let  $(G, k)$  be an instance of the NC problem and let  $\mathcal{S}$  be the encoding of such instance. Then:*

- (i) *if  $G$  has a node cover of size  $k$ , then there exists a standard alignment  $\mathcal{A}$  of  $\mathcal{S}$  such that  $D(\mathcal{A}_{S,X}) \leq C_1(l + 2l(r - 2))$ ;*
- (ii) *if  $G$  has a minimum node cover of size  $k_1 > k$ , then for each standard alignment  $\mathcal{A}$  of  $\mathcal{S}$  it holds that  $D(\mathcal{A}_{S,X}) > U_S + C_1(l + 2l(r - 2))$ .*

**PROOF.** Let  $\mathcal{A}$  be a standard alignment of  $\mathcal{S}$ , and let  $I$  be the set of indices of the columns of  $\mathcal{A}$  containing some  $d$ 's. By the definition of standard alignment the value  $D(\mathcal{A}_{S,X})$  can be computed as the sum of the value of the column of index 1 and  $r$  and the value of all other columns of  $\mathcal{A}_{S \cup X}$ . Then,  $D(\mathcal{A}_{S,X}) = C_1(ld(c, \Delta) + ld(c, a)) + C_1(\sum_{i \in I} (d(d, b)n_b(i) + d(d, a)(l - n_b(i))) + \sum_{i \notin I \cup \{1, r\}} (d(\Delta, b)n_b(i) + d(\Delta, a)(l - n_b(i)))) = C_1(2l + 2l(r - 2) - \sum_{i \in I} n_b(i))$ .

Let us assume that  $G$  has a node cover  $K$  of size  $k$ , then we will construct a standard alignment  $\mathcal{A}$  such that  $D(\mathcal{A}_{S,X}) \leq C_1(l + 2l(r - 2))$ . From the node cover  $K$  we construct the set  $K_1$  consisting of the indices of the columns in  $\mathcal{A}_T$  that contain the 1's encoding the vertices in  $K$ . Since  $K$  is a node cover of  $G$  each edge  $(v_i, v_j)$  has at least an end vertex  $v_h$  in  $K$ , for  $h \in \{i, j\}$ , moreover it is possible to align, in each edge sequence, the  $b$  encoding the vertex  $v_h$  with the  $h$ -th 1 of each template sequence. The alignment of the test sequences in  $\mathcal{A}_{S \cup X}$  is obtained by aligning the  $d$ 's exactly in the columns whose index is in  $K_1$ . By Proposition 5, since only a  $b$  for each edge sequence can be aligned

	$a$	$b$	$\Delta$
$a$	0	1	2
$b$	1	0	1
$\Delta$	2	1	0

Table 2

$SP$ -score for binary alphabet

with a 1 of each template sequence. It follows that  $\sum_{i \in I} n_b(i) = l$ . Substituting this value in the above relation for  $D(\mathcal{A}_{S,X})$ , then (i) easily follows.

Let us assume that  $G$  has a node cover of minimum size  $k_1 > k$ . By Proposition 5 for each edge sequence encoding the edge  $(v_i, v_j)$ , one  $b$  of each edge sequence, encoding the end vertex  $v_h$ , for  $h \in \{i, j\}$ , is aligned with the  $h$ -th 1 of each template sequence, hence there must be at least  $k_1$  columns of  $\mathcal{A}$  that contain some 1's of the template sequences and at least a  $b$  of the edge sequences. By properties (i), (iii) and (v) of standard alignment, in all test sequence each  $d$  is aligned with distinct 1's of the template sequences: it follows that there is at least one edge sequence such that both  $b$ 's are in columns of  $\mathcal{A}_{S \cup X}$  that do not contain any  $d$ 's. Consequently, given  $I$  the set of indices of the columns of  $\mathcal{A}$  that contain some  $d$ 's of the test sequences,  $\sum_{i \in I} n_b(i) \leq l - 1$ , hence  $D(\mathcal{A}_{S,X}) \geq C_1(l + 2l(r - 2) + 1)$ . Since  $C_1 > U_S$  we obtain that  $D(\mathcal{A}_{S,X}) > U_S + C_1(l + 2l(r - 2))$ , which proves (ii).

**Corollary 12** *The graph  $G$  has a node cover of size  $k$  iff the set  $\mathcal{S}$  has an optimal alignment  $\mathcal{A}$  of value  $D(\mathcal{A}) < D_{SD} + U_S + C_1(l + 2l(r - 2))$ .*

**PROOF.** Let  $\mathcal{A}$  be an optimal alignment of  $\mathcal{S}$ . By Lemma 10,  $\mathcal{A}$  is a standard alignment. Then  $D(\mathcal{A}) = D_{SD} + D(\mathcal{A}_S) + D(\mathcal{A}_{S,X})$ . By Theorem 11, if  $G$  has a node cover of size  $k$ , then  $D(\mathcal{A}_{S,X}) \leq C_1(l + 2l(r - 2))$ .

Assume now that  $G$  has a minimum node cover of size  $k_1 > k$ : by Theorem 11,  $D(\mathcal{A}_{S,X}) > U_S + C_1(l + 2l(r - 2))$ . Consequently,  $D(\mathcal{A}) > D_{SD} + U_S + C_1(l + 2l(r - 2))$ , which proves what required.

#### 4 Multiple alignment over binary alphabet

In this section, we show that multiple sequence alignment problem is  $NP$ -complete even when the sequences are over a binary alphabet and the score scheme, which is a metric, is given in Table 2:

Given  $(G, k)$  the  $NC$  instance, where  $G$  is the graph  $(V, E)$ , with  $V = \{v_1, \dots, v_n\}$

and  $E = \{e_1, \dots, e_l\}$ , while  $1 \leq k \leq n$ , we construct the following sequences over alphabet  $\Sigma = \{a, b\}$ :

the *edge sequence*  $s(i, j)$  of length  $3(n + 1)$ , for each edge  $(v_i, v_j)$ ,

$$s(i, j) = a^{3i}ba^{3(j-i)-2}ba^{3(n+1-j)}$$

the *template sequence*  $t$  of length  $3(n + 1) + 1$ ,

$$t = b((a^2)b)^n(a^2)b$$

the *fixed sequence*  $q$  of length  $3(n + 1) + 1$ ,

$$q = ba^{3n+2}b,$$

the *test sequence*  $x(k)$ , given  $k$  the integer of the *NC* instance,

$$x(k) = a^{3n+2-k}.$$

Then, let  $S$  be the set  $\{s(i, j) : (v_i, v_j) \in E\}$  of all possible edge sequences,  $T$  the set of  $C_1$  template sequences  $t$ ,  $Q$  the set of  $C_2$  fixed sequences  $q$  and  $X$  the set of  $C_3$  test sequences  $x(k)$ . The constants  $C_1, C_2$  and  $C_3$  are related to the number of edges, and will be fixed later in the paper.

Finally, the sequences in  $S \cup T \cup Q \cup X$  give the set  $\mathcal{S}$  that is instance of the alignment problem.

Fig. 2. Alignment  $\mathcal{A}$  for  $\mathcal{S}$  in the case of binary alphabet

In the following, we give some properties that allow us to show that a node cover for  $G$  is of size  $k$  iff the cost of an alignment of  $\mathcal{S}$  can be bounded by a value  $C$ , depending on  $k$  and on the graph  $G$ , as stated in Theorem 25. By this result, the proof that the construction of the instance  $(\mathcal{S}, C)$  for the alignment problem is a polynomial reduction is immediate.

**Definition 13** An alignment  $\mathcal{A}$  of the set  $\mathcal{S}$  of sequences is a standard alignment if it satisfies properties (i), (ii), (iii) and (iv):

- (i) in all columns  $\mathcal{A}[i]$  such that  $\mathcal{A}_{T \cup Q}[i]$  contains some  $\Delta$ 's there are only  $\Delta$ 's in  $\mathcal{A}_{T \cup Q \cup X}[i]$  and no  $a$ 's in  $\mathcal{A}_S[i]$ ;
- (ii) all  $\Delta$ 's in  $\mathcal{A}_S$  are opposite only to  $\Delta$ 's or to  $b$ 's in  $\mathcal{A}_Q$ ;
- (iii) in  $\mathcal{A}_X$  there is no column with both  $a$ 's and  $\Delta$ 's, and the first and last column of  $\mathcal{A}_X$  consist of  $\Delta$ 's;
- (iv) the  $\Delta$ 's of  $\mathcal{A}_X$  are contained only in columns that do not contain any  $a$ 's of  $\mathcal{A}_T$ .

We will show that an optimal alignment must be a standard alignment; the properties of Definition 13 will allow us to relate the value of the alignment to the size of the node cover of the graph.

Let  $\mathcal{A}$  be a standard alignment. Then, by the previous definition, it is immediate to verify that conditions (i) and (ii) imply that in  $\mathcal{A}$ , all  $\Delta$ 's in internal columns of  $\mathcal{A}_S$  have a mismatch only with  $b$ 's of sequences in  $S$ . Moreover,  $\Delta$ 's in the first and last column of  $\mathcal{A}_S$  mismatch only with  $b$ 's of  $\mathcal{A}_Q$ , otherwise by condition (i), (ii) and (iv), there is a column in  $\mathcal{A}$  of only  $\Delta$ 's, which is not possible. By this fact and Definition 13, the Proposition 14 easily follows.

**Proposition 14** Let  $\mathcal{A}$  be an alignment of  $\mathcal{S}$  that satisfies properties (i) and (ii) of a standard alignment. For each edge sequence  $s(i, j)$  in  $S$ , one of the two  $b$ 's encoding one end vertex  $v_h$  of the edge  $(v_i, v_j)$  is aligned in  $\mathcal{A}$  with the  $(h + 1)$ -th  $b$  of each template sequence in  $T$ , while the other  $b$  of  $s(i, j)$  has a mismatch with each symbol of the template sequences to which it is opposite.

**Lemma 15** Let  $\mathcal{A}$  be a standard alignment of  $\mathcal{S}$ . Then  $D(\mathcal{A}_S) < 6l^2$ ,  $D(\mathcal{A}_{S,X}) < 8lkC_3$  and  $D(\mathcal{A}_{S,T}) < 4l^2C_1$ .

**PROOF.** Let us first prove the upper bound for  $D(\mathcal{A}_S)$ . By condition (i) and (ii) of Definition 13,  $\Delta$ 's in internal columns of  $\mathcal{A}_S$  have a mismatch only with  $b$ 's. Then, mismatches occur only in columns with  $b$ 's. Since there are exactly  $2l$   $b$ 's, it follows that the cost of all internal columns is bounded by  $2l^2$ , while the cost of the first and last column of  $\mathcal{A}_S$  is bounded by  $4l^2$ , as in such columns  $\Delta$ 's have mismatch with  $a$ 's. Consequently,  $D(\mathcal{A}_S) < 6l^2$ .

Now, let us prove that  $D(\mathcal{A}_{S,X}) < 8lkC_3$ . Let  $s$  and  $x$  be two arbitrary sequences in  $S$  and  $X$ , respectively. By condition (i), (ii) and (iii),  $\Delta$ 's in  $\mathcal{A}_S$  are opposite only to  $\Delta$ 's of  $\mathcal{A}_X$ . Since  $|s| = 3(n + 1)$ , while  $|x| = 3n + 2 - k$  and  $s$  contains two  $b$ 's which are not in  $x$ , it follows that  $\mathcal{A}_{\{s,x\}}$  contains at most  $k + 3$  mismatches, from which we prove that  $D(\mathcal{A}_{S,X}) < 2(k + 3)lC_3$ . Thus the required bound follows.

By Proposition 14 and just as in the above proof, it easily follows that  $D(\mathcal{A}_{S,T}) < 4l^2C_1$ .

In the rest of the paper, we will denote the upper bounds given in Lemma 15, respectively as  $U_S$ ,  $U_{S,X}$  and  $U_{S,T}$ . We pose that  $C_1 > l$ ,  $C_2 > U_S + U_{S,X} + U_{S,T}$  and  $C_3 > U_S$ .

**Lemma 16** *Let  $s, q$  be two sequences with  $s \in S$ ,  $q \in Q$  and let  $\mathcal{A}$  be an alignment of  $\mathcal{S}$ . Then  $\mathcal{A}_{\{s,q\}}$  contains at least four mismatches (and  $D(\mathcal{A}_{\{s,q\}}) \geq 4$ ). Moreover if  $\mathcal{A}$  is a standard alignment, then  $\mathcal{A}_{\{s,q\}}$  contains exactly four mismatches and  $D(\mathcal{A}_{\{s,q\}}) = 4$ .*

**PROOF.** By construction of sequences  $s, q$ , and by definition of standard alignment, it is immediate to note that, in a standard alignment  $\mathcal{A}_1$ , both  $b$ 's in  $s$  have a mismatch with some  $a$ 's or  $\Delta$ 's of  $\mathcal{A}_{1\{q\}}$ , and both  $b$ 's in  $q$  have a mismatch with some  $a$ 's or  $\Delta$ 's of  $\mathcal{A}_{1\{s\}}$ . By the SP-score  $D(\mathcal{A}_{\{s,q\}}) = 4$ . Along the same line it is immediate to note that if  $\mathcal{A}_{\{s,q\}}$  is an arbitrary alignment where all  $b$ 's have a mismatch, then  $D(\mathcal{A}_{\{s,q\}}) \geq 4$ .

Assume now that in a non standard alignment a  $b$  of  $s$  and a  $b$  of  $q$  are aligned in the same column; we will prove that  $D(\mathcal{A}_{\{s,q\}}) > 4$ . Clearly, the smallest number of mismatches is given by assuming that the first  $b$  of  $s$  is aligned with the first  $b$  of  $q$ , or the second  $b$  of  $s$  is aligned with the last  $b$  of  $q$ . Then, by construction of  $s$  and  $q$  the first three or last three  $a$ 's of  $s$  have a mismatch with some  $\Delta$ 's, hence  $D(\mathcal{A}_{\{s,q\}}) \geq 6$ .

The following results are direct consequences of Definition 13 and Lemma 16.

**Lemma 17** *Let  $\mathcal{A}$  be a standard alignment of  $\mathcal{S}$ . Then  $D(\mathcal{A}_X)$ ,  $D(\mathcal{A}_{T,Q})$ ,  $D(\mathcal{A}_T)$ ,  $D(\mathcal{A}_Q)$ ,  $D(\mathcal{A}_{X,T \cup Q})$  and  $D(\mathcal{A}_{S,Q})$  are fixed and minimum over all possible alignments.*

**Lemma 18** *Let  $\mathcal{A}$  be a standard alignment of  $\mathcal{S}$ . Then in  $\mathcal{A}_{S \cup X}$  there are  $C_3l(k+1)$  mismatches of the form  $(\sigma, \Delta)$ , where  $\sigma \in \mathcal{A}_S$  and  $\Delta \in \mathcal{A}_X$ .*

**PROOF.** Let  $x$  and  $s$  be respectively a test sequence and an edge sequence, and let us consider the alignment  $\mathcal{A}_{\{s,x\}}$ . By properties (i), (ii) and (iii) of standard alignment each  $\Delta$ 's in  $\mathcal{A}_{\{s\}}$  is opposite only to  $\Delta$ 's of  $\mathcal{A}_{\{x\}}$  in  $\mathcal{A}_{\{s,x\}}$ . Since, by construction, each edge sequence contains  $k+1$  symbols more than each test sequence, it follows that in each test sequence  $x$  there are exactly  $k+1$   $\Delta$ 's that have a mismatch with a symbol of  $s$  in  $\mathcal{A}_{\{s,x\}}$ . The claim follows immediately.

**Lemma 19** *Let  $\mathcal{A}$  be a standard alignment of  $\mathcal{S}$ . Then the sum  $D(\mathcal{A}_{X \cup T \cup Q}) + D(\mathcal{A}_{S, T \cup Q})$  is fixed over all possible standard alignments of  $\mathcal{S}$ .*

**PROOF.** By Lemma 17,  $D(\mathcal{A}_{X \cup T \cup Q})$  and  $D(\mathcal{A}_{S, Q})$  are fixed. Let  $s, t$  be respectively an edge sequence and a template sequence, and let  $\mathcal{A}$  be a standard alignment of  $\mathcal{S}$ . Since in  $t$  there is one symbol more than in  $s$  it follows that in  $\mathcal{A}_{\{s\}}$  there is one  $\Delta$  more than in  $\mathcal{A}_{\{t\}}$ . By Proposition 14, all  $b$ 's of  $t$ , except for one, have a mismatch with the symbol of  $s$  to which they are opposite. Moreover, by Proposition 14, there is an  $a$  or a  $\Delta$  inserted in  $t$  that has a mismatch with a  $b$  of  $s$ . By definition of standard alignment, there cannot be any other mismatch in  $\mathcal{A}_{\{s, t\}}$ .

By previous Lemma 19, the sum  $D(\mathcal{A}_{X \cup T \cup Q}) + D(\mathcal{A}_{S, T \cup Q})$  is fixed for every standard alignment  $\mathcal{A}$ ; *in the following we will denote such sum as  $D_{SD}$* . Moreover, by Lemma 15 it is immediate that every standard alignment  $\mathcal{A}$  and hence every optimal alignment has a value  $D(\mathcal{A}) \leq D_{SD} + U_S + U_{S, X}$ .

**Lemma 20** *Let  $\mathcal{A}$  be an optimal alignment of  $\mathcal{S}$ . Then  $\mathcal{A}$  must satisfy property (i) of a standard alignment.*

**PROOF.** Let  $\mathcal{A}_1$  be an arbitrary standard alignment and let  $\mathcal{A}$  be an optimal alignment of  $\mathcal{S}$  that does not satisfy property (i) of standard alignment. Then the following cases must be considered.

Case 1) There is a column in  $\mathcal{A}_Q$  containing some  $\Delta$ 's and some  $\sigma$ 's, for  $\sigma \in \{a, b\}$ . By Lemma 3,  $\mathcal{A}$  cannot be optimal.

Case 2) Let us assume that there is a column of  $\mathcal{A}_{T \cup Q}$  that contains at least a symbol  $\sigma$  and  $\Delta$ . Clearly, if  $\Delta$  is in  $\mathcal{A}_T$ , then a symbol  $\Delta$  must be in  $\mathcal{A}_Q$ . By case 1, since each column of  $\mathcal{A}_Q$  contains either  $a$ 's or  $b$ 's or  $\Delta$ 's, it follows that there is a mismatch  $(\Delta, \sigma)$  in  $\mathcal{A}_{T \cup Q}$ , consisting of a  $\Delta$  in  $\mathcal{A}_Q$  and a  $\sigma$  in  $\mathcal{A}_T$ , that occurs in the  $i^{\text{th}}$  column of  $\mathcal{A}$ .

Then, we can show that  $D(\mathcal{A}_{T, Q}) \geq D(\mathcal{A}_{1T, Q}) + C_2$ . In fact, let  $t$  be the sequence of  $T$  that contains the symbol  $\sigma$  in the  $i^{\text{th}}$  column and let  $q$  be an arbitrary sequence in  $Q$ . Then,  $\mathcal{A}_{\{t, q\}}$  contains at least  $n + 1$  mismatches, as  $|t| = |q|$  and  $t$  contains  $n + 2$   $b$ 's, while  $\mathcal{A}_{\{q\}}$  contains 2  $b$ 's and a  $\Delta$  not in  $\mathcal{A}_{\{t\}}$ . Clearly,  $\mathcal{A}_{1\{t, q\}}$  contains exactly  $n$  mismatches, all of value 1. Since  $|Q| = C_2$ , it follows that  $D(\mathcal{A}_{T, Q}) \geq D(\mathcal{A}_{1T, Q}) + C_2$ .

Case 3) Assume now that every column containing  $\Delta$ 's in  $\mathcal{A}_{T \cup Q}$  has only  $\Delta$ 's in  $\mathcal{A}_{T \cup Q}$ , but at least one  $a$  in  $\mathcal{A}_X$ . Let  $y$  be the sequence in  $X$  that has at least one  $a$  in such column and let  $q$  be a sequence in  $Q$ . Since  $\mathcal{A}_{\{q\}}$  contains

at least a  $\Delta$  opposite to an  $a$  in  $\mathcal{A}_{\{y,q\}}$ , while  $y$  has at least  $k+2$   $\Delta$ 's, it follows that  $\mathcal{A}_{\{y,q\}}$  contains at least  $k+3$  mismatches, of which two are of value 1, while the other ones are of value 2. By condition (i) of standard alignment, for any arbitrary sequence  $x \in X$ ,  $\mathcal{A}_{\{x,q\}}$  contains exactly  $k+2$  mismatches. It follows that  $D(\mathcal{A}_{X,Q}) \geq D(\mathcal{A}_{1X,Q}) + C_2$ .

**Case 4)** Each column containing  $\Delta$ 's in  $A_{T \cup Q}$  consists of only  $\Delta$ 's in  $\mathcal{A}_{T \cup Q \cup X}$  and has at least an  $a$  in  $\mathcal{A}_S$ . We now show that  $D(\mathcal{A}_{S,Q}) \geq D(\mathcal{A}_{1S,Q}) + C_2$ . By Lemma 16, for each sequence  $s \in S$  and  $q \in Q$ ,  $D(\mathcal{A}_{1\{s,q\}}) = 4$ .

Let  $s_1$  be a sequence in  $S$  that contains an  $a$  in a column where there are only  $\Delta$ 's in  $\mathcal{A}_{T \cup Q \cup X}$ . Then  $\mathcal{A}_{\{s_1,q\}}$  must contain at least a mismatch ( $a, \Delta$ ) besides four mismatches of value 1. Hence  $D(\mathcal{A}_{\{s_1,q\}}) \geq 5$ . It follows that  $D(\mathcal{A}_{S,Q}) \geq D(\mathcal{A}_{1S,Q}) + C_2$ .

By previous cases, Lemma 17 and Lemma 15, since  $C_2 > U_{S,T} + U_{S,X} + U_S$  the Lemma follows.

**Lemma 21** *Let  $\mathcal{A}$  be an optimal alignment of  $\mathcal{S}$ . Then  $\mathcal{A}$  must satisfy property (ii) of a standard alignment.*

**PROOF.** By Lemma 20,  $\mathcal{A}$  must satisfy property (i). Assume that  $\mathcal{A}$  satisfies property (i) and assume to the contrary that  $\mathcal{A}$  does not have property (ii). Let  $\mathcal{A}_1$  be a standard alignment. As in the proof of Lemma 20, case 4, it is easy to show that  $D(\mathcal{A}_{S,Q}) \geq D(\mathcal{A}_{1S,Q}) + C_2$ . Since  $C_2 > U_{S,T} + U_S + U_{S,X}$ , it follows that  $D(\mathcal{A}_{S,T \cup Q}) > D(\mathcal{A}_{1S,T \cup Q}) + U_S + U_{S,X}$ . By applying Lemma 17 and Lemma 15, we obtain that  $D(\mathcal{A}) > D(\mathcal{A}_1)$ , which is a contradiction.

**Lemma 22** *Let  $\mathcal{A}$  be an optimal alignment of  $\mathcal{S}$ . Then  $\mathcal{A}$  must satisfy property (iii) of a standard alignment.*

**PROOF.** By Lemma 3, there is no column of  $\mathcal{A}_X$  containing  $\Delta$ 's and  $a$ 's. Moreover, by the *SP*-score, in an optimal alignment it is more advantageous that  $\Delta$ 's of  $\mathcal{A}_X$  are opposite to  $b$ 's of  $\mathcal{A}_{T \cup Q}$ . Consequently, in the first and last column of  $\mathcal{A}_X$  there are only  $\Delta$ 's.

**Lemma 23** *Let  $\mathcal{A}$  be an optimal alignment of  $\mathcal{S}$ . Then  $\mathcal{A}$  must satisfy property (iv) of a standard alignment.*

**PROOF.** By Lemmas 20, 21 and 22,  $\mathcal{A}$  must satisfy properties (i), (ii) and (iii). Consequently, if (iv) does not hold, it means that there is a column of index  $l_1$  in  $\mathcal{A}$  containing only  $\Delta$ 's of  $X$  and only  $a$ 's of  $T \cup Q$ , and eventually  $b$ 's of  $S$ . Moreover, since for each sequence  $x$ ,  $x \in X$ ,  $|x| = 3n+2-k$ , while for

each sequence  $t \in T$ ,  $t$  contains  $n + 2$   $b$ 's, it follows that there is a column of index  $l_2$  in  $\mathcal{A}$  containing  $a$ 's of each test sequence in  $X$  and  $b$ 's of the template sequences. Then let  $\mathcal{A}_1$  be the alignment obtained from  $\mathcal{A}$  as follows: in  $\mathcal{A}_X$  substitute the  $\Delta$ 's in the column of index  $l_1$  with the  $a$ 's in the column with index  $l_2$  and vice versa.

By construction of  $\mathcal{A}_1$ ,  $D(\mathcal{A}) - D(\mathcal{A}_1)$  is equal to the sum  $D(\mathcal{A}[l_1]) - D(\mathcal{A}_1[l_1]) + D(\mathcal{A}[l_2]) - D(\mathcal{A}_1[l_2])$ . We will prove that  $D(\mathcal{A}) - D(\mathcal{A}_1) > 0$ , thus obtaining a contradiction with the assumption that  $\mathcal{A}$  is an optimal alignment of  $\mathcal{S}$ . In the following, by  $n_\sigma(l_i)$  we will denote the number of  $\sigma$  symbols in the column of index  $l_i$  of  $\mathcal{A}_S$ . By construction of the sequences, and by the SP-score it is easy to note that:

$$\begin{aligned} D(\mathcal{A}[l_1]) &= (C_1 + C_2 + n_a(l_1))n_b(l_1)d(a, b) + (C_1 + C_2 + n_a(l_1))C_3d(a, \Delta) + n_b(l_1)C_3d(\Delta, b) \\ D(\mathcal{A}[l_2]) &= (C_1 + n_b(l_2))(C_2 + C_3 + n_a(l_2))d(a, b) \\ D(\mathcal{A}_1[l_1]) &= (C_1 + C_2 + C_3 + n_a(l_1))n_b(l_1)d(a, b) \\ D(\mathcal{A}_1[l_2]) &= (C_1 + n_b(l_2))(C_2 + n_a(l_2))d(a, b) + (C_2 + n_a(l_2))C_3d(a, \Delta) + (C_1 + n_b(l_2))C_3d(\Delta, b) \end{aligned}$$

Consequently  $D(\mathcal{A}) - D(\mathcal{A}_1) = 2C_3(C_1 - (n_a(l_2) - n_a(l_1)))$ . Since  $n_a(l_2) - n_a(l_1) \leq l$ , it follows  $D(\mathcal{A}) - D(\mathcal{A}_1) \geq 2C_3(C_1 - l)$ . By posing  $C_1 > l$ , we obtain that  $D(\mathcal{A}_1) < D(\mathcal{A})$ , which contradicts the fact that  $\mathcal{A}$  is optimal.

Thus,  $\mathcal{A}$  must satisfies property (iv).

By Lemmas 20, 21, 22 and 23 it follows directly that:

**Corollary 24** *An optimal alignment of  $\mathcal{S}$  is a standard alignment.*

The result of Theorem 25, relates the value of an alignment of  $\mathcal{S}$  to the size of a node cover.

**Theorem 25** *Let  $G$  be a graph and  $\mathcal{S}$  the encoding of  $G$ . Then:*

- (1) *if  $G$  has a node cover of size  $k$ , then there is a standard alignment  $\mathcal{A}$  of  $\mathcal{S}$  such that  $D(\mathcal{A}_{S,X}) \leq 2lC_3 + 2lkC_3$ ,*
- (2) *if  $G$  has a minimum node cover of size  $k_1 > k$ , then for every standard alignment  $\mathcal{A}$  of  $\mathcal{S}$  it holds that  $D(\mathcal{A}_{S,X}) > U_S + 2lC_3 + 2lkC_3$ .*

**PROOF.** (1) Assume first that  $G$  has a node cover of size  $k$ . Let  $\mathcal{A}$  be a standard alignment of  $\mathcal{S}$ , where the sequences in  $S \cup X$  are aligned as follows:  $\mathcal{A}_S$  does not contain  $\Delta$ 's in internal columns. For each test sequence the first and last  $\Delta$ 's are respectively aligned in the first and last column of  $\mathcal{A}_S$ . Moreover, for each edge sequence  $s(i, j)$  encoding the edge  $e$ ,  $e = (v_i, v_j)$ , one of the two

$b$ 's encoding one end vertex of  $e$  is aligned in a column of  $\mathcal{A}$  containing  $\Delta$ 's of each test sequence  $x(k)$ . Such an alignment is possible since each edge has one end in the node cover, and the number of  $\Delta$ 's in each test sequence is equal to  $k + 2$ , where  $k$  is the size of a node cover. In fact, if the set  $K$  of vertices, with  $K = \{v_{i_1}, \dots, v_{i_k}\}$  is a node cover for  $G$ , then  $\mathcal{A}$  can be obtained by aligning for each  $1 \leq h \leq k$ , the  $(h + 1)$ -th  $\Delta$ 's of each test sequence with the  $(i_h + 1)$ -th  $b$  of each template sequence, and with a  $b$  of an edge sequence. In fact, every edge sequence encodes an edge  $(v_i, v_j)$  such that either  $v_i \in K$  or  $v_j \in K$ , (Fig. 2). It follows, by Proposition 14, that the total number of  $b$ 's in  $\mathcal{A}_S$  opposing  $\Delta$ 's in  $\mathcal{A}_X$  is equal to the number  $l$  of edges of the graph.

Let us determine  $D(\mathcal{A}_{S,X})$ . Let  $I$  be the set of indices of the columns of  $\mathcal{A}$  containing  $\Delta$ 's of  $X$  and let  $n_\sigma(i)$  be the number of  $\sigma$ 's in the column of  $\mathcal{A}_{S \cup X}$  of index  $i$ . Moreover, let  $r$  be the number of columns in  $\mathcal{A}$ . Then  $D(\mathcal{A}_{S,X}) = \sum_{i \in I - \{1,r\}} C_3(d(\Delta, b)n_b(i) + d(\Delta, a)(l - n_b(i)) + \sum_{i \notin I \cup \{1,r\}} C_3 d(a, b)n_b(i) + D(\mathcal{A}_{S,X}[1]) + D(\mathcal{A}_{S,X}[r]))$ . Since  $\sum_{i \in I} n_b(i) = l$ , it follows that  $D(\mathcal{A}_{S,X}) = 2klC_3 + 2lC_3$ .

(2) Now, assume that  $G$  has a minimum node cover of size  $k_1$ , with  $k_1 > k$ .

Let  $\mathcal{A}$  be an arbitrary standard alignment of  $\mathcal{S}$ . Let us compute  $D(\mathcal{A}_{S,X})$ : since  $G$  has a node cover of size  $k_1 > k$  and, by Proposition 14, for each edge sequence  $s(i, j)$  encoding the edge  $(v_i, v_j)$ , exactly one  $b$  of  $s(i, j)$  encoding an end vertex  $v_h$  is aligned with the  $(h + 1)$ -th  $b$  of each template sequence, there must be at least  $k_1$  columns of  $\mathcal{A}$  that contain  $b$ 's of the template sequences opposing one  $b$  of at least an edge sequence. By properties (iii) and (iv) of standard alignment, in  $\mathcal{A}_{X \cup T}$ , each  $\Delta$  of  $\mathcal{A}_X$  is aligned with a  $b$  of  $\mathcal{A}_T$ . Since  $\mathcal{A}_X$  contains  $k$   $\Delta$ 's internal columns, it follows that there is at least one edge sequence such that no one of the two  $b$ 's of these sequences is in a column of  $\mathcal{A}_{S \cup X}$  containing  $\Delta$ 's of  $X$ . Consequently, given  $I_1$  the set of indices of the columns of  $\mathcal{A}$  that contain  $\Delta$ 's of the test sequences,  $\sum_{i \in I_1} n_b(i) \leq l - 1$ . Clearly,  $D(\mathcal{A}_{S,X}) = C_3(\sum_{i \in I_1} (d(\Delta, a)n_a(i) + d(\Delta, b)n_b(i)) + \sum_{i \notin I_1} (d(a, \Delta)n_\Delta(i) + d(a, b)n_b(i)))$ .

By Lemma 18 the number of mismatches  $(\sigma, \Delta)$ , where  $\Delta \in \mathcal{A}_X$  is  $C_3l(k + 1)$ . Consequently, there are  $C_3l(k + 1) - C_3 \sum_{i \in I_1} n_b(i)$  mismatches  $(a, \Delta)$ . It follows that  $D(\mathcal{A}_{S,X}) = C_3(2l(k + 1) - 2 \sum_{i \in I_1} n_b(i) + \sum_{i \in I_1} n_b(i) + \sum_{i \notin I_1} n_b(i) + 2 \sum_{i \notin I} n_\Delta(i)) \geq 2lkC_3 + 2lC_3 + 2C_3$ , as by properties (i) and (ii) of standard alignment  $\sum_{i \notin I} n_\Delta(i) = 0$  and  $\sum_i n_b(i) = 2l$ . By Lemma 19, since  $C_3 > U_S$  it follows that  $D(\mathcal{A}_{S,X}) > U_S + 2lC_3 + 2lkC_3$ .

By Corollary 24 and Theorem 25, the following result is immediate.

**Corollary 26** *The graph  $G$  has a node cover of size  $k$  iff the set  $\mathcal{S}$  has an optimal alignment  $\mathcal{A}$  of value  $D(\mathcal{A})$ , with  $D(\mathcal{A}) < D_{SD} + U_S + 2lC_3 + 2lkC_3$ .*

By the previous result it follows that the construction of the set  $\mathcal{S}$  of sequences

from an instance  $G$  and  $k$  of  $NC$  is a reduction to sequence alignment.

**Theorem 27** *Multiple alignment with metric  $SP$ -score is  $NP$ -complete even over a binary alphabet.*

## 5 Conclusions

We think that the approach developed here can be generalized to prove the  $NP$ -completeness of the decision version of the alignment problem under further restrictions on the  $SP$ -score matrices. For example, we conjecture that our proof can be extended to show that the problem remains intractable also in the case of a  $SP$ -score in which the distance between distinct letters is 1, while the distance of  $\Delta$  with all other letters is 2, i.e. the metric is also an *ultrametric*.

It is not known if the multiple sequence alignment problem with a fixed metric  $SP$ -score admits a polynomial time approximation scheme. While it is interesting to understand whether the problem is MAXSNP-hard, it seems quite difficult to modify the structure of the reduction given in the paper to obtain an  $L$ -reduction [8]. At the same time even describing an approximation algorithm whose error ratio is a constant strictly less than 2 is a challenging problem.

**Acknowledgments.** We thank Tao Jiang for having pointed out the problem.

## References

- [1] V. Bafna, E.L. Lawler, and P.A. Pevzner. Approximation algorithms for multiple sequence alignment. *Theoretical Computer Science*, 182:233–244, 1997.
- [2] H. Carrillo and D. Lipman. The multiple sequence alignment in biology. *SIAM Journal of Applied Mathematics*, 48:1073–1082, 1988.
- [3] S.C. Chan, A.K.C. Wong, and D.K.T. Chiu. A survey of multiple sequence comparison methods. *Bulletin of Mathematical Biology*, 54(4):563–598, 1992.
- [4] W.M. Fitch. Letter to the editor: Commentary on the letter by Ward C. Wheeler. *Mol. Biol. Evol.*, 10(3):713–714, 1993.
- [5] M.R. Garey and D.S. Johnson. *Computers and Intractability: A Guide to the Theory of  $NP$ -Completeness*. W.H. Freeman, 1979.

- [6] D. Gusfield. Efficient methods for multiple sequence alignment with guaranteed error bounds. *Bulletin Mathematical Biology*, 55:141–154, 1993.
- [7] R.M. Karp. Mapping the genome: Some combinatorial problems arising in molecular biology. In *ACM Symp. on Theory of Computing*, pages 278–285, 1993.
- [8] C.H. Papadimitriou and M. Yannakakis. Optimization, approximation and complexity classes. *Journal of Computer and System Sciences*, 43:425–440, 1991.
- [9] L. Wang and T. Jiang. On the complexity of multiple sequence alignment. *Journal of Computational Biology*, 1(4):337–348, 1994.
- [10] H. Todd Wareham. A simplified proof of the np- and max snp-hardness of multiple sequence tree alignments. *Journal of Computational Biology*, 2(4):509–514, 1995.
- [11] W.C. Wheeler. Letter to the editor: The triangle inequality and character analysis. *Mol. Biol. Evol.*, 10(3):707–712, 1993.