Non-Approximability of Weighted Multiple Sequence Alignment

Bodo Siebert[*]

Institut für Theoretische Informatik
Med. Universität zu Lübeck
Wallstraße 40, 23560 Lübeck, Germany
siebert@tcs.mu-luebeck.de

Abstract. We consider a weighted generalization of multiple sequence alignment with sum-of-pair score. Multiple sequence alignment without weights is known to be $\mathcal{NP}$-complete and can be approximated within a constant factor, but it is unknown whether it has a polynomial time approximation scheme. Weighted multiple sequence alignment can be approximated within a factor of $O(\log^2 n)$ where n is the number of sequences.

We prove that weighted multiple sequence alignment is MAX $\mathcal{SNP}$-hard and establish a numerical lower bound on its approximability, namely $\frac{324}{323} - \epsilon$. This lower bound is obtained already for the simple binary weighted case where the weights are restricted to 0 and 1. Furthermore, we show that weighted multiple sequence alignment and its restriction to binary weights can be approximated exactly to the same degree.

1 Introduction

Multiple sequence alignment (MSA) is an important problem in computational biology. The alignment of a group of protein or nucleotide sequences yields information about the relationships between these sequences and it is also used to detect similarities (so called "homologous regions") between them. This information is applied in constructing evolutionary trees and finding coherences between the function and structure of proteins and their sequences.

Many objective functions have been suggested to measure the quality of a multiple sequence alignment. One of the most widely used is the so called sum-of-pair score (SP-score, Carrillo et al. [6]).

MSA with SP-score is known to be $\mathcal{NP}$-complete (Wang et al. [12]). For the case that the scoring function does not have to be a metric, Just has shown that MSA with SP-score is MAX $\mathcal{SNP}$-hard [9]. Akutsu et al. have investigated the multiple sequence alignment problem under several scoring functions, namely $\#LOG\#$-score and IC-score [1]. They have shown that a variant of the multiple sequence alignment problem called local multiple alignment is MAX $\mathcal{SNP}$-hard under these scoring schemes.

However, if the scoring function fulfils the triangle inequality, no lower bound for this problem is known so far. The complexity of MSA over an alphabet of fixed size with

[*] supported by DFG research grant RE 672/3

7th Ann. Int. Conf. on Computing and Combinatorics (COCOON 2001) © Springer

metric SP-scoring functions is of main interest. According to Jiang et al. the approximability of MSA with metric SP-score is an important open problem in computational biology [8].

To represent existing knowledge about the relationships of the sequences considered, a weighted variant of MSA was introduced by Wu et al. [13]. Each pair of sequences is assigned a nonnegative value reflecting their degree of relationship. This means that a pair which is assumed to be closely related will be assigned a high weight while a less related pair will be assigned a smaller weight. This generalization of MSA is called weighted MSA, or WMSA for short.

In this paper we also examine a restricted version of WMSA called binary weighted MSA (BMSA), where the weights are restricted to 0 and 1. The binary weights can be used to represent an arbitrary graph over which multiple sequence alignments can be determined. We will prove that BMSA is equivalent to WMSA with respect to their approximability. Thus, an approximation algorithm for BMSA directly yields an approximation algorithm for the general case with the same performance ratio. Moreover, we prove the MAX $\mathcal{SNP}$-hardness and a numerical lower bound for the approximability of BMSA. These results are obtained even if the sequences are of fixed length and the alphabet is of fixed size. Thus, the difficulty of multiple sequence alignment is caused by the number of sequences, not by their length.

In the next section we give a formal definition of the problems considered. The reduction from WMSA to BMSA is presented in section 3. In section 4 we prove a lower bound for the approximability of a problem called MAX-E2-neg-Lin2. This result will be used in section 5 to prove a lower bound for the approximability of BMSA.

2 Definitions and Notations

Let Σ be an alphabet and $\Sigma' := \Sigma \cup \{-\}$, where "$-$" denotes a gap symbol. $S[l]$ denotes the l-th symbol of a sequence S. Let $\mathcal{S} = \{S_1, \ldots, S_n\}$ be a family (a multiset) of sequences over Σ. An *alignment* of $\mathcal{S}$ is a family $\mathcal{A} = \{\tilde{S}_1, \ldots, \tilde{S}_n\}$ of sequences over Σ' such that all $\tilde{S}_i$ have equal length and $\tilde{S}_i$ is obtained from S_i by inserting gaps. The following is an example of an alignment of three sequences $ATTCTG$, $TTCTTTG$ and $ATTGTT$.

$$\begin{array}{c} ATTCT{-}{-}G \\ {-}TTCTTTG \\ ATTGT{-}T{-} \end{array}$$

A function $d : {\Sigma'}^2 \to \mathbb{N}$ will be called *scoring function* if it is a metric, i.e. for any $x, y, z \in \Sigma'$ we have $d(x,y) = 0$ iff $x = y$, $d(x,y) = d(y,x)$, and $d(x,z) \leq d(x,y) + d(y,z)$. We define the distance of two sequences $\tilde{S}_i$ and $\tilde{S}_j$ of length l as $D(\tilde{S}_i, \tilde{S}_j) := \sum_{k=1}^{l} d(\tilde{S}_i[k], \tilde{S}_j[k])$.

Carrillo and Lipman introduced a scoring scheme for alignments called *sum-of-pair score* (SP-score, [6]). The SP-score of an alignment $\mathcal{A} = \{\tilde{S}_1, \ldots, \tilde{S}_n\}$ is defined by $D(\mathcal{A}) := \sum_{1 \leq i < j \leq n} D(\tilde{S}_i, \tilde{S}_j)$. Multiple sequence alignment (MSA) is the problem of finding an alignment with minimum SP-score.

Wu et al. generalized MSA to weighted sum-of-pair score [13]. The weights are given by $W := (w_{S_i,S_j})_{S_i,S_j \in \mathcal{S}}$, a symmetric matrix of nonnegative integers. Then the weighted SP-score of an alignment $\mathcal{A}$ is $D_W(\mathcal{A}) := \sum_{1\leq i<j\leq n} w_{S_i,S_j} \cdot D(\tilde{S}_i, \tilde{S}_j)$. This generalization is called *weighted multiple sequence alignment* (WMSA). The aim is to find an alignment with minimum weighted SP-score.

An instance of WMSA is a 4-tuple $(\Sigma, \mathcal{S}, d, W)$. We consider the case of a fixed alphabet Σ and a fixed scoring function d. Thus, a problem instance of WMSA is given by a pair $(\mathcal{S}, W)$. It is easy to see that any lower bound for this case also holds if we allow arbitrary scoring functions and alphabets.

A special case of WMSA is *binary weighted MSA* (BMSA), where the weights are restricted to 0 and 1.

It has been shown that MSA with SP-score is $\mathcal{NP}$-complete [12]. For an arbitrary fixed constant r, MSA can be approximated in polynomial time within a factor of $2-\frac{r}{n}$, where $n \geq r$ is the number of sequences [4]. It is unknown whether MSA admits a polynomial time approximation scheme (PTAS, see e.g. Ausiello et al. [3]). WMSA with arbitrary weights can be approximated within a factor of $O(\log^2 n)$ [13]. Using a technique of Bartal [5] one can obtain a randomized $O(\log n \cdot \operatorname{llog} n)$ approximation.

Papadimitriou et al. introduced a class of optimization problems called MAX $\mathcal{SNP}$ [10]. They showed that there exist problems which are MAX $\mathcal{SNP}$-complete with respect to L-reductions. In the following, $\operatorname{opt}(I)$ denotes the optimal score of an instance I of an optimization problem. For example, $\operatorname{opt}(\mathcal{S})$ denotes the score of an optimal (weighted) alignment of $\mathcal{S}$.

Definition 1. *Let Π and Π' be two optimization problems. Then Π L-reduces to Π' if there exist polynomial time computable functions f_1, f_2 and constants $\gamma_1, \gamma_2 > 0$ such that for each instance I of Π:*

1. *Function f_1 produces an instance $I' = f_1(I)$ of Π' such that $\operatorname{opt}(I') \leq \gamma_1 \cdot \operatorname{opt}(I)$.*
2. *Given a solution S' of I' with cost $c'(S')$, function f_2 produces a solution $S = f_2(I, S')$ of I with cost $c(S)$ such that $|c(S) - \operatorname{opt}(I)| \leq \gamma_2 \cdot |c'(S') - \operatorname{opt}(I')|$.*

No MAX $\mathcal{SNP}$-hard problem has a PTAS, unless $\mathcal{NP} = \mathcal{P}$ (Arora et al. [2]).

3 Reduction from WMSA to BMSA

Let $\mathcal{S} = \{S_1, \ldots, S_n\}$ be a family of sequences over Σ and $W = (w_{S_i,S_j})$ be a weight matrix. Let l be the maximal length of the sequences in $\mathcal{S}$ and $d_{\max}$ be the maximum of the scoring function d. We assume that the weights and the scoring function are unary coded. This does not seem to be a restriction because in practice the weights are very small and the scoring function is fixed.

We construct a family of sequences $\mathcal{S}'$ as an instance of BMSA as follows. Let $K := 2 \cdot d_{\max} \cdot l$. For a sequence $S_j \in \mathcal{S}$ generate K copies $T_j^k \in \mathcal{S}'$ $(1 \leq k \leq K)$ of this sequence. Furthermore, for each $1 \leq i \leq n$ construct w_{S_i,S_j} copies $S_j^{i,\mu} \in \mathcal{S}'$ $(1 \leq \mu \leq w_{S_i,S_j})$ of S_j. The weight matrix $W' = (w'_{I,J})_{I,J\in\mathcal{S}'}$ is given by

$$w'_{I,J} := \begin{cases} 1 & \text{if } I \equiv S_j^{i,\mu} \text{ and } J \equiv S_i^{j,\mu}\ , \\ 1 & \text{if } I \equiv S_j^{i,\mu} \text{ and } J \equiv T_j^k \text{ or vice versa}\ , \\ 0 & \text{otherwise}\ , \end{cases}$$

where $A \equiv B$ means that A and B are not only equal but denote the same sequence.

Since the weights and the scoring function are unary coded, the input size N of the instance of WMSA fulfils the bound $N \in \Omega\big(n \cdot l + \sum_{i,j=1}^{n} w_{S_i,S_j}\big)$. On the other hand, the input size N' of the constructed instance of BMSA satisfies

$$N' \in O\big(\underbrace{n \cdot K \cdot l}_{T_\cdot^\cdot} + \underbrace{l \cdot \textstyle\sum_{i,j=1}^{n} w_{S_i,S_j}}_{S_\cdot^{\cdot,\cdot}} + \underbrace{(n \cdot K + \textstyle\sum_{i,j=1}^{n} w_{S_i,S_j})^2}_{W'}\big) \, .$$

Note that N' is polynomially bounded by N.

Lemma 1. *If $\mathcal{S}$ has an alignment $\mathcal{A}$ with weighted score $D_W(\mathcal{A})$ then $\mathcal{S}'$ has an alignment $\mathcal{A}'$ with score $D_{W'}(\mathcal{A}') = D_W(\mathcal{A})$.*

Proof. Let $\mathcal{A} = \{\tilde{S}_1, \ldots, \tilde{S}_n\}$ be an alignment of $\mathcal{S}$ with weighted score D. We obtain an alignment $\mathcal{A}' = \{\tilde{A} | A \in \mathcal{S}'\}$ of $\mathcal{S}'$ by setting $\tilde{T}_j^k = \tilde{S}_j$ and $\tilde{S}_j^{i,\mu} = \tilde{S}_j$ for all j, k, i, μ. The score of $\mathcal{A}'$ with respect to the weight matrix W' is

$$D_{W'}(\mathcal{A}') = \sum_{i,j=1}^{n} \sum_{\mu=1}^{w_{S_i,S_j}} \sum_{k=1}^{K} \underbrace{D(\tilde{S}_j^{i,\mu}, \tilde{T}_j^k)}_{=0} + \sum_{1 \le i < j \le n} \sum_{\mu=1}^{w_{S_i,S_j}} \underbrace{D(\tilde{S}_i^{j,\mu}, \tilde{S}_j^{i,\mu})}_{=D(\tilde{S}_i,\tilde{S}_j)} = D_W(\mathcal{A}) \, .$$

□

Lemma 2. *Given an alignment $\mathcal{A}'$ of $\mathcal{S}'$ with weighted score $D_{W'}(\mathcal{A}')$ we can construct an alignment $\mathcal{A}$ of $\mathcal{S}$ with less or equal score in polynomial time.*

Proof. Let $\mathcal{A}' = \{\tilde{A} | A \in \mathcal{S}'\}$ be an arbitrary alignment of $\mathcal{S}'$ with score $D_{W'}(\mathcal{A}')$. The copies of a sequence $S_j \in \mathcal{S}$ will be called *consistent* if there exists a sequence B_j with $\tilde{T}_j^k = B_j$ and $\tilde{S}_j^{i,\mu} = B_j$ for all k, i, μ. The sequence B_j is called *block*.

We consider the case that for some j_0 the copies of S_{j_0} are not consistent and distinguish two cases. First, if not all $\tilde{T}_{j_0}^k$ are equal, let $D_k := \sum_{i=1}^{n} \sum_{\mu=1}^{w_{S_i,S_j}} D(\tilde{T}_{j_0}^k, \tilde{S}_{j_0}^{i,\mu})$ be the score of $\tilde{T}_{j_0}^k$ with the sequences $\tilde{S}_{j_0}^{i,\cdot}$. Choose k_0 such that D_{k_0} is minimal among all D_k and set $\tilde{T}_{j_0}^k = \tilde{T}_{j_0}^{k_0}$ for all $k \neq k_0$. This way we obtain a new alignment with less or equal score.

Now we consider the case that there exists a B_{j_0} such that $\tilde{T}_{j_0}^k = B_{j_0}$ for all k. Then there exists a sequence $\tilde{S}_{j_0}^{i_0,\mu_0} \neq B_{j_0}$. This sequence yields at least score K with the sequences $\tilde{T}_{j_0}^{\cdot}$, because it yields a score of at least 1 with every $\tilde{T}_{j_0}^k$. Set $\tilde{S}_{j_0}^{i_0,\mu_0} = B_{j_0}$. Then $\tilde{S}_{j_0}^{i_0,\mu_0}$ yields score 0 with any $\tilde{T}_{j_0}^k$ and at most score K with $\tilde{S}_{i_0}^{j_0,\mu_0}$. Thus, the new alignment has less or equal score.

By these modifications we iteratively obtain a new alignment of $\mathcal{S}'$ such that for any $j \in \{1, \ldots, n\}$ the copies of S_j are consistent with block B_j. The blocks of $\mathcal{S}'$ induce an alignment $\mathcal{A} = \{B_1, \ldots, B_n\}$ of $\mathcal{S}$ with score $D_W(\mathcal{A}) = \sum_{1 \le i < j \le n} w_{S_i,S_j} \cdot D(B_i, B_j) \le D_{W'}(\mathcal{A}')$. □

With these results we have shown that a λ-approximation for BMSA can be used as a λ-approximation for WMSA. Thus, the following theorem holds.

Theorem 1. *If* BMSA *can be approximated within a constant factor λ in polynomial time, then* WMSA *can also be approximated within λ in polynomial time.* □

4 The Non-Approximability of MAX-E2-neg-Lin2

We consider the multiplicative group $\{1, -1\}$. Let $\mathcal{G} = \{G_1, \dots, G_t\}$ be a multiset of linear equations over the variables $U = \{x_1, \dots, x_r\}$, $G_i \widehat{=} x_{\alpha_{i,1}} \cdot \ldots \cdot x_{\alpha_{i,k}} = a_i$, $k \geq 2$, $\alpha_{i,q} \in \{1, \dots, r\}$, and $a_i \in \{1, -1\}$ is a constant. MAX-Ek-Lin2 is the optimization problem of finding the maximum number of simultaneously satisfiable equations. A restriction of MAX-Ek-Lin2 is MAX-Ek-neg-Lin2, where $a_i = -1$ for all $1 \leq i \leq t$.

MAX-E2-neg-Lin2 is exactly the problem MAX-Cut (see e.g. [3]) where the equations correspond to the edges, the variables correspond to the nodes, and multiple edges are allowed. Therefore, MAX-E2-neg-Lin2 is MAX $\mathcal{SNP}$-complete [10]. We use MAX-E2-neg-Lin2 here due to the simpler notation.

An instance of MAX-Ek-Lin2 or MAX-Ek-neg-Lin2 consisting of t equations will be called *η-satisfiable* iff $\eta \cdot t$ is the maximum number of simultaneously satisfiable equations. Håstad proved in [7] that it is $\mathcal{NP}$-hard to distinguish $\big(1 - \epsilon\big)$-satisfiable and $\big(\frac{1}{2} + \epsilon\big)$-satisfiable instances of MAX-E3-Lin2 for any $\epsilon > 0$.

Instead of the known lower bound for the approximability of MAX-Cut (Håstad [7] and Trevisan et al. [11]) we will construct a reduction from MAX-E3-Lin2 to MAX-E2-neg-Lin2 to prove that it is $\mathcal{NP}$-hard to distinguish $\big(\frac{18}{22} - \epsilon\big)$- and $\big(\frac{17}{22} + \epsilon\big)$-satisfiable instances of MAX-E2-neg-Lin2 for any $\epsilon > 0$; the gadget used by Trevisan et al. [11] does not yield such a gap directly. This result will be used in section 5 to establish the lower bound for the approximability of BMSA.

We will now reduce MAX-E3-Lin2 to MAX-E2-neg-Lin2. Let $\mathcal{G} = \{G_1, \dots, G_t\}$ be a multiset of equations over variables U, $G_i \widehat{=} x_{\alpha_{i,1}} \cdot x_{\alpha_{i,2}} \cdot x_{\alpha_{i,3}} = a_i$.

We construct an instance $\mathcal{G}'$ of MAX-E2-neg-Lin2 with $22 \cdot t$ equations and $4 \cdot t + 2 \cdot r + 2$ variables. The reduction is similar to the reduction from MAX-E3-Lin2 to MAX-E2-Lin2 in [7]. The set of variables U' is given by

$$U' = \{x_j^+, x_j^- | 1 \leq j \leq r\} \cup \{z^+, z^-\} \cup \{p_{i,1}, p_{i,2}, p_{i,3}, p_{i,z} | 1 \leq i \leq t\} .$$

Note that if an assignment satisfies an equation of an instance of MAX-E2-neg-Lin2, then the negated assignment also satisfies the equation. So without loss of generality we assume that in any case $z^+ = 1$.

We interpret $x_j^+ = x_j$. We call an assignment *consistent for* x_j if $x_j^+ \neq x_j^-$ and therefore $x_j^+ = x_j = (-x_j^-)$. An assignment that is consistent for every x_j and where $z^+ \neq z^-$ is called *consistent*.

For an equation $G_i \widehat{=} x_{\alpha_{i,1}} \cdot x_{\alpha_{i,2}} \cdot x_{\alpha_{i,3}} = a_i$ we construct the twelve equations

$$\begin{aligned}
x^+_{\alpha_{i,q}} \cdot p_{i,q'} &= -1 \quad \text{for } q, q' = 1, 2, 3 \text{ and } q \neq q' \ , \\
x^+_{\alpha_{i,q}} \cdot p_{i,z} &= -1 \quad \text{for } q = 1, 2, 3 \ , \\
x^-_{\alpha_{i,q}} \cdot p_{i,q} &= -1 \quad \text{for } q = 1, 2, 3 \ .
\end{aligned}$$

We add either the four equations $z^+ \cdot p_{i,q} = -1$ ($q = 1, 2, 3$) and $z^- \cdot p_{i,z} = -1$ if $a_i = 1$ or the four equations $z^- \cdot p_{i,q} = -1$ ($q = 1, 2, 3$) and $z^+ \cdot p_{i,z} = -1$ if $a_i = -1$. For every equation in $\mathcal{G}$ we construct the three equations $x^+_{\alpha_{i,q}} \cdot x^-_{\alpha_{i,q}} = -1$ ($q = 1, 2, 3$). Finally, we add the equation $z^+ \cdot z^- = -1$ three times. Note that $\mathcal{G}'$ contains $3 \cdot t$ times the equation $z^+ \cdot z^- = -1$. Let n_j be the number of occurrences of the variable x_j in $\mathcal{G}$. Then $\mathcal{G}'$ contains n_j times the equation $x_j^+ \cdot x_j^- = -1$.

For every equation $G_i \in \mathcal{G}$ we have constructed 22 equations for $\mathcal{G}'$. These 22 equations are called the *representation of* G_i.

Lemma 3. *Let an arbitrary assignment for* U *be given. Assign* $z^- = -1$ *and* $x_j^+ = x_j$, $x_j^- = (-x_j)$ *for* $j = 1, \ldots, r$. *Then for any* $i \in \{1, \ldots, t\}$ *there exists an assignment for* $p_{i,1}$, $p_{i,2}$, $p_{i,3}$, *and* $p_{i,z}$ *such that 18 equations of the representation of* G_i *are satisfied if* G_i *is satisfied by the given assignment and 16 equations of the representation are satisfied if* G_i *is not satisfied.*

It is not possible to satisfy more than 18 equations of the representation if G_i *is satisfied by the assignment and to satisfy more than 16 equations if* G_i *is not satisfied by the assignment.*

Proof. The lemma can be proved by testing all possible assignments. □

If an assignment for U satisfies g of the t equations of $\mathcal{G}$, then the corresponding consistent assignment for U' satisfies $16 \cdot t + 2 \cdot g$ equations of $\mathcal{G}'$. This assignment can be found efficiently by adjusting the assignment for $p_{i,1}$, $p_{i,2}$, $p_{i,3}$, and $p_{i,z}$. On the other hand, a consistent assignment for U' that satisfies $16 \cdot t + 2 \cdot g$ equations of $\mathcal{G}'$ yields an assignment for U that satisfies g equations of $\mathcal{G}$.

Lemma 4. *Given an arbitrary assignment for* U' *that satisfies* $16 \cdot t + 2 \cdot g$ *equations of* $\mathcal{G}'$*, a consistent assignment that satisfies at least this amount of equations of* $\mathcal{G}'$ *can be computed in polynomial time.*

Proof. First assume that $z^+ = z^-$ in the given assignment. Then the $3 \cdot t$ equations $z^+ \cdot z^- = -1$ are not satisfied by the assignment. Let $z^- = (-z^+)$. Then these $3 \cdot t$ equations will be satisfied. On the other hand, z^- occurs in only $3 \cdot t$ other equations. Thus, at most $3 \cdot t$ equations are no longer satisfied. Altogether the number of satisfied equations is not decreased by this modification.

If there exists a j with $x_j^+ = x_j^-$, then there are n_j equations $x_j^+ \cdot x_j^- = -1$ that are not satisfied by the assignment. Let $x_j^- = (-x_j^+)$. Then the n_j equations $x_j^+ \cdot x_j^- = -1$ are satisfied by the modified assignment. On the other hand x_j^- occurs in only n_j other equations. Thus, at most n_j equations are no longer satisfied. The number of satisfied equations is thus not decreased by this modification.

This way we iteratively obtain a consistent assignment. Obviously, the modifications can be computed in polynomial time. □

Now we can prove the following theorem used in section 5.

Theorem 2. *For any* $\epsilon > 0$ *it is* $\mathcal{NP}$*-hard to distinguish* $\left(\frac{18}{22} - \epsilon\right)$*- and* $\left(\frac{17}{22} + \epsilon\right)$*-satisfiable instances of* MAX-E2-neg-Lin2.

Proof. An instance of MAX-E3-Lin2 is η-satisfiable iff the corresponding instance of MAX-E2-neg-Lin2 is $\left(\frac{16+2\cdot\eta}{22}\right)$-satisfiable. According to Håstad [7] it is $\mathcal{NP}$-hard to distinguish $\left(1-\xi\right)$- and $\left(\frac{1}{2}+\xi\right)$-satisfiable instances of MAX-E3-Lin2 for any $\xi > 0$. Thus, it is $\mathcal{NP}$-hard to distinguish $\left(\frac{16+2\cdot(1-\xi)}{22}\right)$- and $\left(\frac{16+2\cdot(\frac{1}{2}+\xi)}{22}\right)$-satisfiable instances of MAX-E2-neg-Lin2. Choosing $\xi = 11 \cdot \epsilon$ completes the proof. □

Since MAX-Cut and MAX-E2-neg-Lin2 are exactly the same problem, we obtain the same approximability gap for MAX-Cut.

Corollary 1. *For any $\epsilon > 0$ it is $\mathcal{NP}$-hard to decide whether the maximum cut of an instance $G = (V, E)$ (where multiple edges are allowed) of* MAX-Cut *consists of at most $\left(\frac{17}{22} + \epsilon\right) \cdot |E|$ or at least $\left(\frac{18}{22} - \epsilon\right) \cdot |E|$ edges.* □

5 The Non-Approximability of BMSA

In this section we reduce MAX-E2-neg-Lin2 to BMSA. Let $\mathcal{G} = \{G_1, \ldots, G_t\}$ be an instance of MAX-E2-neg-Lin2 over a set of variables $U = \{x_1, \ldots, x_r\}$, $G_i \widehat{=} x_{\alpha_{i,1}} \cdot x_{\alpha_{i,2}} = -1$, $\alpha_{i,q} \in \{1, \ldots, r\}$. We construct a family of sequences $\mathcal{S} = \{Z\} \cup \{X_j | j = 1, \ldots, r\} \cup \{Y_{i,1}, Y_{i,2} | i = 1, \ldots, t\}$ over the alphabet $\Sigma = \{\bullet, \circ, \times\}$. Let $Z :=$ ○○○○○○○○ be a sequence of length 8. Z will be used as a control sequence. For $j \in \{1, \ldots, r\}$ let $X_j :=$ ●○○○○○○○● be a sequence of length 9 that represents the variable $x_j \in U$. For each $i \in \{1, \ldots, t\}$ create two sequences $Y_{i,1} :=$ ●○○×○×○○● and $Y_{i,2} :=$ ●○○○×○○○●, each of length 9. $Y_{i,q}$ represents the variable $x_{\alpha_{i,q}}$ in G_i.

The scoring function is given in the following table. Note that it is a metric.

	-	●	○	×
-	0	1	2	5
●	1	0	1	4
○	2	1	0	3
×	5	4	3	0

The weight matrix $W = (w_{I,J})_{I,J \in \mathcal{S}}$ is given by

$$w_{I,J} := \begin{cases} 1 & \text{if } I \equiv Y_{i,q} \text{ and } J \equiv Y_{i,q'} \text{ ,} \\ 1 & \text{if } I \equiv Z \text{ and } J \equiv Y_{i,q} \text{ or vice versa ,} \\ 1 & \text{if } I \equiv Y_{i,q} \text{ and } J \equiv X_{\alpha_{i,q}} \text{ or vice versa ,} \\ 0 & \text{otherwise .} \end{cases}$$

The set $\mathcal{S}_i = \{Y_{i,1}, Y_{i,2}, X_{\alpha_{i,1}}, X_{\alpha_{i,2}}\}$ will be called the *representation of G_i*. Note that in general a sequence X_j occurs in more than one representation.

Let $\mathcal{A} = \{\tilde{S} | S \in \mathcal{S}\}$ be an alignment of $\mathcal{S}$. Then $D_i(\mathcal{A})$ denotes the score of the equation G_i, $D_i(\mathcal{A}) = D(\tilde{Y}_{i,1}, \tilde{Y}_{i,2}) + D(\tilde{Y}_{i,1}, \tilde{X}_{\alpha_{i,1}}) + D(\tilde{Y}_{i,2}, \tilde{X}_{\alpha_{i,2}}) + D(\tilde{Y}_{i,1}, \tilde{Z}) + D(\tilde{Y}_{i,2}, \tilde{Z})$. By the construction of the weight matrix, $D_W(\mathcal{A}) = \sum_{i=1}^{t} D_i(\mathcal{A})$ holds.

Definition 2. *An alignment $\mathcal{A} = \{\tilde{S} | S \in \mathcal{S}\}$ of $\mathcal{S}$ will be called* variable-consistent *with respect to an assignment for U if, after eliminating all columns consisting solely of gaps (which do not affect the score), the following holds for all j, i, and q:*

1. $\tilde{Z} = -Z-$
2. $\tilde{X}_j = \begin{cases} X_j- & \textit{if } x_j = -1 \\ -X_j & \textit{if } x_j = 1 \end{cases}$
3. $\tilde{Y}_{i,q} = \begin{cases} Y_{i,q}- & \textit{if } x_{\alpha_{i,q}} = -1 \\ -Y_{i,q} & \textit{if } x_{\alpha_{i,q}} = 1 \end{cases}$

The following lemma follows immediately from this definition.

Lemma 5. *An alignment is variable-consistent iff for all $i = 1, \ldots, t$ and $q = 1, 2$ the following properties hold:*

A. Either $Y_{i,q}[1]$ or $Y_{i,q}[9]$ matches a gap in Z. No other character of Z or $Y_{i,q}$ matches a gap in the other sequence.
B. No character in either of the two sequences $Y_{i,q}$, $X_{\alpha_{i,q}}$ matches a gap in the other sequence. □

These properties are referred to as property A and B. The following is an example of a variable-consistent alignment representing the equation $G_i \hat{=} x_1 \cdot x_2 = -1$ which is satisfied by $x_1 = -1$ and $x_2 = 1$.

$$
\begin{array}{lcllllllllll}
\tilde{Y}_{i,1} & = & \bullet & \circ & \circ & \times & \circ & \times & \circ & \circ & \bullet & - \\
\tilde{Y}_{i,2} & = & - & \bullet & \circ & \circ & \circ & \times & \circ & \circ & \circ & \bullet \\
\tilde{X}_1 & = & \bullet & \circ & \circ & \circ & \circ & \circ & \circ & \circ & \bullet & - \\
\tilde{X}_2 & = & - & \bullet & \circ & \circ & \circ & \circ & \circ & \circ & \circ & \bullet \\
\tilde{Z} & = & - & \circ & \circ & \circ & \circ & \circ & \circ & \circ & \circ & -
\end{array}
$$

Note the functional region of a pair $Y_{i,1}, Y_{i,2}$ given by the triples $\times\circ\times$ and $\circ\times\circ$. If $Y_{i,1}$ and $Y_{i,2}$ represent the same value, the functional region yields a weighted score of 9. Otherwise, it yields a weighted score of 3. If an alignment $\mathcal{A}$ is variable-consistent, we have $D_i(\mathcal{A}) = 29$ if G_i is satisfied by the represented assignment and $D_i(\mathcal{A}) = 31$ otherwise.

The next two lemmas have similar proofs. Thus, we only give a proof of the first.

Lemma 6. *Alignments of the pairs $\{Y_{i,1}, Z\}$ and $\{Y_{i,2}, Z\}$ yield scores of 8 and 5, respectively, if they fulfil property A. Violating property A yields scores of at least 10 and 7, respectively.*

Proof. An alignment of $\{Y_{i,1}, Z\}$ that fulfils property A yields score 8.

Let us consider an alignment of $\{Y_{i,1}, Z\}$ that does not fulfil property A. Then at least one of the characters $Y_{i,1}[2], \ldots, Y_{i,1}[8], Z[1], \ldots, Z[8]$ matches a gap in the other sequence.

We distinguish two cases. If there is an "$\times$" in $Y_{i,1}$ matching a gap in Z, then the alignment yields a score of 5 for this "$\times$" plus 3 for the other "$\times$" plus 1 for each "$\bullet$". So altogether it yields a score of at least 10.

On the other hand consider the case that no "$\times$" in $Y_{i,1}$ matches a gap in Z. Then there is a "$\circ$" in $Y_{i,1}$ or Z matching a gap in the other sequence. So the alignment yields a score of 3 for each "$\times$" plus 1 for each "$\bullet$" plus 2 for the "$\circ$" matching a gap. So the alignment again yields a score of at least 10.

The statement about $Y_{i,2}$ and Z can be proved in a similar fashion. □

Lemma 7. *Alignments of the pairs $\{Y_{i,1}, X_{\alpha_{i,1}}\}$ and $\{Y_{i,2}, X_{\alpha_{i,2}}\}$ yield scores of 6 and 3, respectively, if they fulfil property B. Violating property B yields scores of at least 8 and 5, respectively.* □

With the fact that an optimal alignment of a pair $\{Y_{i,1}, Y_{i,2}\}$ has score 7 we can prove the following.

Lemma 8. *Given an arbitrary alignment with score $31 \cdot t - 2 \cdot g$ we can construct a variable-consistent alignment with less or equal score in polynomial time.*

Proof. Let $\mathcal{A}$ be an arbitrary alignment with $D_W(\mathcal{A}) = 31 \cdot t - 2 \cdot g$.

Let I be the set of all i such that $Y_{i,1}$ and $Y_{i,2}$ fulfil properties A and B. This implies an assignment for the variables $U_I = \{x_j \in U | \exists i \in I : X_j \in \mathcal{S}_i\}$. Let $\overline{I} = \{1, \ldots, t\} \setminus I$. Because in every set $\mathcal{S}_i$ for $i \in \overline{I}$ there exists a sequence $Y_{i,q}$ that violates property A or B, we have $D_i(\mathcal{A}) \geq 31$ for each $i \in \overline{I}$ due to Lemmas 6 and 7.

For $i \in \overline{I}$ if $x_{\alpha_{i,q}} \in U_I$ ($q \in \{1, 2\}$), we realign $Y_{i,q}$ with respect to $x_{\alpha_{i,q}}$. Then we assign an arbitrary value to the variables in $U \setminus U_I$ and realign the corresponding $Y_{i,q}$ and X_j.

By these modifications we obtain an alignment $\mathcal{A}'$. Then $D_i(\mathcal{A}') = D_i(\mathcal{A})$ for $i \in I$ and $D_i(\mathcal{A}') \leq 31 \leq D_i(\mathcal{A})$ otherwise. Thus, $D_W(\mathcal{A}') \leq D_W(\mathcal{A})$. $\mathcal{A}'$ is variable-consistent due to its construction and can be computed in polynomial time. □

The alignment obtained yields an assignment that satisfies at least g equations of $\mathcal{G}$.

Theorem 3. BMSA *is* MAX $\mathcal{SNP}$*-hard.*

Proof. We reduce MAX-E2-neg-Lin2 to BMSA. f_1 is given by the construction of $\mathcal{S}$ from a family $\mathcal{G}$ of t equations. One can see that $\operatorname{opt}(\mathcal{S}) \leq 31 \cdot t$.

An equation of $\mathcal{G}$ will be satisfied by 2 of the 4 possible assignments of its variables. Therefore, for every multiset $\mathcal{G}$ of t equations an assignment exists that satisfies at least $\frac{1}{2} \cdot t$ equations. Then for $\gamma_1 = 62$ we have $\operatorname{opt}(\mathcal{S}) \leq \gamma_1 \cdot \operatorname{opt}(\mathcal{G})$.

Given an alignment of $\mathcal{S}$ with score $31 \cdot t - 2 \cdot g'$ for some g' we can find an assignment satisfying $g \geq g'$ equations of $\mathcal{G}$ due to Lemma 8. Let $\gamma_2 = \frac{1}{2}$, then $|g - \operatorname{opt}(\mathcal{G})| \leq \gamma_2 \cdot |(31 \cdot t - 2 \cdot g') - \operatorname{opt}(\mathcal{S})|$ holds. □

Theorem 4. BMSA *has no polynomial time approximation with performance ratio $\frac{324}{323} - \epsilon$ for any $\epsilon > 0$, unless $\mathcal{NP} = \mathcal{P}$.*

Proof. An instance of MAX-E2-neg-Lin2 consisting of t equations is η-satisfiable iff the corresponding instance of BMSA has an alignment with score $(31 - 2 \cdot \eta) \cdot t$.

The optimal alignment of a BMSA instance corresponding to a $\left(\frac{18}{22} - \xi\right)$-satisfiable instance of MAX-E2-neg-Lin2 has score $\left(31 - 2 \cdot \left(\frac{18}{22} - \xi\right)\right) \cdot t = \frac{323 + 22 \cdot \xi}{11} \cdot t$. Using the $\left(\frac{324}{323} - \epsilon\right)$-approximation algorithm for BMSA we are able to find an alignment with score at most $\left(\frac{324}{323} - \epsilon\right) \cdot \frac{323 + 22 \cdot \xi}{11} \cdot t =: K_1$.

The optimal alignment of a BMSA instance corresponding to a $\left(\frac{17}{22} + \xi\right)$-satisfiable instance of MAX-E2-neg-Lin2 has score $\left(31 - 2 \cdot \left(\frac{17}{22} + \xi\right)\right) \cdot t =: K_2$. We have $K_1 < K_2$ iff $\xi < \frac{1}{22} \cdot \frac{323^2 \cdot \epsilon}{647 - 323 \cdot \epsilon}$. Choose ξ with $0 < \xi < \frac{1}{22} \cdot \frac{323^2 \cdot \epsilon}{647 - 323 \cdot \epsilon}$. Then the $\left(\frac{324}{323} - \epsilon\right)$-approximation for BMSA can be used to distinguish $\left(\frac{18}{22} - \xi\right)$- and $\left(\frac{17}{22} + \xi\right)$-satisfiable instances of MAX-E2-neg-Lin2. This would imply $\mathcal{NP} = \mathcal{P}$ due to Theorem 2. □

Since WMSA is a generalization of BMSA it is also MAX $\mathcal{SNP}$-hard and we obtain the same non-approximability result.

6 Conclusions

We have shown MAX $\mathcal{SNP}$-hardness and proved a numerical lower bound for the approximability of weighted multiple sequence alignment (WMSA). These results hold even if we restrict the problem to binary weights (BMSA). Furthermore, BMSA and WMSA are equivalent with respect to their approximability. But the distance to the best known upper bound is huge. An obvious goal is to reduce this gap.

Finally, we would like to know how well the unweighted version of the multiple sequence alignment problem with metric SP-score can be approximated.

Acknowledgements

I wish to thank Martin Böhme, Andreas Jakoby, and Rüdiger Reischuk for valuable discussions and comments on this paper.

References

1. T. Akutsu, H. Arimura, and S. Shimozono. On approximation algorithms for local multiple alignment. In *Proc. 4th Ann. ACM Int. Conf. on Comput. Mol. Biol.*, pages 1–7, 2000.
2. S. Arora, C. Lund, R. Motwani, M. Sudan, and M. Szegedy. Proof verification and the hardness of approximation problems. *J. ACM*, 45(3):501–555, 1998.
3. G. Ausiello, P. Crescenzi, G. Gambosi, V. Kann, A. Marchetti-Spaccamela, and M. Protasi. *Complexity and Approximation*. Springer-Verlag, 1999.
4. V. Bafna, E. L. Lawler, and P. A. Pevzner. Approximation algorithms for multiple sequence alignment. *Theor. Comp. Sc.*, 182(1–2):233–244, 1997.
5. Yair Bartal. On approximating arbitrary metrics by tree metrics. In *Proc. 30th Ann. Symp. Theor. Comput.*, pages 161–168. ACM, 1998.
6. H. Carrillo and D. Lipman. The multiple sequence alignment problem in biology. *SIAM J. Appl. Math.*, 48:1073–1082, 1988.
7. J. Håstad. Some optimal inapproximability results. In *Proc. 29th Ann. Symp. Theor. Comput.*, pages 1–10. ACM, 1997.
8. T. Jiang, P. Kearney, and M. Li. Some open problems in computational molecular biology. *SIGACT News*, 30(3):43–49, 1999.
9. W. Just. Computational complexity of multiple sequence alignment with SP-score. Technical report, Ohio University, 1999.
10. C. H. Papadimitriou and M. Yannakakis. Optimization, approximation, and complexity classes. *J. Comp. Sys. Sc.*, 43(3):425–440, 1991.
11. L. Trevisan, G. B. Sorkin, M. Sudan, and D. P. Williamson. Gadgets, approximation, and linear programming. *SIAM J. Comput.*, 29(6):2074–2097, 2000.
12. L. Wang and T. Jiang. On the complexity of multiple sequence alignment. *J. Comp. Biol.*, 1(4):337–348, 1994.
13. B. Y. Wu, G. Lancia, V. Bafna, K. Chao, R. Ravi, and C. Y. Tang. A polynomial-time approximation scheme for minimum routing cost spanning trees. *SIAM J. Comput.*, 29(3):761–778, 1999.