# Complexity of the Multiple Sequence Alignment Problem

Winfried Just

Department Mathematics,
Ohio University

April 6, 2005

# Scoring matrices

Let Σ be an alphabet (e.g., $\{A, C, G, T\}$), and let $\Delta \notin \Sigma$ denote the space symbol. A *scoring scheme* is a symmetric *scoring function* $d_M : (\Sigma \cup \Delta) \times (\Sigma \cup \Delta) \mapsto \mathbb{N}$ together with specifications on how to handle gaps. A scoring function $d_M$ can be conveniently represented by a *scoring matrix* $M$. The cost of a pair of symbols $s_1, s_2$ under the scoring matrix $M$ is $d_M(s_1, s_2)$.

We will not in general assume that the scoring function is a metric; if we do, then this assumption will be explicitly stated.

# Gaps

A *gap* is a string of the form $\Delta^i$. The cost of an alignment of two sequences is obtained summing the cost of all columns and the costs of all gaps, also called *gap penalties*. Most scoring schemes used in practice are *affine*, i.e., each gap penalty is the sum of a fixed *gap opening penalty* $g$ (possibly 0) and all *gap extension penalties* occurring in the alignment of the gap, where the gap extension penalties are the values of $d_M(s, \Delta)$ for $s \in \Sigma$. If all gap extension penalties are zero, then we have a scoring scheme with *fixed gap penalties.*

# Scoring alignments of two sequences

Suppose we are given a scoring matix $M$ and a gap opening penalty $g$, the cost of aligning two sequences $s_1$ and $s_2$ of equal length $m$ is $d_M(s_1, s_2) = g(G_1 + G_2) + \sum_{i=1}^{m} d_M(s_1[i], s_2[i])$, where $G_j$ is the number of gaps in $s_j$.

# Multiple sequence alignments

Given a set $< t_1, \ldots, t_k >$ of sequences over the alphabet $\Sigma \cup \{\Delta\}$, a *multiple alignment* is a set $< at_1, \ldots, at_k >$ of equal-length sequences (where $at_i$ stands for *aligned* $t_i$) over the alphabet $\Sigma \cup \{\Delta\}$ such that each $at_i$ can be obtained from $t_i$ by inserting some space symbols into the sequences without altering the order of symbols of $t_i$.

# The SP-Alignment Problem

Given two equal-length sequences $at_1, at_2$, their *induced pairwise alignment* is the pair of sequences $bt_1, bt_2$ that is obtained from $at_1, at_2$ by removing all columns containing only $\Delta$'s.

The SP-Alignment problem for a given scoring scheme $(d_M, g)$ is to find the multiple alignment $< at_1, \ldots, at_k >$ that minimizes the score $SP(< at_1, \ldots at_k >) = \sum_{1 \leq i < j \leq k} d(bt_i, bt_j)$ among all possible multiple alignments of $< t_1, \ldots, t_k >$.

# Variants of the SP-Alignment Problem

- In the Gap-0-Alignment Problem, spaces may be inserted only at the beginning or the end of sequences.

- In the Gap-0-1-Alignment Problem, all sequences have the same length, and only one space may be inserted at he beginning or end of each sequence.

- In the Space-L-Alignment Problem, into each sequence at most $L$ spaces may be inserted (with no restrictions on where these insertions may occur).

These variants have some relevance to molecular biology, since optimal alinements of biomolecular sequences typically contain relatively few space symbols, and frequently do have gaps at the beginning or end of sequences.

# An ancient result

**Theorem 1.** *(Wang and Jiang, 1994) There exists a (nonmetric) scoring scheme for which the* SP-Alignment *Problem is NP-hard.*

# A medieval result

**Theorem 2.** *(Just; proved 1999, appeared 2001) "For all scoring schemes actually used in molecular biology," each of the following problems is NP-hard:*

- *The* SP-Alignment *Problem.*
- *The* Gap-0-Alignment *Problem.*
- *The* Gap-0-1-Alignment *Problem.*

The proof of this theorem shows that the analogous result also holds for Space-L-Alignment, but this has not been stated explicitly in the paper.

# An old open question

In a 1999 survey paper, Jiang, Kearney, and Li asked: “Does SP-Alignment admit a PTAS if we assume that the scoring matrix is metric?”

# Another medieval result

**Theorem 3.** *(Just; proved 1999, appeared 2001) There exists a nonmetric scoring scheme such that each of the following problems is MAX-SNP-hard:*

- *The* SP-Alignment *Problem.*
- *The* Gap-0-Alignment *Problem.*
- *The* Gap-0-1-Alignment *Problem.*

The proof of this theorem shows that the analogous result also holds for Space-L-Alignment; this has not been stated explicitly in (Just, 2001), but a corresponding result (for fixed gap penalties) appears in (Just and Della Vedova, 2004).

# An excursion into Operations Research

Suppose a communication network is to be set up in a country that consists of $k$ regions. In each region, there should be one switchboard of the network, and each switchboard is to be connected by expensive, high quality cable to every other switchboard. If in each region there are several possible locations for the switchboard that are equally good for the operation of the network within this region, then the locations of switchboards should be chosen in such a way as to minimize overall cost of cable between them.

# The Switchboard Location Problem

The Switchboard Location problem has as instance some disjoint sets $R_1, \ldots, R_k$ called *regions*, as well as a distance function $d$ between all pairs $< x_i, x_j >$ of points in $R_1 \cup \cdots \cup R_k$. Unlike in (Just and Della Vedova, 2004), we will assume throughout this talk that $d$ is a metric. A *feasible solution* is a set $< x_1, \ldots, x_k >$ of points such that $x_i \in R_i$ for $1 \leq i \leq k$. The problem asks for a feasible solution that minimizes $\sum_{1 \leq i < j \leq k} d(x_i, x_j)$.

# The Switchboard Location$_P$ Problem

Let $P$ be a positive integer. The Switchboard Location$_P$ problem is the restriction of the Switchboard Location problem to regions of size at most $P$.

The *spread* of an instance $I$ of Switchboard Location$_P$ problem is the quotient of largest distance between points from different regions and the smallest distance between points from different regions. The Switchboard Location$_P(\sigma)$ problem is the restriction of the Switchboard Location$_P$ problem instances of spread at most $\sigma$.

# An easy theorem

**Theorem 4.** *(Just and Della Vedova, 2004) The* Switchboard Location$_2$ *problem is NP-hard.*

**Proof:** Given a graph $G =< V, E >$ with vertices $V = \{v_1, \ldots, v_k\}$, we construct a metric space $X = \{x_1, \ldots, x_k, y_1, \ldots, y_k\}$ as follows: For $i \neq j$, we let $d(x_i, x_j) = d(y_i, y_j) = 2$. If $\{v_i, v_j\} \in E$, then $d(x_i, y_j) = 1$; if $\{v_i, v_j\} \notin E$, then $d(x_i, y_j) = 2$. For $1 \leq i \leq k$, the region $R_i$ is defined as $\{x_i, y_i\}$. This gives us an instance $I$ of the Switchboard Location$_2$ problem. Note that every optimal solution of $I$ induces a cut in $G$ of maximal size. Now the theorem follows from NP-hardness of MAX-CUT. □

# A harder theorem

**Theorem 5.** *(Just and Della Vedova, 2004) For every fixed $P$, the* Switchboard Location$_P$ *problem (restricted to metric distances) admits a PTAS.*

First the theorem was proved without the restiction on metricity for Switchboard Location$_P(\sigma)$ for any fixed $\sigma > 1$. (Just and Della Vedova, 2000). The elegant one-page proof uses a powerful theorem of (Arora, Karger, and Karpinski, 1999). Several years later I showed how the assumption of metricity allows one to remove the restriction on instances of small spread. This part of the proof takes up five pages of calculations in (Just and Della Vedova, 2004).

# Don’t get too excited

The exponent of the expected running time for the PTAS contains a factor of $\frac{1}{\varepsilon^2}$, where $\varepsilon$ is the desired relative accuracy.

# Back to the Gap-0-1 Alignment Problem

Suppose we are given sequences $t_1, \ldots, t_k$ of equal length and a metric scoring scheme. We can then construct an abstract metric space $\{\Delta t_1, t_1\Delta, \ldots, \Delta t_k, t_k\Delta\}$ that consists of the original sequences with space symbols inserted either to their right or to their left, and the distance defined by the scoring function. If we define regions $R_i = \{\Delta t_i, t_i\Delta\}$, then finding the optimal gap-0-1 alignment for these sequences clearly becomes an instance of the Switchboard Location$_2$ Problem! This proves:

**Theorem 6.** *(Just and Della Vedova, 2004) The* Gap-0-1 Alignment *Problem for metric scoring schemes admits a PTAS.*

# Back to the Space-L Alignment Problem

Suppose we are given sequences $t_1, \ldots, t_k$ of length at most $n$, a metric scoring scheme, and a positive integer $L$. We can then construct for each $i$ a set $R_i$ that consists of all possible sequences that can be obtained from $t_i$ by inserting at most $L$ space symbols into it. The union of these $R_i$'s forms an abstract metric space whose distance is defined by the scoring function. If $P$ denotes the maximum cardinality of all these regions $R_i$, then finding the optimal space-L alignment for these sequences clearly becomes an instance of the Switchboard Location$_P$ Problem! This should prove that the Space-L Alignment Problem for metric scoring schemes admits a PTAS.

# Not so fast!

There are $\binom{n+L+1}{L}$ ways of inserting at most $L$ spaces into a sequence of length $n$, which makes the $P$ in our previous slide equal to $\binom{n+L+1}{L}$, and thus causes the exponent in our running time to grow for fixed $\varepsilon$ with increasing sequence length. This is not what we had in mind when we said "PTAS."

# A partial rescue

Roughly speaking, the *spread* $\sigma$ of an instance of the Space-L Alignment Problem is the ratio between the optimal Space-L alignment of the two most distance sequences and the optimal Space-L alignment of the two closest sequences. The Space-L Alignment ($\sigma$) Problem is the restriction of the Space-L Alignment Problem to instances of spread at most $\sigma$.

**Theorem 7.** *(Just and Della Vedova, 2004) Let $\sigma$ be a constant. Then the* Space-$L$ Multiple Alignment*($\sigma$) Problem has a PTAS. This is true even if metricity of the scoring scheme is not assumed.*

# Biological significance of the last theorem

It is an empirical fact that multiple sequence alignment is easy if all the sequences are closely related (except in cases where they have long repeats or inversions). Our last theorem shows that, in a sense, Space-L Alignment is also "easy" if *none* of the sequences are closely related to each other. If the Space-$L$ Multiple Alignment Problem with metric scoring schemes does not have a PTAS, then the really hard cases must be mixtures of closely related and very distant sequences.

# References

1. Arora, S., D. Karger, M. Karpinski; *Polynomial-time approximation schemes for dense instances of NP-hard problems.* Journal of Computational Systems Science **58** (1999) 193-210.

2. Jiang, T., P. Kearney, M. Li; *Some open problems in computational molecular biology.* SIGACT News **30** (1999) 43–49.

3. Just, W.;*Computational complexity of multiple sequence alignment with SP-score.* Journal of Computational Biology **8** (2001) 615–623.

4. Just, W. and G. Della Vedova; *Multiple Sequence Alignment as a Facility Location*

*Problem.* INFORMS Journal on Computing **16** (2004) 430–440.

5. Wang, L. and T. Jiang; *On the complexity of multiple sequence alignment.* Journal of Computational Biology **1** (1994) 337–348.