Approximating an optimal Sum-of-Pairs multiple alignment
Sum-of-Pairs Multiple Alignment

Problem

Let \( F \) be a set of \( k \) strings, each of length \( \leq n \), we know how to an optimal SP-alignment \( M^* \) in time \( O(n^k) \) using dynamic programming.

We will show how to compute an alignment \( M \) in time \( O(k^2n^2) \) s.t.

\[
SP(M) < 2 \cdot SP(M^*)
\]

Notation

Let \( d(x,y) \) be a metric between characters

Let \( D(S,S') \) be the induced metric between strings as given by the optimal score of a global pairwise alignment (with linear gap cost)
Alignments consistent with a tree

\[ \text{Score}(M(S_1, S_4)) = \text{Score}(\begin{array}{cccc} A & - & - & C \ G & - & T \\ A & T & T & C & - & - & T \\ C & T & - & C & G & - & A \\ A & - & - & C & G & G & T \end{array}) \geq D(S_1, S_4) \]
Alignments consistent with a tree

**Definition (Gusfield, p. 347):** Let $F$ be a set of strings, and let $T$ be a tree where each node is labeled with a distinct string from $F$. Then, a multiple alignment $M$ of $F$ is called *consistent* with $T$ if the induced pairwise alignment of $S_i$ and $S_j$ has score $D(S_i, S_j)$ for each pair of strings $(S_i, S_j)$ that label adjacent nodes in $T$.
Definition (Gusfield, p. 347): Let $F$ be a set of strings, and let $T$ be a tree where each node is labeled with a distinct string from $F$. Then, a multiple alignment $\mathcal{M}$ of $F$ is called consistent with $T$ if the induced pairwise alignment of $S_i$ and $S_j$ has score $D(S_i, S_j)$ for each pair of strings $(S_i, S_j)$ that label adjacent nodes in $T$. 

The “guide” tree

$\mathcal{M}$:

$\begin{align*}
\text{Score}(\mathcal{M}(S_1, S_2)) &= \text{Score}(A - - C G - T, A - - C G G T) \\
&\geq D(S_1, S_4)
\end{align*}$

“=” if consistent with “guide tree”
Lemma 14.6.1 (Gusfield, p. 347): For any set of strings $F$ and for any tree $T$ whose nodes are labeled by distinct strings of $F$, we can efficiently find a multiple alignment $M(T)$ of $F$ that is consistent with $T$. 

Definition (Gusfield, p. 347): Let $F$ be a set of strings, and let $T$ be a tree where each node is labeled with a distinct string from $F$. Then, a multiple alignment $M$ of $F$ is called consistent with $T$ if the induced pairwise alignment of $S_i$ and $S_j$ has score $D(S_i, S_j)$ for each pair of strings $(S_i, S_j)$ that label adjacent nodes in $T$. 

The “guide” tree

The image illustrates the concept of alignments consistent with a tree. The score of an alignment $M(S_1, S_2)$ is compared to the score $D(S_1, S_4)$ for any pair of strings $(S_1, S_4)$. This comparison ensures that the alignment is consistent with the “guide tree.”
Algorithm

Input: A set $F$ of $k$ strings, each of length $\leq n$

Step 1 – Find the “center” string

Find $S_1$ such that $\sum_{S \in F - S_1} D(S_1, S)$ is minimized.

Call the remaining strings $S_2, S_3, \ldots, S_k$
**Algorithm**

**Input:** A set $\mathcal{F}$ of $k$ strings, each of length $\leq n$

**Step 1 – Find the “center” string**

Find $S_1$ such that $\sum_{S \in \mathcal{F} - S_1} D(S_1, S)$ is minimized.

Call the remaining strings $S_2, S_3, ..., S_k$
Algorithm

Input: A set $\mathcal{F}$ of $k$ strings, each of length $\leq n$

Step 1 – Find the “center” string $S_1$

Find $S_1$ such that $\sum_{S \in \mathcal{F} - S_1} D(S_1, S)$ is minimized

The “guide” tree

Call the remaining strings $S_2, S_3, \ldots, S_k$

Takes time $O(n^2)$ for each of the $k(k-1)$ pairs of strings
Algorithm

Input: A set $\mathcal{F}$ of $k$ strings, each of length $\leq n$

**Step 1 – Find the “center” string**

Find $S_1$ such that

$$\sum_{S \in \mathcal{F} - S_1} D(S_1, S)$$

is minimized.

Call the remaining strings $S_2, S_3, ..., S_k$.

**Step 2 – Construct alignment $\mathcal{M}$ cf. “guide tree”**

$$M_1 = [S_1]$$

for $i = 2$ to $k$:

$A = \text{optalign}(S_1, S_i)$

$M_i = "M_{i-1} \text{ extended with } A"$

$\mathcal{M} = M_k$

The “guide” tree

$S_2$

$S_3$

$S_4$

Takes time $O(n^2)$ for each of the $k(k-1)$ pairs of strings.
**Algorithm**

Step 1 – Find the “center” string

Call the remaining strings $S_2, S_3, \ldots, S_k$ is minimized.

Find $S_1$ such that $\sum_{S \in \mathcal{F} - S_1} D(S_1, S)$ is minimized

Step 2 – Construct alignment $\mathcal{M}$ cf. “guide tree”

- Construct alignment $\mathcal{M}$ of $k$ strings, each of length $\leq n$

Example

Assume that $i=4$:

$S_1 = A - - C G T$
$S_2 = A T T C - - T$
$M_3 = C T - - C G A$

$S_4 = A C G G T$
$A C G - - T$
$A = A C G G T$

Extend $M_3$ with $A$ gives:

$A - - C G - - T$
$A T T C - - T$
$C T - - C G - - A$
$M_4 = A - - C G G T$

Note that the new column does not affect $\text{Score}(S_1, S_i)$ for $i < 4$

Takes time $O(n^2)$ for each of the $k(k-1)$ pairs of strings

$M_1 = [S_1]$

for $i = 2$ to $k$:

$A = \text{optalign}(S_1, S_i)$

$M_i = "M_{i-1} \text{ extended with } A"

M = M_k$
**Algorithm**

Input: A set $F$ of $k$ strings, each of length $\leq n$

Step 1 – Find the “center” string $S_1$ such that $\sum_{S \in F - S_1} D(S, S_1)$ is minimized.

- $S_4 = A\ C\ G\ G\ T$
- $S_3 = A\ C\ G\ -\ T$
- $S_2 = A\ -\ -\ C\ G\ T$
- $S_1 = C\ T\ -\ C\ G\ A$

Construct alignment $M$ of “guide tree” $T$.

- $M_1 = [S_1]$
- for $i = 2$ to $k$: $A = \text{optalign}(S_1, S_i)$
- $M_i = \text{"}M_{i-1}\ \text{extended with } A\text{"}$
- $M = M_k$

Takes time $O(n^2)$ for each of the $k(k-1)$ pairs of strings

Takes time $O(kn^2)$

Example

Assume that $i=4$:

<table>
<thead>
<tr>
<th>$S_1$</th>
<th>$S_2$</th>
<th>$S_3$</th>
<th>$S_4$</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>A</td>
<td>C</td>
<td>A</td>
</tr>
<tr>
<td>C</td>
<td>T</td>
<td>G</td>
<td>C</td>
</tr>
<tr>
<td>G</td>
<td>T</td>
<td>G</td>
<td>G</td>
</tr>
<tr>
<td>T</td>
<td>C</td>
<td>A</td>
<td>T</td>
</tr>
</tbody>
</table>

$M_3$ = $C\ T\ -\ C\ G\ A$

$M_4$ = $A\ -\ -\ C\ G\ G\ T$

Note that the new column does not affect $\text{Score}(S_1, S_i)$ for $i < 4$.

The “guide” tree

- $S_1$
- $S_2$
- $S_3$
- $S_4$
Algorithm

Input: A set \( F \) of \( k \) strings, each of length \( \leq n \)

### Step 1 – Find the “center” string

Find \( S_1 \) such that \[ \sum_{S \in F - S_1} D(S_1, S) \] is minimized.

Call the remaining strings \( S_2, S_3, \ldots, S_k \)

### Step 2 – Construct alignment \( M \) cf. “guide tree”

\[ M_1 = [S_1] \]

for \( i = 2 \) to \( k \):

\[ A = \text{optalign}(S_1, S_i) \]

\[ M_i = "M_{i-1} \text{ extended with } A" \]

\( M = M_k \)
Algorithm

Input: A set $\mathcal{F}$ of $k$ strings, each of length $\leq n$

Step 1 – Find the “center” string

Find $S_1$ such that $\sum_{S \in \mathcal{F} - S_1} D(S_1, S)$ is minimized.

Call the remaining strings $S_2, S_3, \ldots, S_k$

Running time: $O(k^2n^2 + kn^2) = O(k^2n^2)$

Step 2 – Construct alignment $\mathcal{M}$ cf. “guide tree”

$\mathcal{M}_1 = [S_1]$
for $i = 2$ to $k$:
\[
A = \text{optalign}(S_1, S_i)
\]
$\mathcal{M}_i = \text{“}M_{i-1} \text{ extended with } A\text{”}$
$\mathcal{M} = \mathcal{M}_k$
Approximation Ratio, part 1

Finding an upper bound of the computed alignment $\mathcal{M}$

$$SP(\mathcal{M}) = \frac{1}{2} \sum_{i=1}^{k} \sum_{j=1, i \neq j}^{k} \text{Score}(\mathcal{M}(S_i, S_j))$$

$$= \frac{1}{2} \sum_{i=1}^{k} \sum_{j=1, i \neq j}^{k} d(i, j)$$

The score of the alignment of $S_i$ and $S_j$ as induced by $\mathcal{M}$
Approximation Ratio, part 1

Finding an upper bound of the computed alignment $\mathcal{M}$

$$\text{SP}(\mathcal{M}) = \frac{1}{2} \sum_{i=1}^{k} \sum_{j=1, i \neq j}^{k} \text{Score}(\mathcal{M}(S_i, S_j))$$

$$= \frac{1}{2} \sum_{i=1}^{k} \sum_{j=1, i \neq j}^{k} d(i, j)$$

Using the triangle-inequality and symmetry. Valid because the substitution matrix is metric.

$$\leq \frac{1}{2} \sum_{i=1}^{k} \sum_{j=1, i \neq j}^{k} (d(i, 1) + d(1, j))$$

$$= \frac{1}{2} \sum_{i=1}^{k} \sum_{j=1, i \neq j}^{k} (d(1, i) + d(1, j))$$

The score of the alignment of $S_i$ and $S_j$ as induced by $\mathcal{M}$.
Approximation Ratio, part 1

Finding an upper bound of the computed alignment $\mathcal{M}$

$$SP(\mathcal{M}) = \frac{1}{2} \sum_{i=1}^{k} \sum_{j=1, i \neq j}^{k} \text{Score}(\mathcal{M}(S_i, S_j))$$

$$= \frac{1}{2} \sum_{i=1}^{k} \sum_{j=1, i \neq j}^{k} d(i, j)$$

By the triangle-inequality and symmetry. Valid because the substitution matrix is metric.

$$\leq \frac{1}{2} \sum_{i=1}^{k} \sum_{j=1, i \neq j}^{k} (d(i, 1) + d(1, j))$$

$$= \frac{1}{2} \sum_{i=1}^{k} \sum_{j=1, i \neq j}^{k} (d(1, i) + d(1, j))$$

Expanding and rewriting the sum.

$$= \frac{1}{2} \sum_{l=2}^{k} 2(k - 1)d(1, l)$$

$$= (k - 1) \sum_{l=2}^{k} \text{Score}(\mathcal{M}(S_1, S_l))$$
Approximation Ratio, part 1

Finding an upper bound of the computed alignment $\mathcal{M}$

$$SP(\mathcal{M}) = \frac{1}{2} \sum_{i=1}^{k} \sum_{j=1, i \neq j}^{k} \text{Score}(\mathcal{M}(S_i, S_j))$$

Using the triangle-inequality and symmetry. Valid because the substitution matrix is metric

$$\leq \frac{1}{2} \sum_{i=1}^{k} \sum_{j=1, i \neq j}^{k} (d(i, 1) + d(1, j))$$

Expanding and rewriting the sum

$$= \frac{1}{2} \sum_{l=2}^{k} 2(k - 1)d(1, l)$$

$$= (k - 1) \sum_{l=2}^{k} \text{Score}(\mathcal{M}(S_1, S_l))$$

$\mathcal{M}$ is consistent with the guide tree, where $S_1$ is the center

$$= (k - 1) \sum_{l=2}^{k} D(S_1, S_l)$$
\[
\frac{1}{2} \sum_{i=1}^{k} \sum_{j=1, i \neq j}^{k} (d(1, i) + d(1, j)) = \frac{1}{2} \left( \sum_{i=1}^{k} \sum_{j=1, i \neq j}^{k} d(1, i) + \sum_{i=1}^{k} \sum_{j=1, i \neq j}^{k} d(1, j) \right)
\]
\[
= \frac{1}{2} \left( \sum_{i=1}^{k} (k - 1)d(1, i) + (\sum_{j=1}^{k} \sum_{j=1, i \neq j}^{k} d(1, j) - \sum_{j=1}^{k} d(1, j)) \right)
\]
\[
= \frac{1}{2} \left( \sum_{i=1}^{k} (k - 1)d(1, i) + \sum_{j=1}^{k} k \cdot d(1, j) - \sum_{j=1}^{k} d(1, j) \right)
\]
\[
= \frac{1}{2} \left( \sum_{i=1}^{k} (k - 1)d(1, i) + \sum_{j=1}^{k} (k - 1)d(1, j) \right)
\]
\[
= \frac{1}{2} \sum_{l=2}^{k} (k - 1)d(1, l) + \sum_{l=2}^{k} (k - 1)d(1, l)
\]
\[
= \frac{1}{2} \sum_{l=2}^{k} 2(k - 1)d(1, l)
\]
Approximation Ratio, part 2

Finding a lower bound of the score of an optimal alignment $M^*$

$$SP(M^*) = \frac{1}{2} \sum_{i=1}^{k} \sum_{j=1,i\neq j}^{k} \text{Score}(M^*(S_i, S_j))$$

$$= \frac{1}{2} \sum_{i=1}^{k} \sum_{j=1,i\neq j}^{k} d^*(i, j)$$

The score of the alignment of $S_i$ and $S_j$ as induced by $M^*$
Approximation Ratio, part 2

Finding a lower bound of the score of an optimal alignment $\mathcal{M}^*$

$$\text{SP}(\mathcal{M}^*) = \frac{1}{2} \sum_{i=1}^{k} \sum_{j=1, i \neq j}^{k} \text{Score}(\mathcal{M}^*(S_i, S_j))$$

$$= \frac{1}{2} \sum_{i=1}^{k} \sum_{j=1, i \neq j}^{k} d^*(i, j)$$

Nothing is better than the optimal scores

$$\geq \frac{1}{2} \sum_{i=1}^{k} \sum_{j=1}^{k} D(S_i, S_j)$$

The score of the alignment of $S_i$ and $S_j$ as induced by $\mathcal{M}^*$
Approximation Ratio, part 2

Finding a lower bound of the score of an optimal alignment $M^*$

\[
\text{SP}(M^*) = \frac{1}{2} \sum_{i=1}^{k} \sum_{j=1, i \neq j}^{k} \text{Score}(M^*(S_i, S_j))
\]

\[= \frac{1}{2} \sum_{i=1}^{k} \sum_{j=1, i \neq j}^{k} d^*(i, j)
\]

The score of the alignment of $S_i$ and $S_j$ as induced by $M^*$

Nothing is better than the optimal scores

By choice of $S_1$ we have $D(S_1, S_j) \leq D(S, S_j)$

\[
\geq \frac{1}{2} \sum_{i=1}^{k} \sum_{j=1}^{k} D(S_i, S_j)
\]

\[\geq \frac{1}{2} \sum_{i=1}^{k} \sum_{j=1}^{k} D(S_1, S_j)
\]
Approximation Ratio, part 2

Finding a lower bound of the score of an optimal alignment $M^*$

$$SP(M^*) = \frac{1}{2} \sum_{i=1}^{k} \sum_{j=1, i \neq j}^{k} \text{Score}(M^*(S_i, S_j))$$

$$= \frac{1}{2} \sum_{i=1}^{k} \sum_{j=1, i \neq j}^{k} d^*(i, j)$$

The score of the alignment of $S_i$ and $S_j$ as induced by $M^*$.

Nothing is better than the optimal scores.

By choice of $S_1$ we have $D(S_1, S_j) \leq D(S_p, S_j)$.

$$\geq \frac{1}{2} \sum_{i=1}^{k} \sum_{j=1}^{k} D(S_i, S_j)$$

Rewriting and renaming.

$$\geq \frac{1}{2} \sum_{i=1}^{k} \sum_{j=1}^{k} D(S_1, S_j)$$

$$= \frac{1}{2} \sum_{j=1}^{k} D(S_1, S_j)$$

$$= \frac{1}{2} \sum_{l=2}^{k} D(S_1, S_l)$$
Approximation Ratio, part 3

<table>
<thead>
<tr>
<th>Upper-bound</th>
<th>Lower-bound</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\text{SP}(\mathcal{M}) \leq (k - 1) \sum_{l=2}^{k} D(S_1, S_l)$</td>
<td>$\text{SP}(\mathcal{M}^*) \geq \frac{1}{2} k \sum_{l=2}^{k} D(S_1, S_l)$</td>
</tr>
</tbody>
</table>

Using the upper- and lower-bounds we get

$$\frac{\text{SP}(\mathcal{M})}{\text{SP}(\mathcal{M}^*)} \leq \frac{(k - 1) \sum_{l=2}^{k} D(S_1, S_l)}{\frac{1}{2} k \sum_{l=2}^{k} D(S_1, S_l)} = \frac{2(k - 1)}{k} < 2$$
# Approximation Ratio, part 3

<table>
<thead>
<tr>
<th>Upper-bound</th>
<th>Lower-bound</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\text{SP}(\mathcal{M}) \leq (k - 1) \sum_{l=2}^{k} D(S_1, S_l)$</td>
<td>$\text{SP}(\mathcal{M}^*) \geq \frac{1}{2} k \sum_{l=2}^{k} D(S_1, S_l)$</td>
</tr>
</tbody>
</table>

Using the upper- and lower-bounds we get

$$\frac{\text{SP}(\mathcal{M})}{\text{SP}(\mathcal{M}^*)} \leq \frac{(k - 1) \sum_{l=2}^{k} D(S_1, S_l)}{\frac{1}{2} k \sum_{l=2}^{k} D(S_1, S_l)} = \frac{2(k - 1)}{k} < 2$$

$$\text{SP}(\mathcal{M}) < 2 \cdot \text{SP}(\mathcal{M}^*)$$
Can we do better?

SP-multiple alignment is NP-complete [Wang and Jiang 1994]

PTAS by [Bafna, Lawler, Pevzner 1995] gives

$$\frac{\text{SP}(\mathcal{M})}{\text{SP}(\mathcal{M}^*)} \leq 2 - \frac{q}{k}$$

in time $O(k^3 n^{2q-1})$, where $1 \leq q < k$

Using the upper- and lower-bounds we get

$$\frac{\text{SP}(\mathcal{M})}{\text{SP}(\mathcal{M}^*)} \leq \frac{(k - 1) \sum_{l=2}^{k} D(S_1, S_l)}{\frac{1}{2} k \sum_{l=2}^{k} D(S_1, S_l)} = \frac{2(k - 1)}{k} < 2$$

$$\text{SP}(\mathcal{M}) < 2 \cdot \text{SP}(\mathcal{M}^*)$$