

**Approximating an optimal  
Sum-of-Pairs multiple alignment**

# Sum-of-Pairs Multiple Alignment

## Problem

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

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

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

## 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

$\mathcal{M}$ :

A	-	-	C	G	-	T	$S_1$
A	T	T	C	-	-	T	
C	T	-	C	G	-	A	
A	-	-	C	G	G	T	$S_4$

$$\text{Score}(\mathcal{M}(S_1, S_2)) = \text{Score}\left(\begin{array}{cccccc} \mathbf{A} & \mathbf{-} & \mathbf{-} & \mathbf{C} & \mathbf{G} & \mathbf{-} & \mathbf{T} \\ \mathbf{A} & \mathbf{-} & \mathbf{-} & \mathbf{C} & \mathbf{G} & \mathbf{G} & \mathbf{T} \end{array}\right) \geq D(S_1, S_2)$$

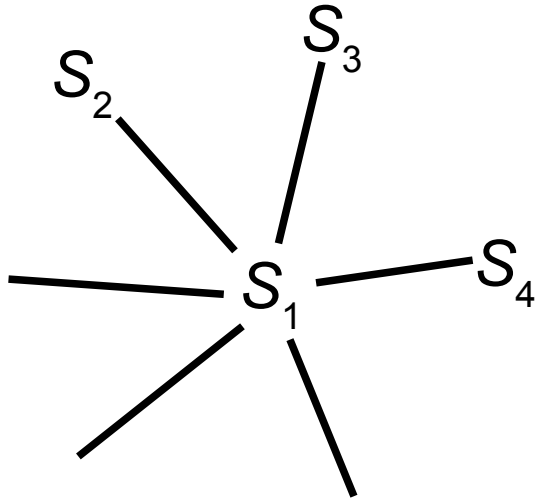
# Alignments consistent with a tree

$$\mathcal{M}: \begin{array}{cccccc} \text{A} & - & - & \text{C} & \text{G} & - & \text{T} & S_1 \\ \text{A} & \text{T} & \text{T} & \text{C} & - & - & \text{T} & \\ \text{C} & \text{T} & - & \text{C} & \text{G} & - & \text{A} & \\ \text{A} & - & - & \text{C} & \text{G} & \text{G} & \text{T} & S_4 \end{array}$$

$$\text{Score}(\mathcal{M}(S_1, S_2)) = \text{Score} \left( \begin{array}{cccccc} \text{A} & - & - & \text{C} & \text{G} & - & \text{T} \\ \text{A} & - & - & \text{C} & \text{G} & \text{G} & \text{T} \end{array} \right) \geq D(S_1, S_2)$$

**Definition (Gusfield, p. 347):** Let  $\mathcal{F}$  be a set of strings, and let  $T$  be a tree where each node is labeled with a distinct string from  $\mathcal{F}$ . Then, a multiple alignment  $\mathcal{M}$  of  $\mathcal{F}$  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



**s consistent with a tree**

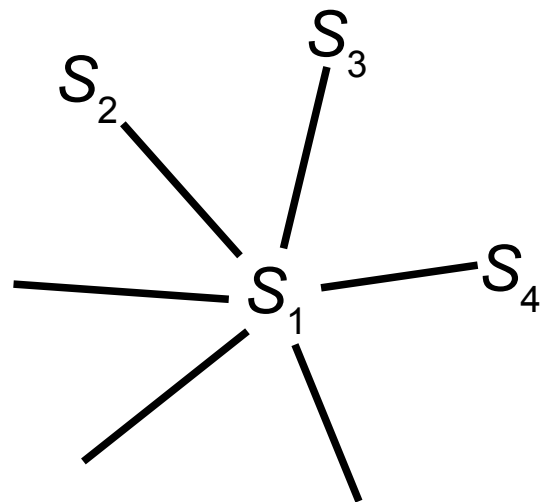
A	-	-	C	G	-	T	$S_1$
A	T	T	C	-	-	T	
C	T	-	C	G	-	A	
A	-	-	C	G	G	T	$S_4$

"=" if consistent with "guide tree"

$$\text{Score}(\mathcal{M}(S_1, S_2)) = \text{Score} \left( \begin{array}{cccccc} \text{A} & - & - & \text{C} & \text{G} & - & \text{T} \\ \text{A} & - & - & \text{C} & \text{G} & \text{G} & \text{T} \end{array} \right) \geq D(S_1, S_2)$$

**Definition (Gusfield, p. 347):** Let  $\mathcal{F}$  be a set of strings, and let  $T$  be a tree where each node is labeled with a distinct string from  $\mathcal{F}$ . Then, a multiple alignment  $\mathcal{M}$  of  $\mathcal{F}$  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



**s consistent with a tree**

A	-	-	C	G	-	T	$S_1$
A	T	T	C	-	-	T	
C	T	-	C	G	-	A	
A	-	-	C	G	G	T	$S_4$

"=" if consistent with "guide tree"

$$\text{Score}(\mathcal{M}(S_1, S_2)) = \text{Score} \left( \begin{array}{cccccc} \text{A} & - & - & \text{C} & \text{G} & - & \text{T} \\ \text{A} & - & - & \text{C} & \text{G} & \text{G} & \text{T} \end{array} \right) \geq D(S_1, S_2)$$

**Definition (Gusfield, p. 347):** Let  $\mathcal{F}$  be a set of strings, and let  $T$  be a tree where each node is labeled with a distinct string from  $\mathcal{F}$ . Then, a multiple alignment  $\mathcal{M}$  of  $\mathcal{F}$  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  $T$

**Lemma 14.6.1 (Gusfield, p. 347):** For any set of strings  $\mathcal{F}$  and for any tree  $T$  whose nodes are labeled by distinct strings of  $\mathcal{F}$ , we can efficiently find a multiple alignment  $\mathcal{M}(T)$  of  $\mathcal{F}$  that is consistent with  $T$

# 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  $S_2, S_3, \dots, S_k$

# Algorithm

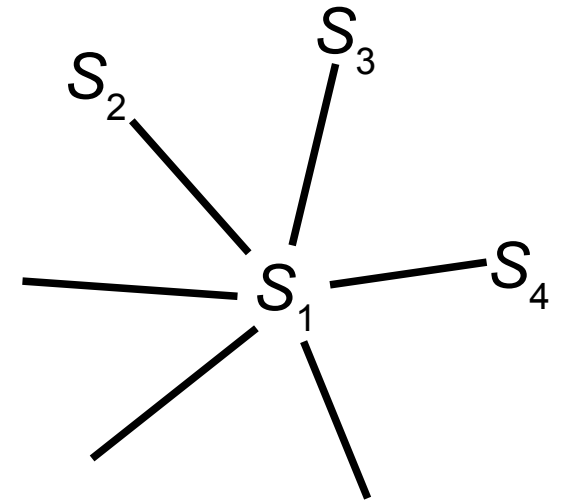
**Input:** A set  $\mathcal{F}$  of  $k$  strings, each of

**Step 1 – Find the “center”  $s$**

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

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

The “guide” tree



# Algorithm

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

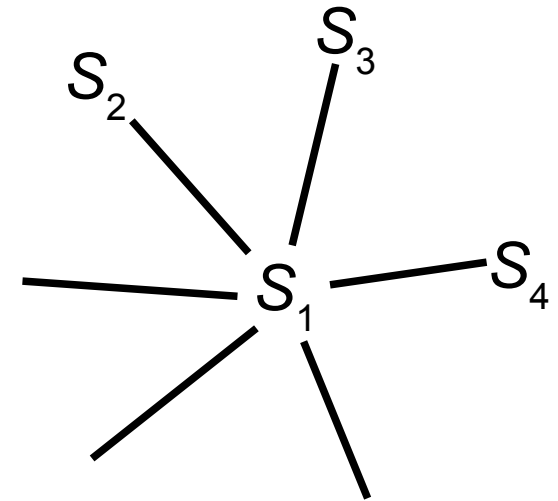
**Step 1 – Find the “center”  $s$**

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

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

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

The “guide” tree



# Algorithm

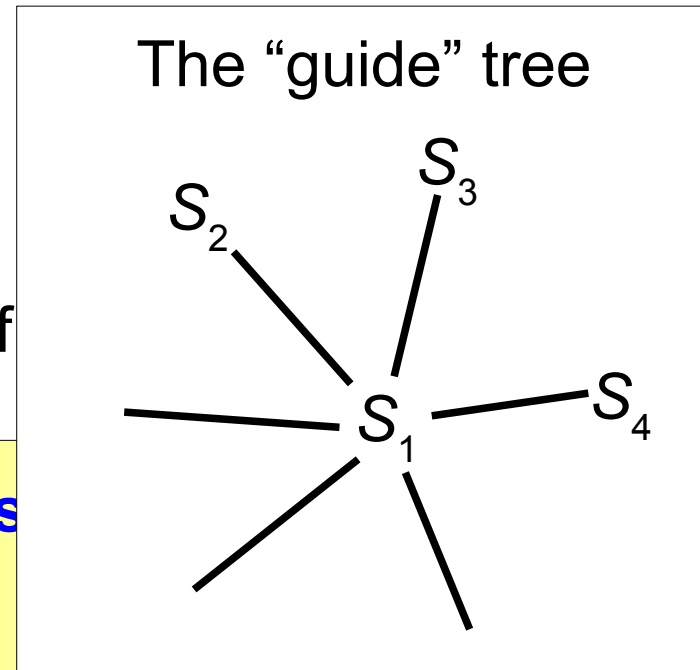
**Input:** A set  $\mathcal{F}$  of  $k$  strings, each of

## Step 1 – Find the “center” $S_1$

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

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

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



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

$M_1 = [S_1]$

for  $i = 2$  to  $k$ :

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

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

$\mathcal{M} = M_k$

# Example

Assume that  $i=4$ :

```

A - - C G T
A T T C - T
M3 = 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
M4 = A - - C G G T
    
```

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

# Algorithm

Let  $F$  of  $k$  strings, each of length  $n$

– Find the “center” string  $S_1$

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

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

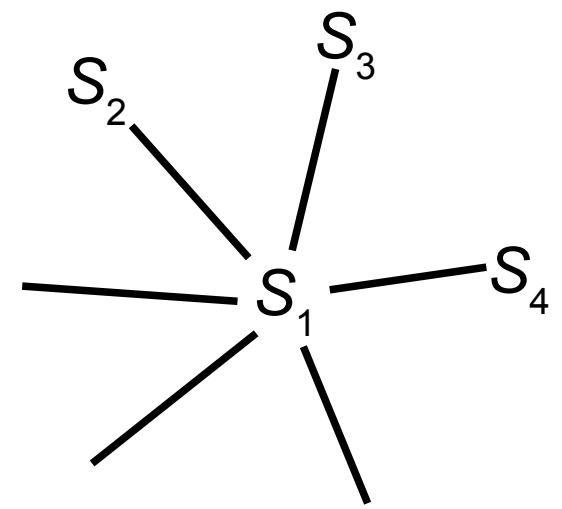
and then align  $S_1$  with each of the remaining strings  $S_2, S_3, \dots, S_k$

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

```

M1 = [S1]
for i = 2 to k:
  A = optalign(S1, Si)
  Mi = "Mi-1 extended with A"
M = Mk
    
```

The “guide” tree



# Example

Assume that  $i=4$ :

```

A - - C G T
A T T C - T
M3 = 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
M4 = A - - C G G T

```

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

# Algorithm

Let  $F$  of  $k$  strings, each of length  $n$

– Find the “center” string  $S_1$

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

for each of the remaining strings  $S_2, S_3, \dots, S_k$

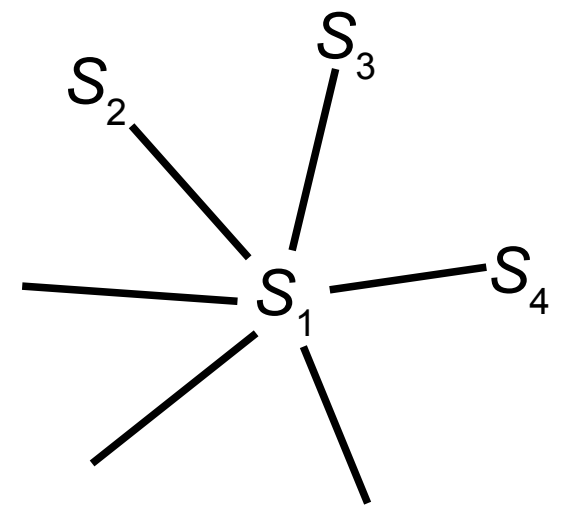
– Construct alignment  $M_i$  of “guide tree”

```

M1 = [S1]
for i = 2 to k:
  A = optalign(S1, Si)
  Mi = "Mi-1 extended with A"
M = Mk

```

The “guide” tree



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

Takes time  $O(kn^2)$

# 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  $S_2, S_3, \dots, S_k$

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

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

# 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  $S_2, S_3, \dots, S_k$

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

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

$M_1 = [S_1]$

for  $i = 2$  to  $k$ :

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

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

$\mathcal{M} = M_k$

# Approximation Ratio, part 1

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

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

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  $\mathcal{M}$

$$\begin{aligned} \text{SP}(\mathcal{M}) &= \frac{1}{2} \sum_{i=1}^k \sum_{j=1, i \neq k}^k \text{Score}(\mathcal{M}(S_i, S_j)) \\ &= \frac{1}{2} \sum_{i=1}^k \sum_{j=1, i \neq j}^k d(i, j) \\ &\leq \frac{1}{2} \sum_{i=1}^k \sum_{j=1, i \neq j}^k (d(i, 1) + d(1, j)) \\ &= \frac{1}{2} \sum_{i=1}^k \sum_{j=1, i \neq j}^k (d(1, i) + d(1, j)) \end{aligned}$$

The score of the alignment of  $S_i$  and  $S_j$  as induced by  $\mathcal{M}$

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

# Approximation Ratio, part 1

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

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

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

The score of the alignment of  $S_i$  and  $S_j$  as induced by  $\mathcal{M}$

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

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

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

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

Expanding and rewriting the sum

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

# Approximation Ratio, part 1

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

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

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

The score of the alignment of  $S_i$  and  $S_j$  as induced by  $\mathcal{M}$

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

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

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

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

Expanding and rewriting the sum

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

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

$$= (k-1) \sum_{l=2}^k D(S_1, S_l)$$

# Approximation Ratio, part 2

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

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

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  $\mathcal{M}^*$

$$\begin{aligned} \text{SP}(\mathcal{M}^*) &= \frac{1}{2} \sum_{i=1}^k \sum_{j=1, i \neq j}^k \text{Score}(\mathcal{M}^*(S_i, S_j)) \\ &= \frac{1}{2} \sum_{i=1}^k \sum_{j=1, i \neq j}^k d^*(i, j) \\ &\geq \frac{1}{2} \sum_{i=1}^k \sum_{j=1}^k D(S_i, S_j) \end{aligned}$$

The score of the alignment of  $S_i$  and  $S_j$  as induced by  $\mathcal{M}^*$

Nothing is better than the optimal scores

# Approximation Ratio, part 2

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

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

The score of the alignment of  $S_i$  and  $S_j$  as induced by  $\mathcal{M}^*$

Nothing is better than the optimal scores

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

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

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

# Approximation Ratio, part 2

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

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

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

The score of the alignment of  $S_i$  and  $S_j$  as induced by  $\mathcal{M}^*$

Nothing is better than the optimal scores

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

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

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

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

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

Rewriting and renaming

# Approximation Ratio, part 3

**Upper-bound**

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

**Lower-bound**

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

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

**Upper-bound**

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

**Lower-bound**

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

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

$$\frac{\text{SP}(\mathcal{M})}{\text{SP}(\mathcal{M}^*)} \leq 2 - \frac{q}{k} \quad \text{in time } O(k^3 n^{2q-1}), \text{ where } 1 \leq q < k$$

Using the upper- and lower-bounds we get

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

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