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.

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

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

M:

```
A - - C G - T   S1
A T T C - - T
C T - C G - A
A - - C G G T   S4
```

$$\text{Score}(M(S_1,S_4)) = \text{Score}\left(\begin{array}{ccccccc} A & - & - & C & G & - & T \\ A & - & - & C & G & G & T \end{array}\right) \geq D(S_1, S_4)$$

Alignments consistent with a tree

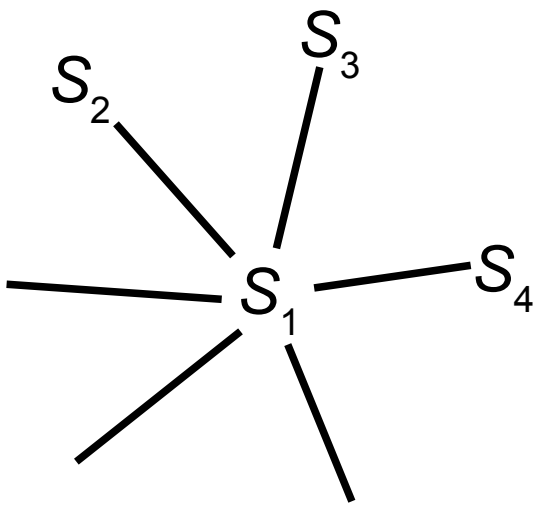
M:

```
A - - C G - T   S1
A T T C - - T
C T - C G - A
A - - C G G T   S4
```

$$\text{Score}(M(S_1,S_4)) = \text{Score}\left(\begin{array}{c} \texttt{A - - C G - T} \\ \texttt{A - - C G G T} \end{array}\right) \geq D(S_1, S_4)$$

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



s consistent with a tree

M:

```
A - - C G - T   S1
A T T C - - T
C T - C G - A
A - - C G G T   S4
```

"=" if consistent with "guide tree"

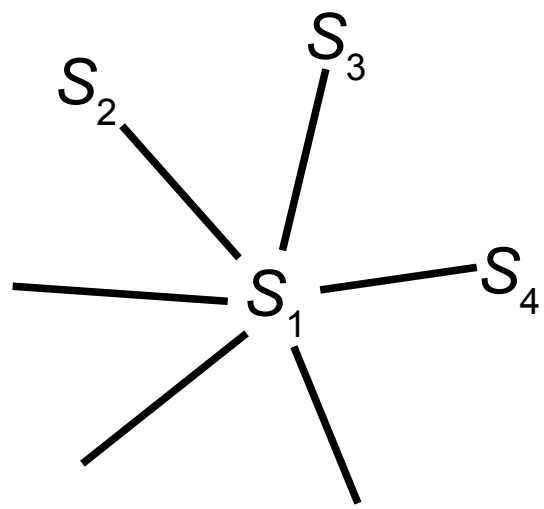
Score($M(S_1,S_4)$) = Score(

```
A - - C G - T
A - - C G G T
```

) ≥ D(S_1, S_4)

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



s consistent with a tree

M:

```
A - - C G - T   S1
A T T C - - T
C T - C G - A
A - - C G G T   S4
```

"=" if consistent with "guide tree"

Score($M(S_1,S_4)$) = Score(

```
A - - C G - T
A - - C G G T
```

) ≥ D(S_1, S_4)

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*

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

Algorithm

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

Step 1 – Find the "center" string

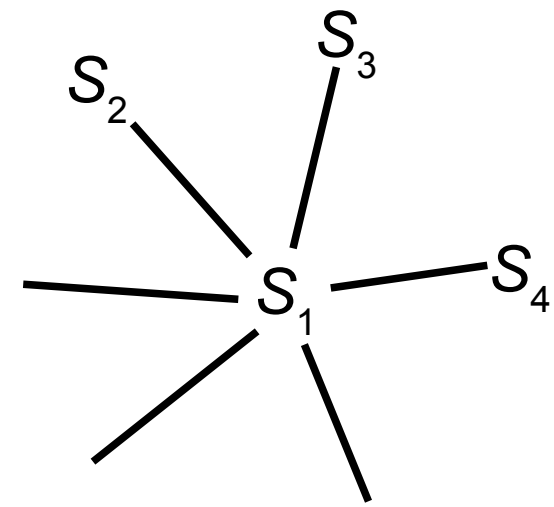
Find S_1 such that $\sum_{S \in \mathcal{F} - S_1} D(S_1, S)$ is minimized.

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

Algorithm

Input: A set F of k strings, each of

The "guide" tree



Step 1 – Find the "center" s

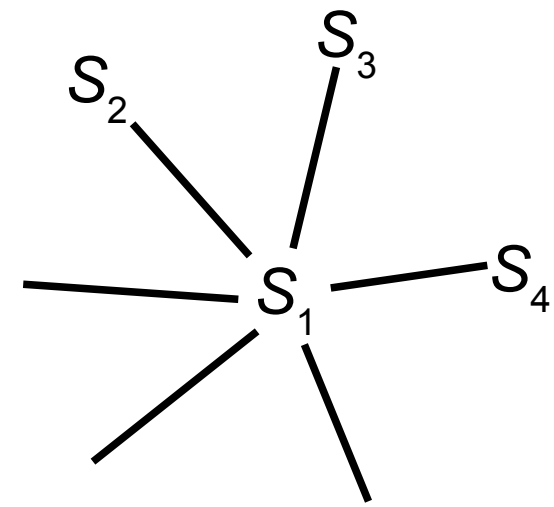
Find S_1 such that $\sum_{S \in \mathcal{F} - S_1} D(S_1, S)$ is minimized.

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

Algorithm

Input: A set F of k strings, each of

The "guide" tree



Step 1 – Find the "center" s

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

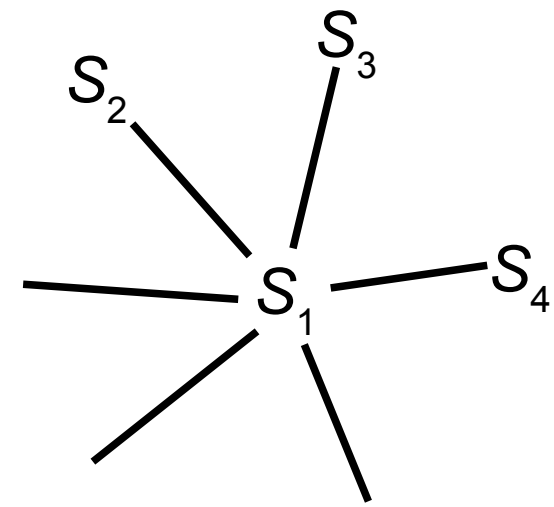
Takes time $O(n^2)$ for each of the $k(k\text{-}1)$ pairs of strings

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

Algorithm

Input: A set F of k strings, each of

The "guide" tree



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 $k(k\text{-}1)$ pairs of strings

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

Step 2 – Construct alignment *M* cf. "guide tree"

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

Algorithm

t F of k strings, each of

– Find the "center" s

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

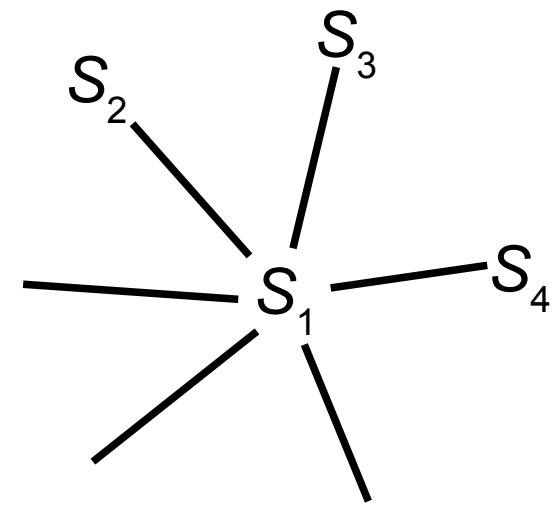
remaining strings $S_2, S_3, \ldots, S_k$

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

struct alignment *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

S4 = 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 Score(S_1, S_i) for i < 4

Algorithm

t F of k strings, each of

– Find the "center" s

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

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

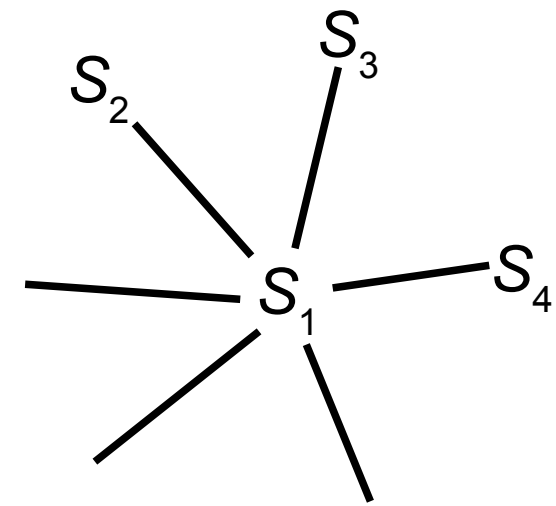
Takes time O(n^2) for each of the $k(k$-1) pairs of strings

struct alignment M of "guide tree"

Takes time O(kn^2)

```
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

S4 = 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 Score(S_1, S_i) for i < 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 $\sum_{S \in \mathcal{F} - S_1} D(S_1, S)$ is minimized.

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

Step 2 – Construct alignment *M* cf. "guide tree"

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

Algorithm

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

Step 1 – Find the "center" string

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

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

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

Step 2 – Construct alignment *M* cf. "guide tree"

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

How to extend M with A?

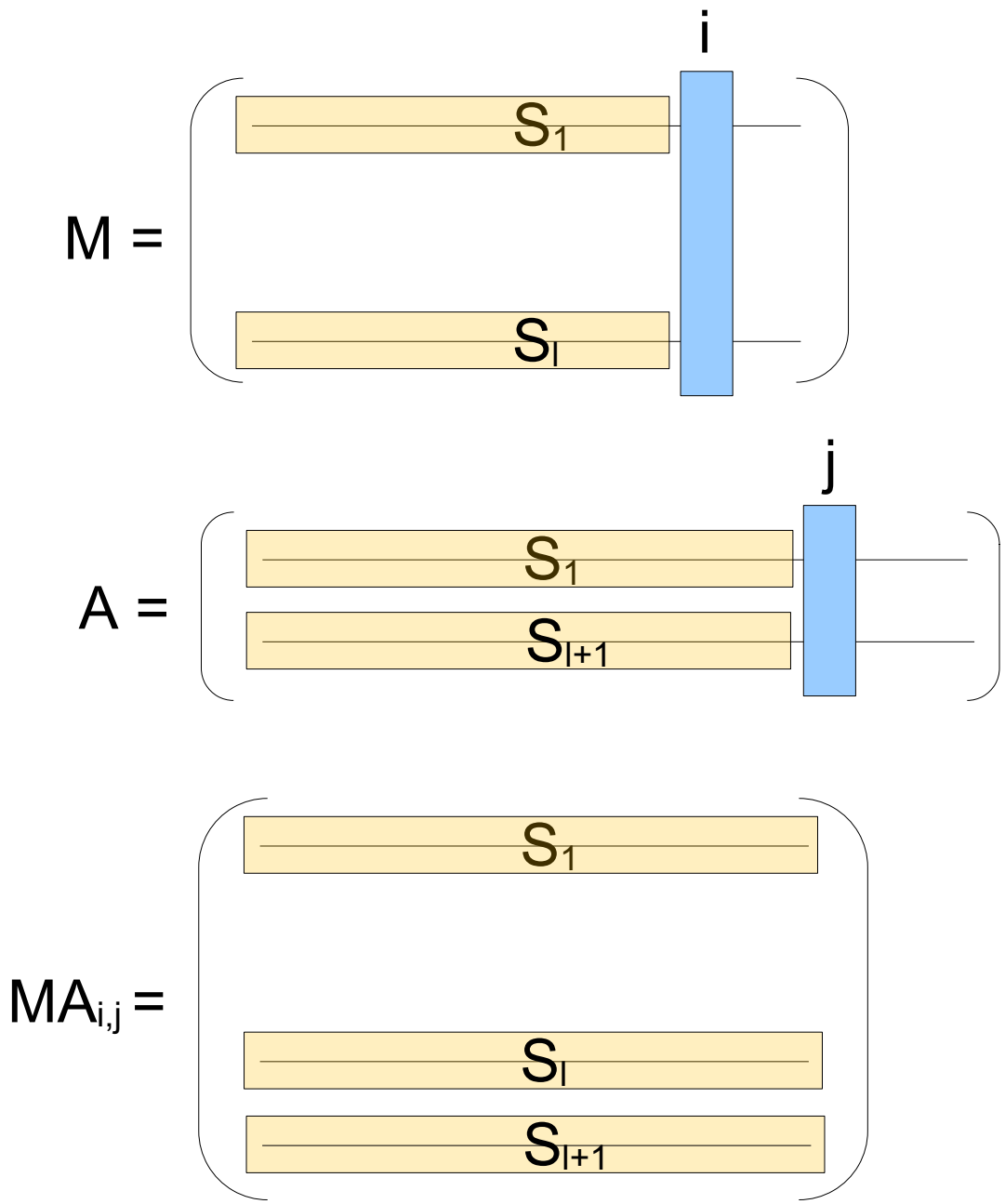
Let M be a multiple alignment, and let A be a pairwise alignment such that the first row of M and the first row of A is the same string if gaps are removed, i.e. like M_3 and A on the previous slides.

Let i be a column in M and j be column in A such the first row of M and A up to (but not including) column i and j respectively is the same string if gaps are removed.

Let MA be an extension of M with A that is consistent with the guide tree.

In particular, let $MA_{i,j}$ be an extension of all columns before column i in M and all columns before column j in A that is consistent with the guide tree.

Clearly $MA_{0,0}$ is the empty alignment, and $MA = MA_{len(M)+1,\ len(A)+1}$



How to extend M with A?

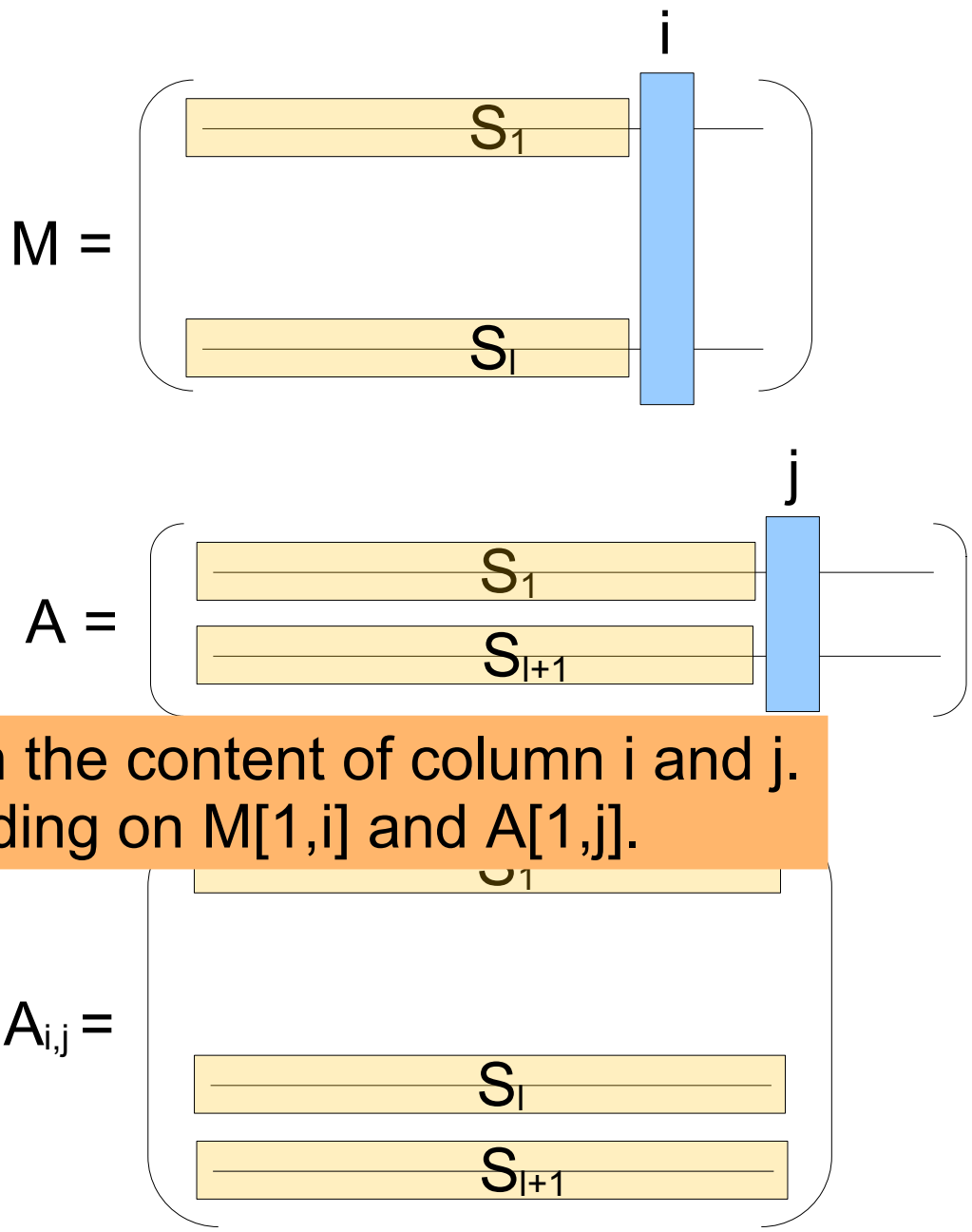
Let M be a multiple alignment, and let A be a pairwise alignment such that the first row of M and the first row of A is the same string if gaps are removed, i.e. like M_3 and A on the previous slides.

Let i be a column in M and j be column in A such the first row of M and A up to (but not including) column i and j respectively is the same string if gaps are removed.

Let MA
that is c

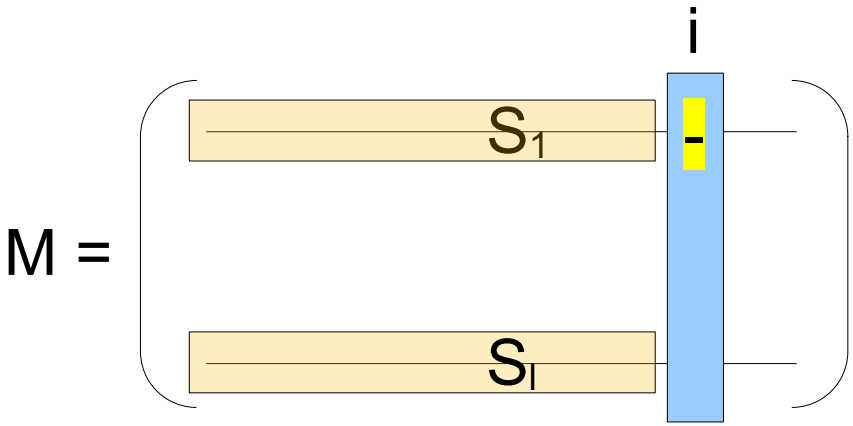
In particular, let $MA_{i,j}$ be an extension of all columns before column i in M and all columns before column j in A that is consistent with the guide tree.

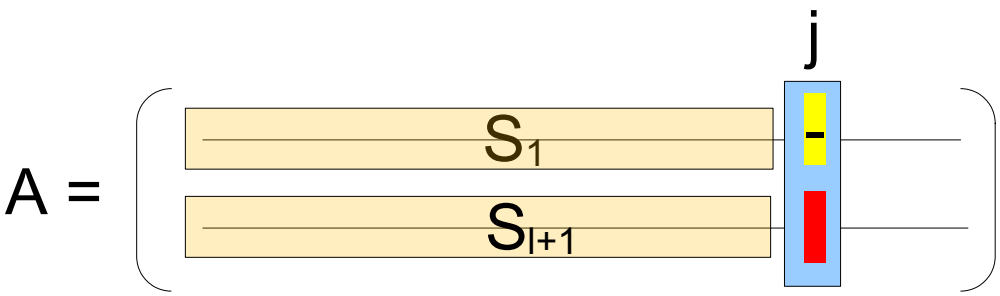
Clearly $MA_{0,0}$ is the empty alignment, and $MA = MA_{len(M)+1,\ len(A)+1}$

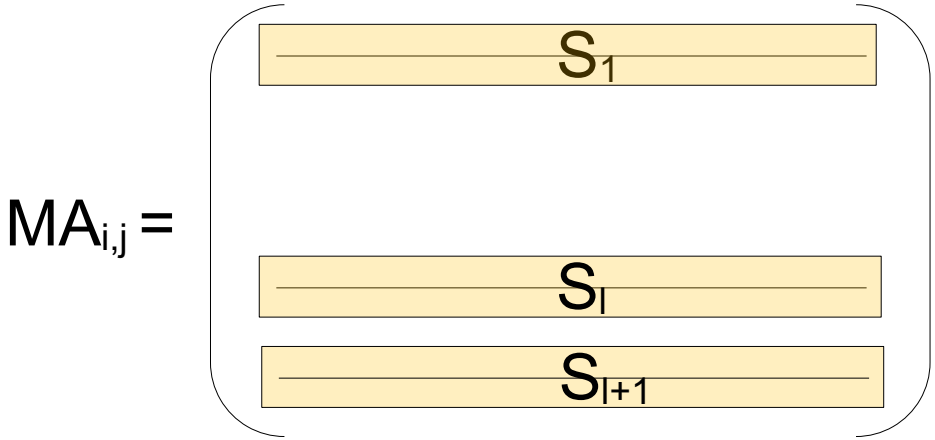


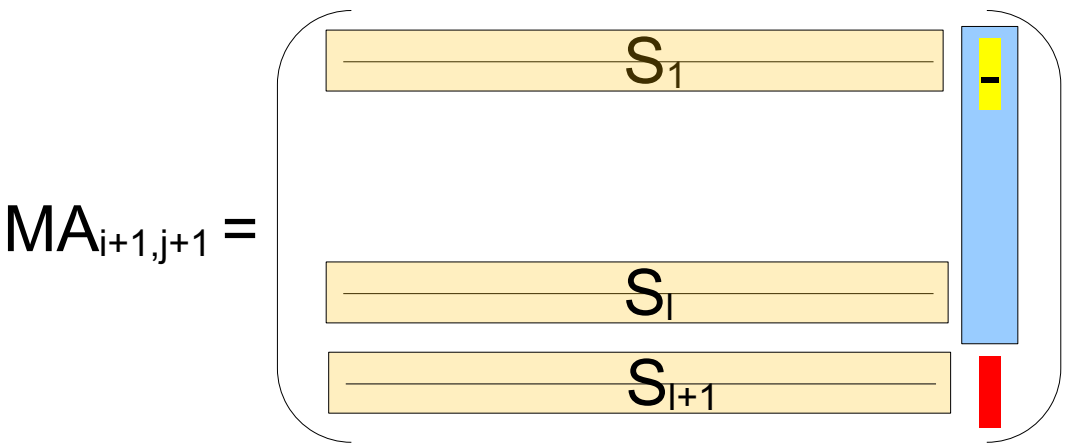
We extend $M_{i,j}$ depending on the content of column i and j. There are four cases depending on M[1,i] and A[1,j].

Case 1



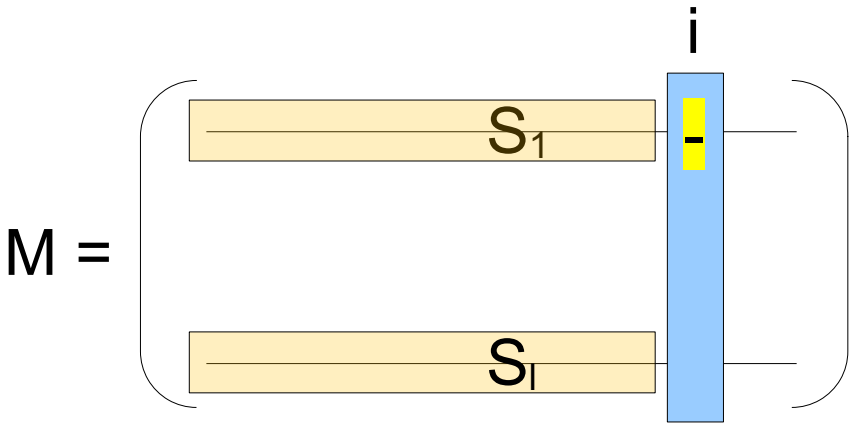


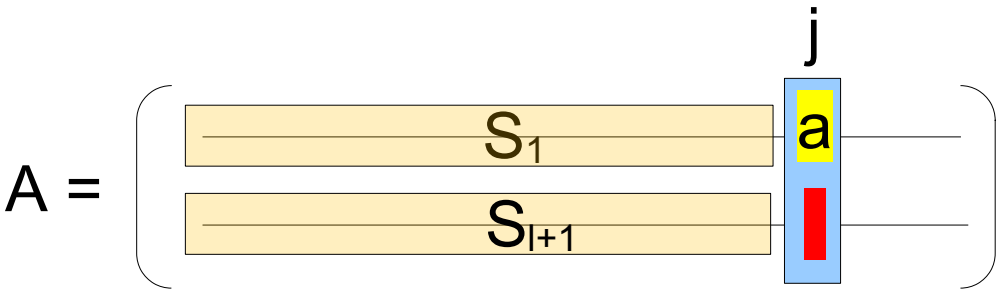


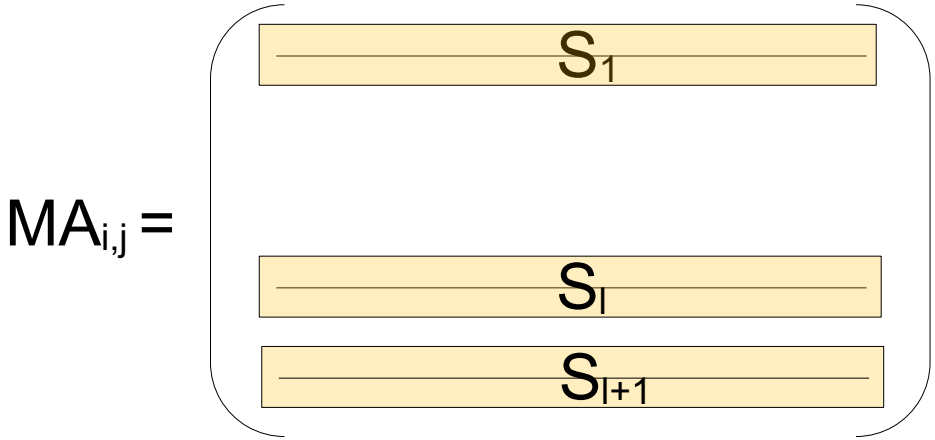


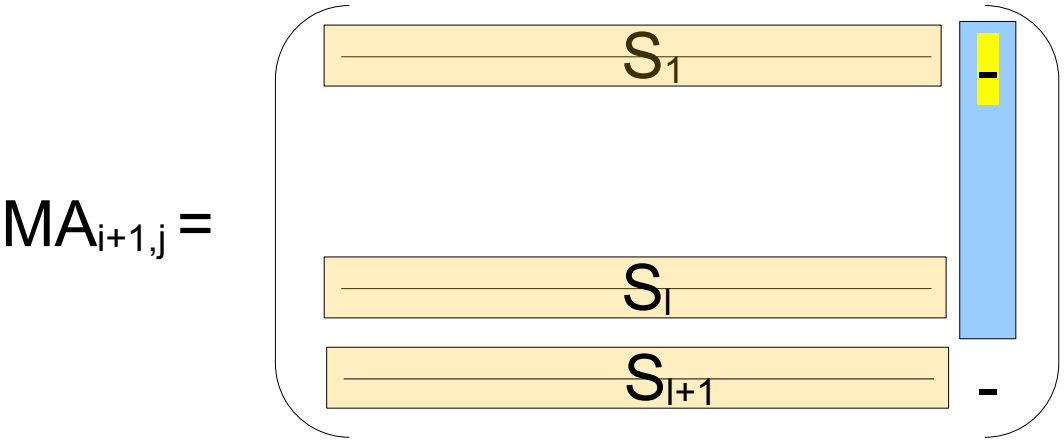
$MA_{i+1,j+1}$ is consistent with the guide tree. Continue with column i+1 in M and j+1 in A

Case 2



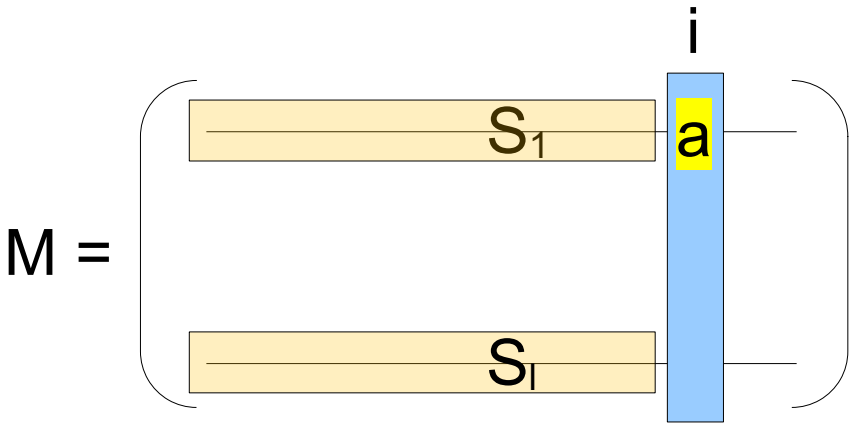


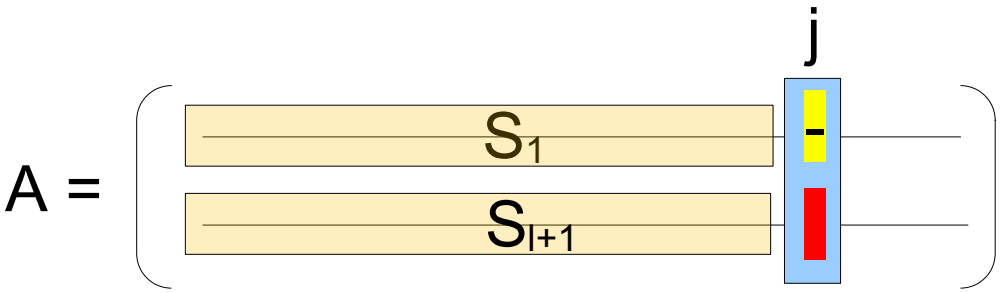


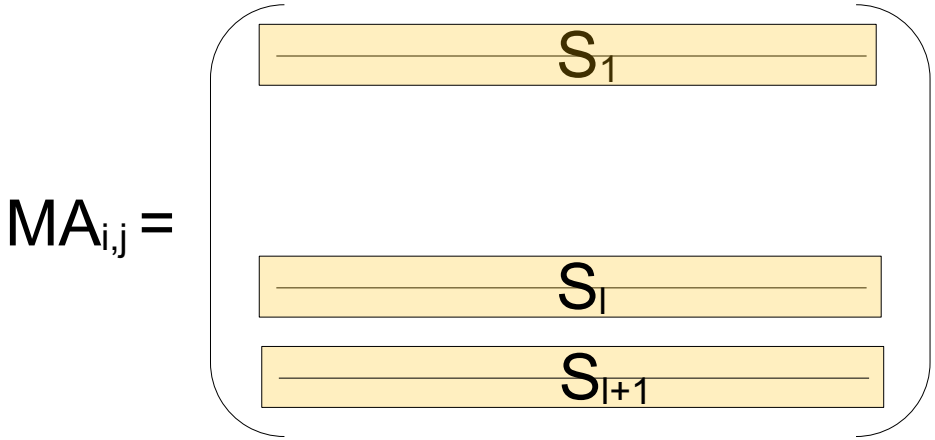


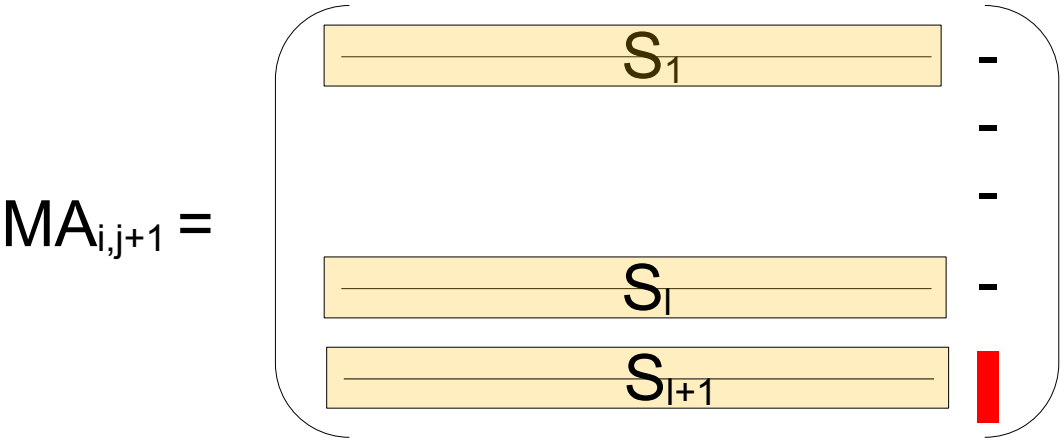
$MA_{i+1,j}$ is consistent with the guide tree. Continue with column i+1 in M and j in A

Case 3



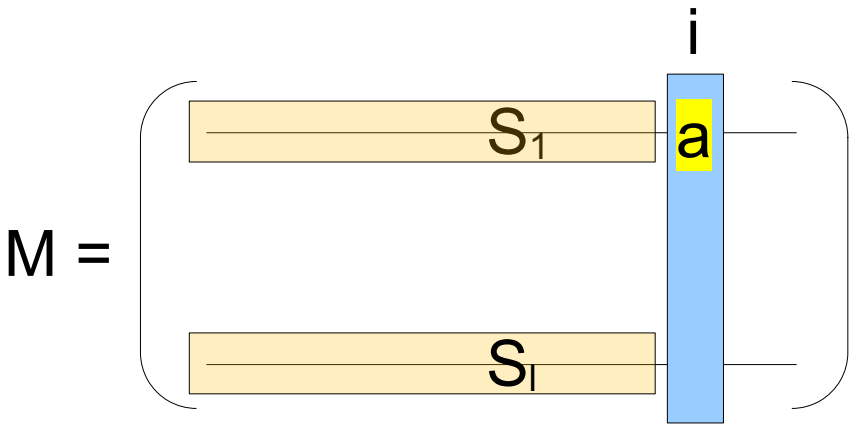


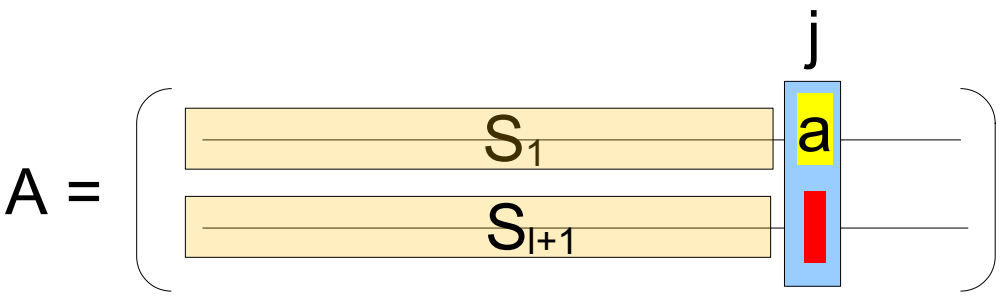


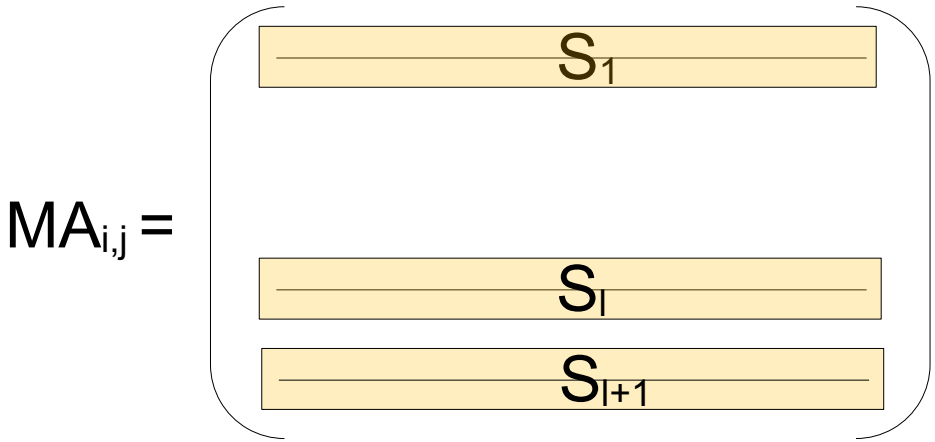


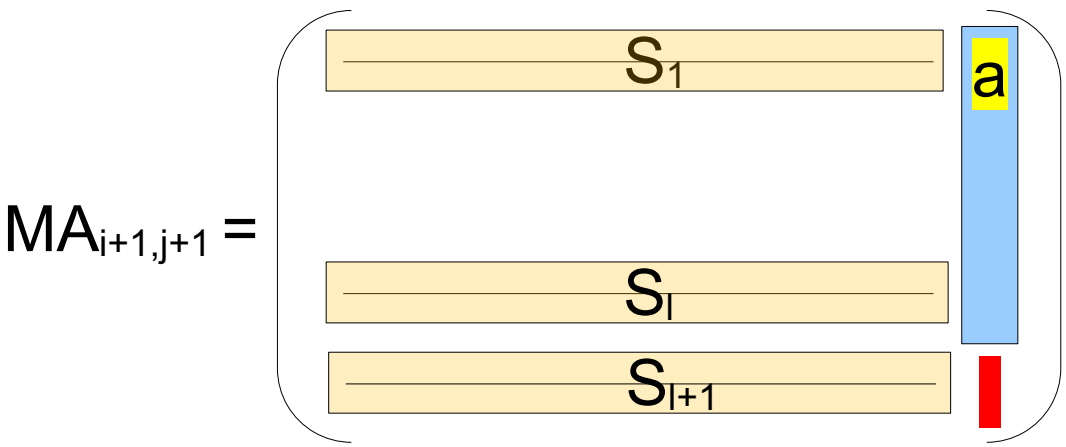
$MA_{i,j+1}$ is consistent with the guide tree. Continue with column i in M and j+1 in A

Case 4







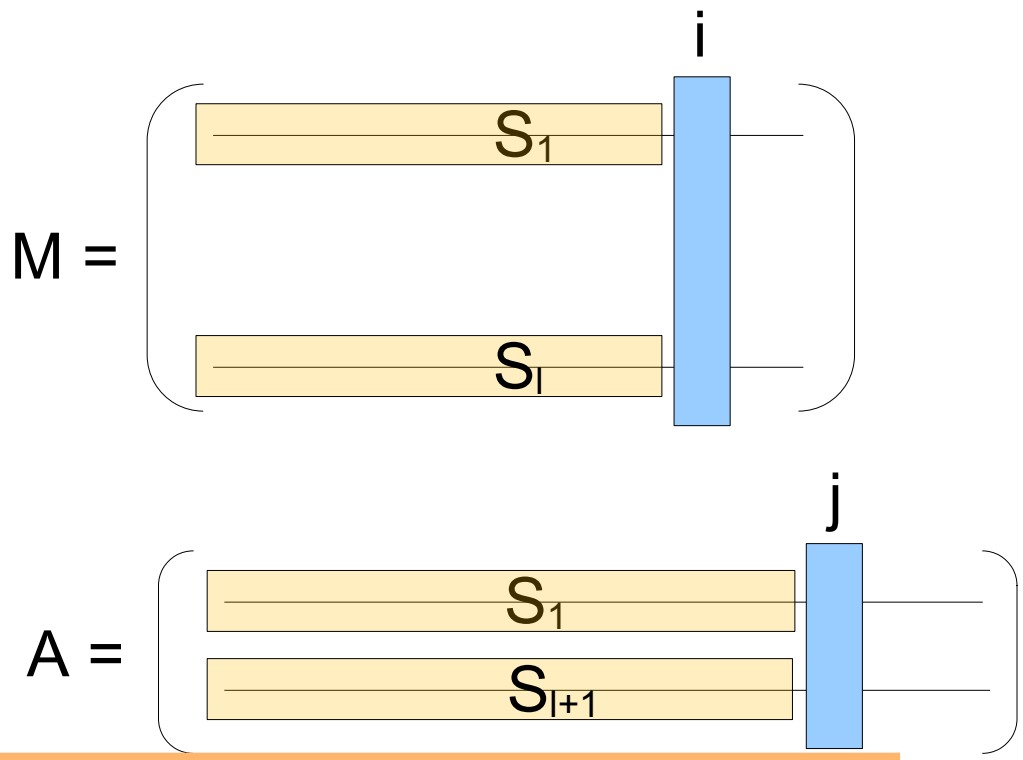


$MA_{i+1,j+1}$ is consistent with the guide tree. Continue with column i+1 in M and j+1 in A

How to extend M with A?

Let M be a multiple alignment, and let A be a pairwise alignment such that the first row of M and the first row of A is the same string if gaps are removed, i.e. like M_3 and A on the previous slides.

Let i be a column in M and j be column in A such the first row of M and A up to (but not including) column i and j respectively is the same string if gaps are removed.



Let M
that is

In par
of all
and a
that is

Clearl
and M len(M)+1, len(A)+1

We extend $M_{i,j}$ depending on the content of column i and j. There are four cases depending on M[1,i] and A[1,j].

Summary: If we can extend $M_{i,j}$ in time O(1), then we can build MA from $MA_{0,0}$ in time **O(#columns in M + #columns in A)** by iteratively applying cases 1-4. Think about how to represent M, A, and MA

Algorithm

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

Step 1 – Find the "center" string

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

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

Step 2 – Construct alignment *M* cf. "guide tree"

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

Algorithm

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

Step 1 – Find the "center" string

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

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

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

Step 2 – Construct alignment *M* cf. "guide tree"

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

Approximation Ratio?

We want to prove that $SP(M) < 2\ SP(M^*)$. How?

Approximation Ratio, part 1

Finding an upper bound of the computed alignment *M*

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

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

Approximation Ratio, part 1

Finding an upper bound of the computed alignment *M*

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

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

$$
\begin{aligned}
\mathrm{SP}(\mathcal{M}) &= \frac{1}{2}\sum_{i=1}^{k}\sum_{j=1,i\neq j}^{k} \mathrm{Score}(\mathcal{M}(S_i,S_j)) \\
&= \frac{1}{2}\sum_{i=1}^{k}\sum_{j=1,i\neq j}^{k} d(i,j) \\
&\leq \frac{1}{2}\sum_{i=1}^{k}\sum_{j=1,i\neq j}^{k} (d(i,1)+d(1,j)) \\
&= \frac{1}{2}\sum_{i=1}^{k}\sum_{j=1,i\neq j}^{k} (d(1,i)+d(1,j)) \\
&= \frac{1}{2}\sum_{l=2}^{k} 2(k-1)d(1,l) \\
&= (k-1)\sum_{l=2}^{k} \mathrm{Score}(\mathcal{M}(S_1,S_l))
\end{aligned}
$$

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

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

Expanding and rewriting the sum

Approximation Ratio, part 1

Finding an upper bound of the computed alignment *M*

$$
\begin{aligned}
\mathrm{SP}(\mathcal{M}) &= \frac{1}{2}\sum_{i=1}^{k}\sum_{j=1,i\neq j}^{k} \mathrm{Score}(\mathcal{M}(S_i,S_j)) \\
&= \frac{1}{2}\sum_{i=1}^{k}\sum_{j=1,i\neq j}^{k} d(i,j) \\
&\leq \frac{1}{2}\sum_{i=1}^{k}\sum_{j=1,i\neq j}^{k} (d(i,1)+d(1,j)) \\
&= \frac{1}{2}\sum_{i=1}^{k}\sum_{j=1,i\neq j}^{k} (d(1,i)+d(1,j)) \\
&= \frac{1}{2}\sum_{l=2}^{k} 2(k-1)d(1,l) \\
&= (k-1)\sum_{l=2}^{k} \mathrm{Score}(\mathcal{M}(S_1,S_l)) \\
&= (k-1)\sum_{l=2}^{k} D(S_1,S_l)
\end{aligned}
$$

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

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

Expanding and rewriting the sum

M is consistent with the guide tree, where S_1 is the center

Expanding and rewriting ...

$$
\begin{aligned}
\frac{1}{2}\sum_{i=1}^{k}\sum_{j=1,i\neq j}^{k}(d(1,i)+d(1,j)) &= \frac{1}{2}(\sum_{i=1}^{k}\sum_{j=1,i\neq j}^{k}d(1,i)+\sum_{i=1}^{k}\sum_{j=1,i\neq j}^{k}d(1,j)) \\
&= \frac{1}{2}(\sum_{i=1}^{k}(k-1)d(1,i)+(\sum_{i=1}^{k}\sum_{j=1}^{k}d(1,j)-\sum_{j=1}^{k}d(1,j))) \\
&= \frac{1}{2}(\sum_{i=1}^{k}(k-1)d(1,i)+(\sum_{j=1}^{k}k\cdot d(1,j)-\sum_{j=1}^{k}d(1,j))) \\
&= \frac{1}{2}(\sum_{i=1}^{k}(k-1)d(1,i)+\sum_{j=1}^{k}(k-1)d(1,j)) \\
&= \frac{1}{2}(\sum_{l=2}^{k}(k-1)d(1,l)+\sum_{l=2}^{k}(k-1)d(1,l)) \\
&= \frac{1}{2}\sum_{l=2}^{k}2(k-1)d(1,l)
\end{aligned}
$$

Approximation Ratio, part 2

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

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

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

Approximation Ratio, part 2

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

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

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

Nothing is better than the optimal scores

Approximation Ratio, part 2

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

$$\begin{aligned}
\mathrm{SP}(\mathcal{M}^*) &= \frac{1}{2}\sum_{i=1}^{k}\sum_{j=1,i\neq j}^{k} \mathrm{Score}(\mathcal{M}^*(S_i,S_j)) \\
&= \frac{1}{2}\sum_{i=1}^{k}\sum_{j=1,i\neq j}^{k} d^*(i,j) \\
&\geq \frac{1}{2}\sum_{i=1}^{k}\sum_{j=1}^{k} D(S_i,S_j) \\
&\geq \frac{1}{2}\sum_{i=1}^{k}\sum_{j=1}^{k} D(S_1,S_j)
\end{aligned}$$

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 $\Sigma_j D(S_1, S_j) \leq \Sigma_j D(S_i, S_j)$ for any i

Approximation Ratio, part 2

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

$$
\begin{aligned}
\mathrm{SP}(\mathcal{M}^*) &= \frac{1}{2}\sum_{i=1}^{k}\sum_{j=1,i\neq j}^{k} \mathrm{Score}(\mathcal{M}^*(S_i,S_j)) \\
&= \frac{1}{2}\sum_{i=1}^{k}\sum_{j=1,i\neq j}^{k} d^*(i,j) \\
&\geq \frac{1}{2}\sum_{i=1}^{k}\sum_{j=1}^{k} D(S_i,S_j) \\
&\geq \frac{1}{2}\sum_{i=1}^{k}\sum_{j=1}^{k} D(S_1,S_j) \\
&= \frac{1}{2}k\sum_{j=1}^{k} D(S_1,S_j) \\
&= \frac{1}{2}k\sum_{l=2}^{k} D(S_1,S_l)
\end{aligned}
$$

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 $\Sigma_j D(S_1, S_j) \leq \Sigma_j D(S_i, S_j)$ for any i

Rewriting and renaming

Approximation Ratio, part 3

Upper-bound	Lower-bound
$\mathrm{SP}(\mathcal{M}) \leq (k-1)\sum_{l=2}^{k} D(S_1, S_l)$	$\mathrm{SP}(\mathcal{M}^*) \geq \frac{1}{2}k\sum_{l=2}^{k} D(S_1, S_l)$

Using the upper- and lower-bounds we get

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

Approximation Ratio, part 3

Upper-bound	Lower-bound
$\mathrm{SP}(\mathcal{M}) \leq (k-1)\sum_{l=2}^{k} D(S_1, S_l)$	$\mathrm{SP}(\mathcal{M}^*) \geq \frac{1}{2}k\sum_{l=2}^{k} D(S_1, S_l)$

Using the upper- and lower-bounds we get

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

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

Can we do better?

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

PTAS by [Bafna, Lawler, Pevzner 1995] gives

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

Using the upper- and lower-bounds we get

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

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