

Multiple sequence alignment generated by center star method has sum of pair distance at most twice the optimal sum of pair distance

↗ optimal MSA which minimizes the sum of pair distance

$$M^* = \sum_{1 \leq i < j \leq k} d_{M^*}(i, j)$$

$$\geq \sum_{1 \leq i < j \leq k} D(S_i, S_j)$$

$$= \frac{1}{2} \sum_{i=1}^k \sum_{j=1}^k D(S_i, S_j) \rightarrow \text{center string } S_c$$

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

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

↗ MSA generated by center star method

$$M = \sum_{1 \leq i \leq j \leq k} d_M(i, j)$$

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

$$d_M(X, Y) \leq d_M(X, Z) + d_M(X, Y)$$

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

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

$$= k \sum_j D(S_c, S_j)$$

# Center star method

## Center Star Method

**Require:** A set  $S$  of sequences

**Ensure:** A multiple alignment of  $M$  with sum of pair distances at most twice that of the optimal alignment of  $S$

- 1: Find  $D(S_i, S_j)$  for all  $i, j$ .
- 2: Find the center sequence  $S_c$  which minimizes  $\sum_{i=1}^k D(S_c, S_i)$ .
- 3: For every  $S_i \in S - \{S_c\}$ , choose an optimal alignment between  $S_c$  and  $S_i$ .
- 4: Introduce spaces into  $S_c$  so that the multiple alignment  $M$  satisfies the alignments found in Step 3.

Example: Find center sequence  $s_1$   $O(k^2)$

$S_1$ : CCTGCTGCAG  
 $S_2$ : GATGTGCCG  
 $S_3$ : GATGTGCAG  
 $S_4$ : CCGCTAGCAG  
 $S_5$ : CCTGTAGG

	$S_1$	$S_2$	$S_3$	$S_4$	$S_5$
$S_1$	0	4	3	2	4
$S_2$		0	1	6	5
$S_3$			0	5	5
$S_4$				0	4
$S_5$					0

$\sum_{i=1, k} D(S_1, S_i) = 13$   
 $\sum_{i=1, k} D(S_2, S_i) = 16$   
 $\sum_{i=1, k} D(S_3, S_i) = 14$   
 $\sum_{i=1, k} D(S_4, S_i) = 17$   
 $\sum_{i=1, k} D(S_5, S_i) = 18$

$S_1$ : CCTGCTGCAG  
 $S_2$ : GATG-TGCCG

$S_1$ : CCTGCTGCAG  
 $S_3$ : GATG-TGCAG  
 $S_1$ : CCTGCT-GCAG  
 $S_4$ : CC-GCTAGCAG  
 $S_1$ : CCTGCT-GCAG  
 $S_5$ : CCTG-TAG--G

$S_1$ : CCTGCT-GCAG  
 $S_5$ : CCTG-TAG--G

(a)  
Five input sequences

(b)  
Find  $O(S_i, S_j)$  for all  $i, j$   
 $O(k^2 n^2)$

(c)  
Pairwise alignment between  $S_1$  and other sequences  
 $O(k n^2)$

Convert pairwise alignment to multiple sequence alignment

$S_1$ : CCTGCTGCAG  
 $S_2$ : GATG-TGCCG  
 $S_1$ : CCTGCTGCAG  
 $S_3$ : GATG-TGCAG

$S_1$ : CCTGCTGCAG  
 $S_2$ : GATG-TGCCG  
 $S_3$ : GATG-TGCAG  
 $S_1$ : CCTGCT-GCAG  
 $S_4$ : CC-GCTAGCAG

$S_1$ : CCTGCTG-CAG  
 $S_2$ : GATG-TG-CCG  
 $S_3$ : GATG-TG-CAG  
 $S_4$ : CC-GCTAGCAG  
 $S_1$ : CCTGCT-GCAG  
 $S_5$ : CCTG-TAG--G

$S_1$ : CCTGCT-GCAG  
 $S_2$ : GATG-T-GCCG  
 $S_3$ : GATG-T-GCAG  
 $S_4$ : CC-GCTAGCAG  
 $S_5$ : CCTG-TAG--G