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

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

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

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

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

$$= \frac{1}{2} \sum_{i=1}^{k} \sum_{j=1}^{k} d_M(i, j)$$ $d_M(X,Y) \leq d_M(X,Z) +$ $d_M(X,Y)$

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

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

$$= k \sum_{j} D(S_c, S_j)$$

Center star method

Example:

Find center sequence S_1 $O(k^2)$

S_1: CCTGCTGCAG
S_2: GATG-TGCCG

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

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

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

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

(a)   (b)   (c)   (d)   (e)

Five input sequences

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

Pairwise alignment between S_1 and other sequences $O(kn^2)$

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