

MA3259 Lecture 6

Approximating MSA in Distance Measure

LX Zhang

Department of Mathematics
National University of Singapore

matzlx@nus.edu.sg

1. Star Alignment Methods

Another progressive method called star alignment has the following two steps:

- Find the sequence that is most similar to the rest in terms of pairwise alignment score, the details of which is given below.
- Use the selected sequence as reference to merge all the optimal pairwise alignments between the reference and the rest to produce a multiple alignment.

For example, we consider the following 5 sequences:

$S_1 =$ attcggatt
 $S_2 =$ atccggatt
 $S_3 =$ atggaatttt
 $S_4 =$ atgttgtt
 $S_5 =$ agtcagg

Step 1: We first compute all the optimal pairwise alignment scores $S(i, j)$, and select the sequence S_k such that $\sum_{j \neq k} S(k, j)$ is the maximum. Assume

	S_1	S_2	S_3	S_4	S_5	$\sum_{j \neq k} S(k, j)$
S_1		14	-4	0	-6	4
S_2	14		-4	0	-8	2
S_3	-4	-4		0	-14	-22
S_4	0	0	0		-6	-6
S_5	-6	-8	-14	-6		-34

Hence, S_1 is selected.

Step 2. Merge all optimal alignments between S1 and each of the rest.
Assume they are

S_1 : a t t c g g a t t

S_2 : a t c c g g a t t

S_1 : a t t c g g a t t - -

S_3 : a t g - g a a t t t t

S_1 : a t t c g g a t t

S_4 : a t g t t g - t t

S_1 a t t c g g a t t

S_5 a g t c a g g - -

We start with the alignment between S1 and S2.

Step 3.1: Add S3 using its alignment to S1.

S_2 : a t c c g g a t t - -

S_1 : a t t c g g a t t - -

S_3 : a t g - g a a t t t t

The last two t's in S3 are aligned with two spaces in the alignment between S1 and S3. So, we add two spaces in S2.

Step 3.2: Add S4 in the alignment in Step 3.1 using its alignment to S1.

S_2 : a t c c g g a t t - -
 S_1 : a t t c g g a t t - -
 S_3 : a t g - g a a t t t t

- -

S_1 : a t t c g g a t t
 S_4 : a t g t t g - t t

Step 3.3: Add S5 in the alignment in Step 3.2 using its alignment to S1.

S_2 : a t c c g g a t t - -
 S_3 : a t g - g a a t t t t
 S_1 : a t t c g g a t t - -
 S_4 : a t g t t g - t t - -

- -

S_1 a t t c g g a t t
 S_5 a g t c a g g - -

2. Performance of Star Alignment under Distance Measure

- Pairwise alignment score is a similarity measure.
 - Defined as the sum of all column scores
 - Each column score is given by a scoring scheme in which
 - match scores $f(x, x) > 0$
 - mismatch or indel scores
$$f(x, y) < 0, \text{ } y \text{ is a different letter or } -.$$
 - The higher the alignment score, the more similar the sequences are.

- Distance measures dissimilarity.
 - The larger the distance, the more different the sequences are
 - Distance of alignment is defined as the sum of distances for columns, each of which is given by a distance function satisfying
$$d(x, x) = 0, d(x, y) = d(y, x) > 0 \text{ for } x \neq y, \text{ and}$$
$$d(x, y) \leq d(x, z) + d(z, y).$$
for x, y is a letter or a space.
 - Hamming distance and edit distance for sequences

Hamming Distance

The Hamming distance between sequences S and S' is equal to the number of positions that differ in S and S' .

For example, the following two sequences

```
a t t g t c t
a c t c t c g
```

They differ in the 2nd, 3rd and 7th position and so the Hamming distance of them is 3.

Hamming distance, while important in information theory, is not useful for comparing DNA or protein sequences. Due to indels, it is unknown in advance whether the k -th letter in one sequence corresponds to the k -th letter in the other or not when two DNA sequences are compared.

Edit (also called Levenshtein) distance

The edit distance between two sequences is the minimum number of editing operations needed to transform one sequence into the other, where editing operations are

- insertion of a letter,
- deletion of a letter,
- substitution of one symbol for another.

t g c a t a t

↓

t g c a t a

↓

t g c a t

↓

a t g c a t

↓

a t c c a t

↓

a t c c g a t

delete last t

delete last a

insert a at the front

substitute c for g in the 3rd position

insert g before the last a

t g c a t a t

↓

a t g c a t a t

↓

a t g c a a t

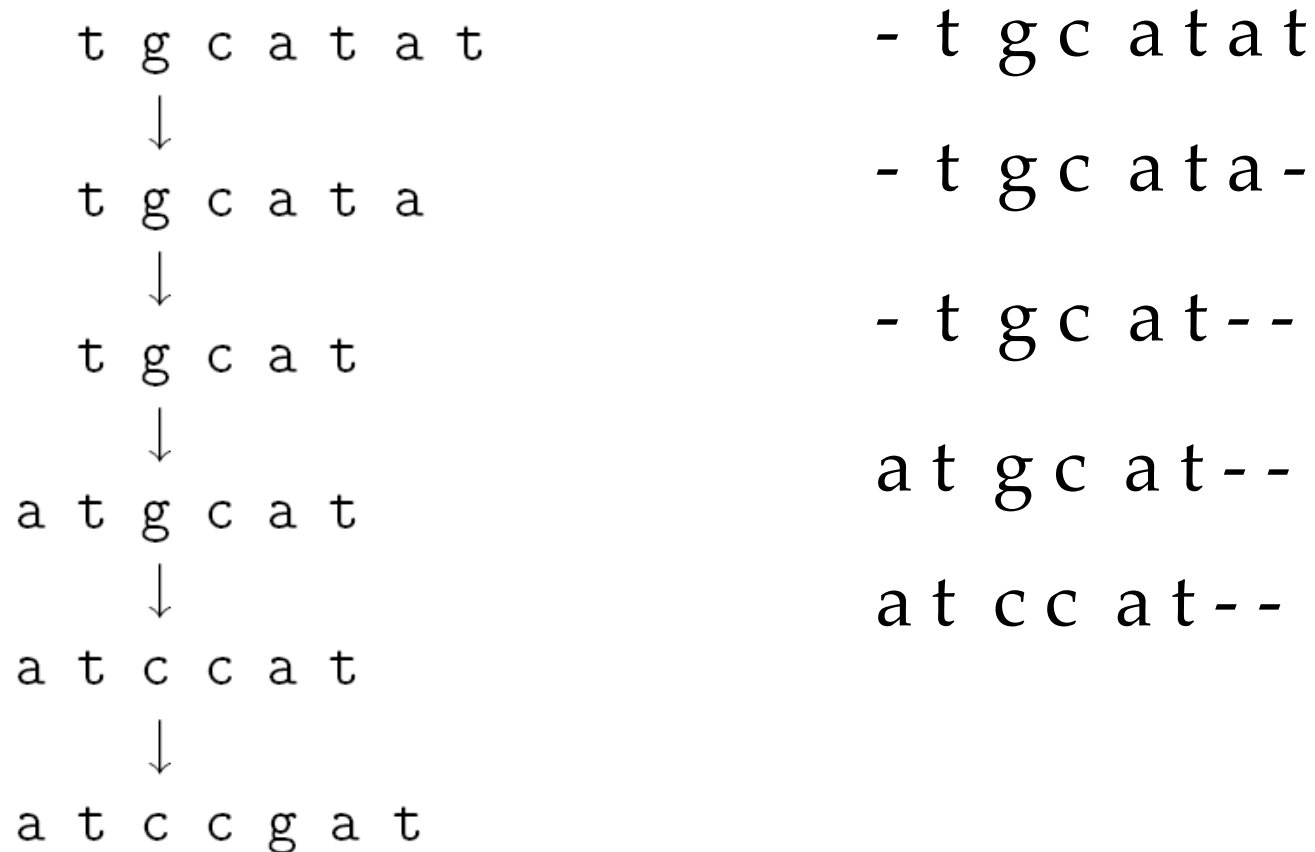
↓

a t g c g a t

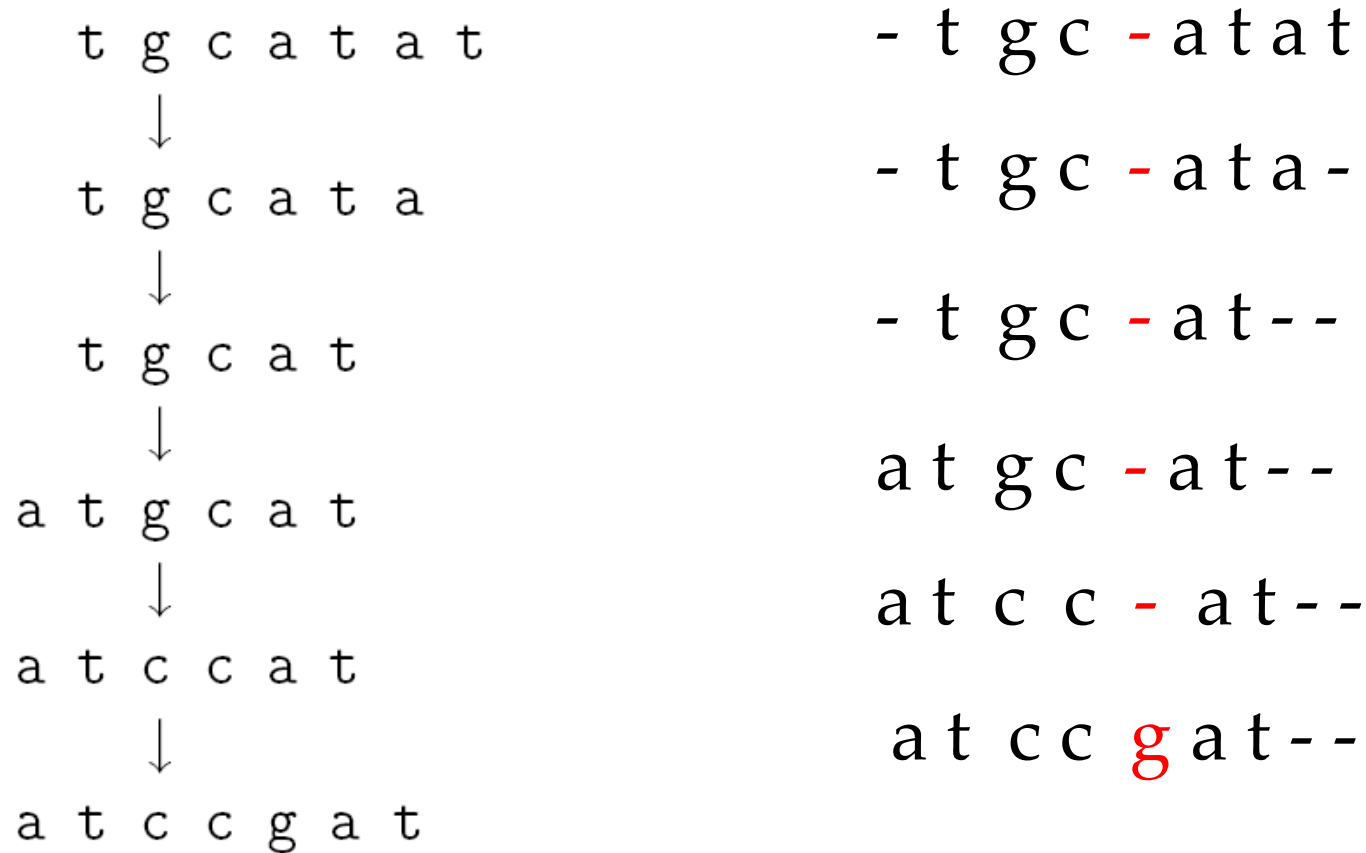
↓

a t c c g a t

Theorem. For any two sequences S and S', their edit distance is equal to the number of mismatches plus the number of indels in an optimal alignment obtained with the scoring scheme that scores match 1 and mismatch/indels -1.



Theorem. For any two sequences S and S', their edit distance is equal to the number of mismatches plus the number of indels in an optimal alignment obtained with the scoring scheme that scores match 1 and mismatch/indels -1.



Alignment with distance measure

Given a distance function $d(,)$ that specifies the distance for mismatch and indel.

-- The distance of an pairwise alignment \mathcal{A} of m columns is

$$D(\mathcal{A}) = \sum_{1 \leq j \leq m} d(a_j, b_j),$$

where a_j and b_j are letters in the j th column of \mathcal{A} .

-- The distance of a MSA is computed using the Sum of Pairs (SP) method.

Distance version of Alignment

Instance: Two or more DNA or Protein sequence.

Solution: An optimal alignment that has the minimum distance score.

Theorem Aligning two sequences with distance can be done in quadratic time by using dynamic approach.

$$D(i, j) = \min \begin{cases} D(i - 1, j) + d(x_i, -) \\ D(i, j - 1) + d(-, y_j) \\ D(i - 1, j - 1) + d(x_i, y_j) \end{cases}$$

Star Alignment with Distance

Input: k sequences S_1, S_2, \dots, S_k .

Step 1: Find the sequence S_i that minimizes $\sum_{j \neq i} D(S_i, S_j)$. For simplicity, we assume S_1 is selected.

Step 2: Progressively align sequences one by one:

-- Assume S_1, S_2, \dots, S_{i-1} are aligned.

We use an optimal alignment of S_1 and S_i to add S_i into the existing alignment.

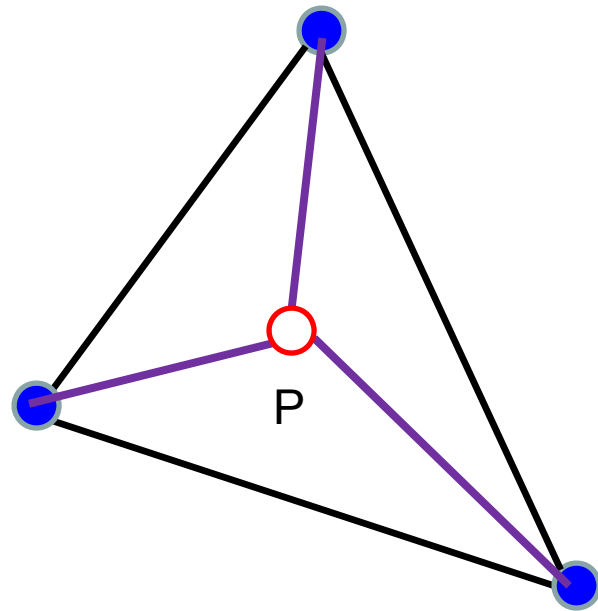
Fermat Problem & Steiner Problem

Fermat Problem:

For three points in a plane, find a point P such that the sum of the distances from P to the given points is minimal.

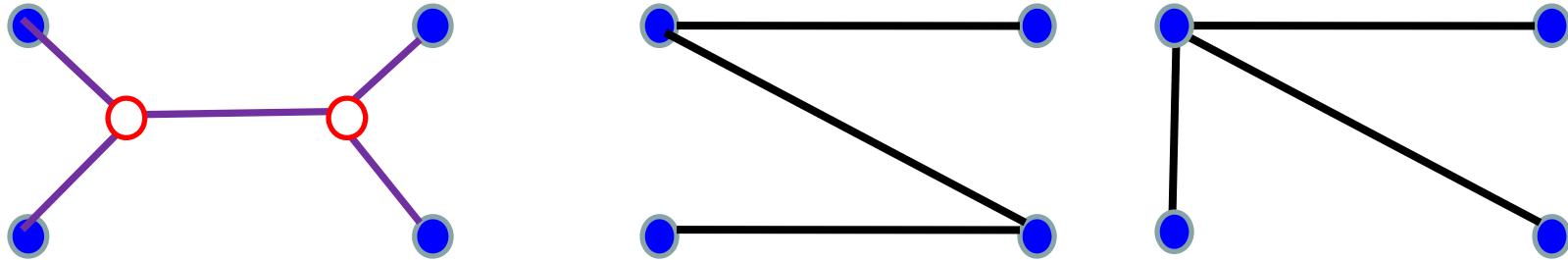


Pierre de Fermat (1601-1665)



What if there are more than 3 points?

Adding more points in the middle may give better solution.



Steiner Problem: For a set S of points in the Euclidean plane, find a shortest network connecting all the given points by adding new points.

Let the best solution for the Steiner problem have cost $W_{\text{opt}}(S)$

And let the best solution without adding points have cost $W_{\text{spanning}}(S)$.

$$W_{\text{opt}}(S) \leq W_{\text{spanning}}(S)$$

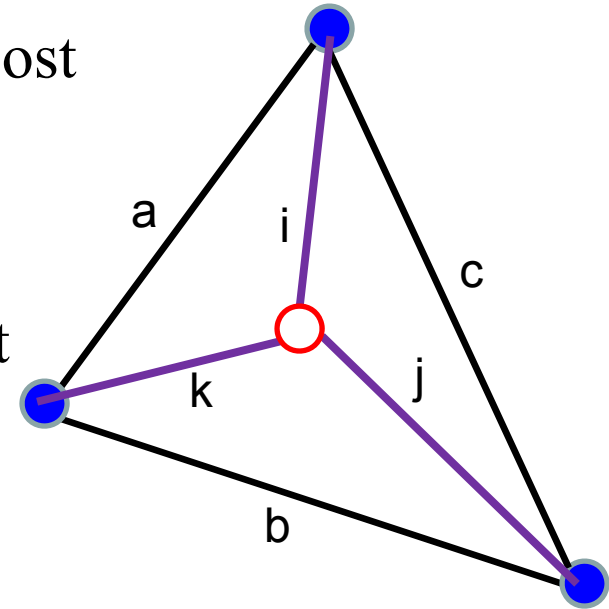
$$W_{\text{spanning}}(S) \leq C W_{\text{opt}}(S) \text{ for a small } C > 1?$$

Gilbert – Pollark Conjecture: $C = \frac{2}{\sqrt{3}}$ is the smallest bound.

Fact: $W_{\text{spanning}}(S) \leq (4/3) W_{\text{opt}}(S)$ for $n=3$

- The best solution without new points has cost $W_{\text{spanning}}(S) = a + b$.
(which is the sum of the two shortest edges).

- The best solution with new points has cost $W_{\text{opt}}(S) = i+j+k$.



- By the triangular inequality,
 $a \leq i + k$,
 $b \leq j + k$,
 $c \leq i + j$.

- Summing these inequalities together, we obtain $a+b+c \leq 2(i+j+k)$

- Multiplying the above inequality by 2, we have

$$3 W_{\text{spanning}}(S) \leq (a+b) + (b+c) + (a+c) \leq 4(i+j+k) = 4 W_{\text{opt}}(S)$$

Let \mathcal{M} be the multiple sequence alignment produced by the **star** algorithm on input sequences S_1, S_2, \dots, S_k . We also let \mathcal{M}^* be an optimal multiple alignment of these k sequences.

Theorem: The distance $d(\mathcal{M})$ satisfies:

$$d(\mathcal{M}^*) \leq d(\mathcal{M}) \leq (2(k-1)/k) d(\mathcal{M}^*)$$

Proof.

Let $D(i, j)$ denote the distance score of the optimal alignment of S_j and S_i

Since we assume S_1 as the center,

$$\sum_{j \neq i} D(i, j) \geq \sum_{j \neq 1} D(1, j) \quad \text{for } i=1, 2, \dots, k \quad \text{(Formula A)}$$

Let $d_{\mathcal{M}}(i, j)$ denote the distance score of the pairwise alignment of S_j and S_i induced by \mathcal{M} .

$$\begin{aligned} \text{Then, } d_{\mathcal{M}}(1, j) &= D(1, j), \\ d_{\mathcal{M}}(i, j) &\leq d_{\mathcal{M}}(1, i) + d_{\mathcal{M}}(1, j) = D(1, i) + D(1, j) \quad \text{(Formula B)} \end{aligned}$$

Let $d_{\mathcal{M}^*}(i, j)$ denote the distance score of the pairwise alignment of S_j and S_i induced by \mathcal{M}^* .

$$\text{Then, } d_{\mathcal{M}^*}(i, j) \geq D(i, j) \quad \text{(Formula C)}$$

$$\begin{bmatrix} 0 & d_{\mathcal{M}^*}(1, 2) & d_{\mathcal{M}^*}(1,3) & d_{\mathcal{M}^*}(1,4) & \dots & d_{\mathcal{M}^*}(1,k) \\ d_{\mathcal{M}^*}(2, 1) & 0 & d_{\mathcal{M}^*}(2, 3) & d_{\mathcal{M}^*}(2,4) & \dots & d_{\mathcal{M}^*}(2,k) \\ d_{\mathcal{M}^*}(3, 1) & d_{\mathcal{M}^*}(3, 2) & 0 & d_{\mathcal{M}^*}(3,4) & \dots & d_{\mathcal{M}^*}(3,k) \\ d_{\mathcal{M}^*}(4, 1) & d_{\mathcal{M}^*}(4, 2) & d_{\mathcal{M}^*}(4,3) & 0 & \dots & d_{\mathcal{M}^*}(4,k) \\ \dots & \dots & \dots & \dots & \dots & \dots \\ d_{\mathcal{M}^*}(k, 1) & d_{\mathcal{M}^*}(k, 2) & d_{\mathcal{M}^*}(k,3) & d_{\mathcal{M}^*}(k,4) & \dots & d_{\mathcal{M}^*}(k,k) \end{bmatrix}$$

By (Formula A) and (Formula C)

$$2d(\mathcal{M}^*) = \sum_{1 \leq i \leq k} \sum_{j \neq i} d_{\mathcal{M}^*}(i, j) \geq \sum_{1 \leq i \leq k} \sum_{j \neq i} D(i, j) \geq k \sum_{j \neq 1} D(1, j)$$

By (Formula B)

$$\begin{aligned} 2d(\mathcal{M}) &= \sum_{1 \leq i \leq k} \sum_{j \neq i} d_{\mathcal{M}}(i, j) \leq \sum_{1 \leq i \leq k} \sum_{j \neq i} [D(1, i) + D(1, j)] \\ &= 2(k-1) \sum_{j \neq 1} D(1, j) \end{aligned}$$

$$d(\mathcal{M}) \leq (k-1) \sum_{j \neq 1} D(1, j) \leq (2(k-1)/k) d(\mathcal{M}^*)$$



Expand the last slide into three slides.

Use matrix terms to express the terms in the
about $d(M)$.