CS 466
Introduction to Bioinformatics
Lecture 6

Mohammed El-Kebir

September 11, 2020

Course Announcements

Instructor:

- Mohammed El-Kebir (melkebir)
- Office hours: Wednesdays, 3:15-4:15pm

TA:

- Sarah Christensen (sac2) – Mondays, 3-4pm
- Wesley Wei Qian (weiqian3) – Fridays, 9-10am

Homework 1 due 9/17 by 11:59pm

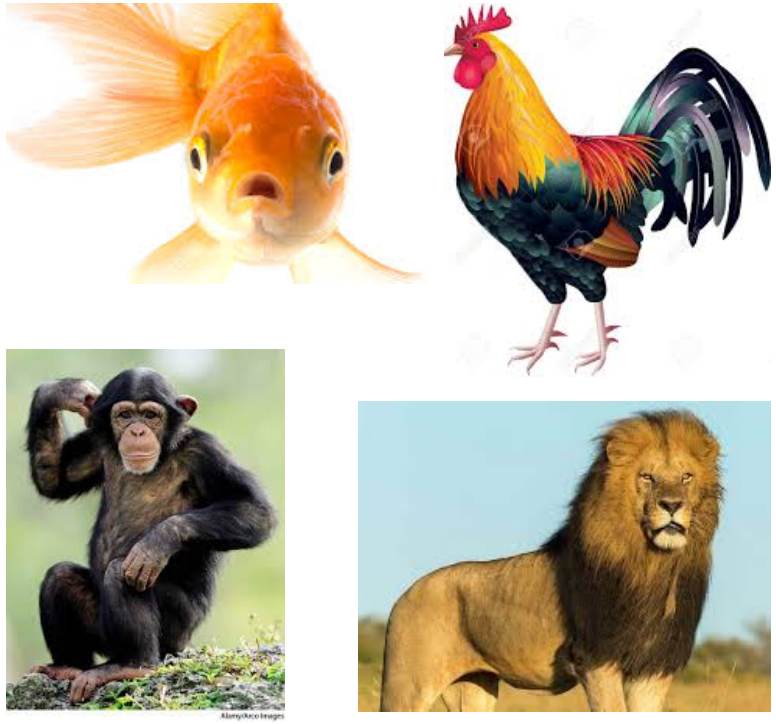
Outline

- Multiple sequence alignment
- Exact algorithm
- Sum-of-pairs (SP) score

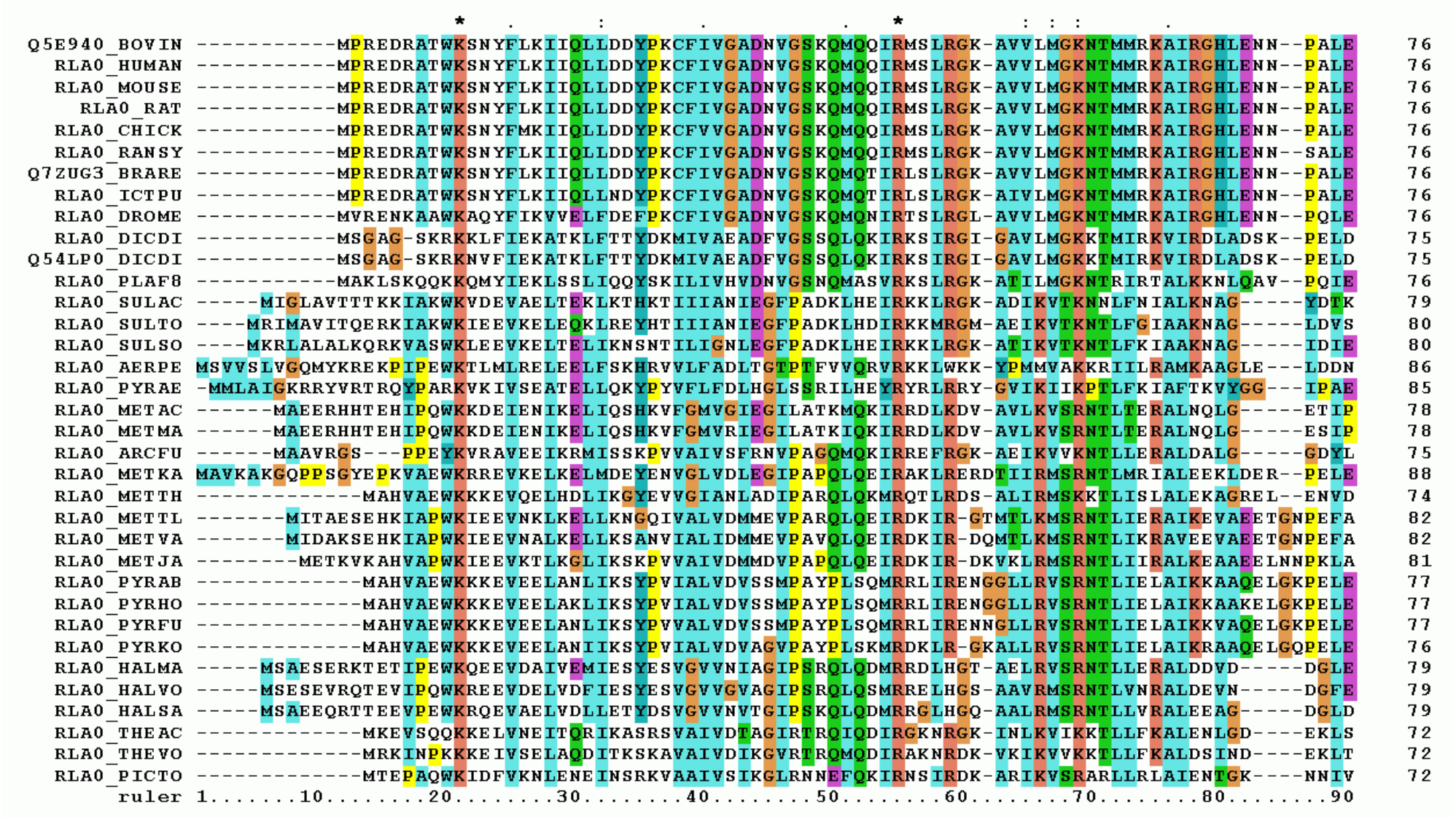
Reading:

- Jones and Pevzner. Chapter 6.10

Motivation



Simultaneous alignment of multiple (> 2) sequences enables inference of subtle similarities that are conserved in more than two species



"Pairwise alignment whispers ... multiple alignment shouts out loud".

Hubbard, Lesk, Tramontano, Nature Structural Biology 1996.

Multiple Sequence Alignment (MSA)

A **multiple sequence alignment** $\mathcal{M}$ between k strings $\mathbf{v}_1, \ldots, \mathbf{v}_k$ is a $k \times q$ matrix, where $q = \{\max\{\, |\mathbf{v}_i| : i \in [k] \,\}, \ldots, \sum_{i=1}^{k} |\mathbf{v}_i|\}$ such that the i-th row contains the characters of $\mathbf{v}_i$ in order with spaces '-' interspersed and no column contains k spaces

$\mathbf{v}_1$	A	T	–	G	C	G	–
$\mathbf{v}_2$	A	–	C	G	T	–	C
$\mathbf{v}_3$	A	T	C	A	C	–	A

Question: How to score a multiple sequence alignment?

Scoring a Multiple Sequence Alignment

$\mathbf{v}_1$	A	T	–	G	C	G	–
$\mathbf{v}_2$	A	–	C	G	T	–	C
$\mathbf{v}_3$	A	T	C	A	C	–	A

Question: How to score a multiple sequence alignment?

Pairwise scoring function:

$$\delta : (\Sigma \cup \{-\}) \times (\Sigma \cup \{-\}) \rightarrow \mathbb{R}$$

Scoring a Multiple Sequence Alignment

$\mathbf{v}_1$	A	T	–	G	C	G	–
$\mathbf{v}_2$	A	–	C	G	T	–	C
$\mathbf{v}_3$	A	T	C	A	C	–	A

Question: How to score a multiple sequence alignment?

Pairwise scoring function:

$$\delta : (\Sigma \cup \{-\}) \times (\Sigma \cup \{-\}) \rightarrow \mathbb{R}$$

k-wise scoring function:

$$\delta : (\Sigma \cup \{-\})^k \rightarrow \mathbb{R}$$

Outline

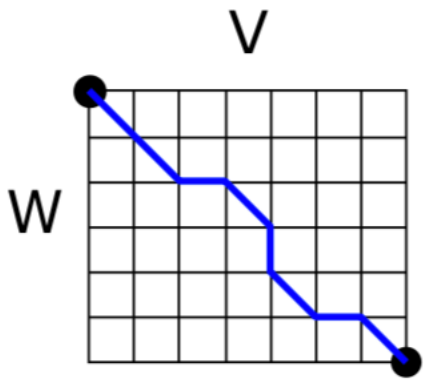
- Multiple sequence alignment
- Exact algorithm
- Sum-of-pairs (SP) score

Reading:

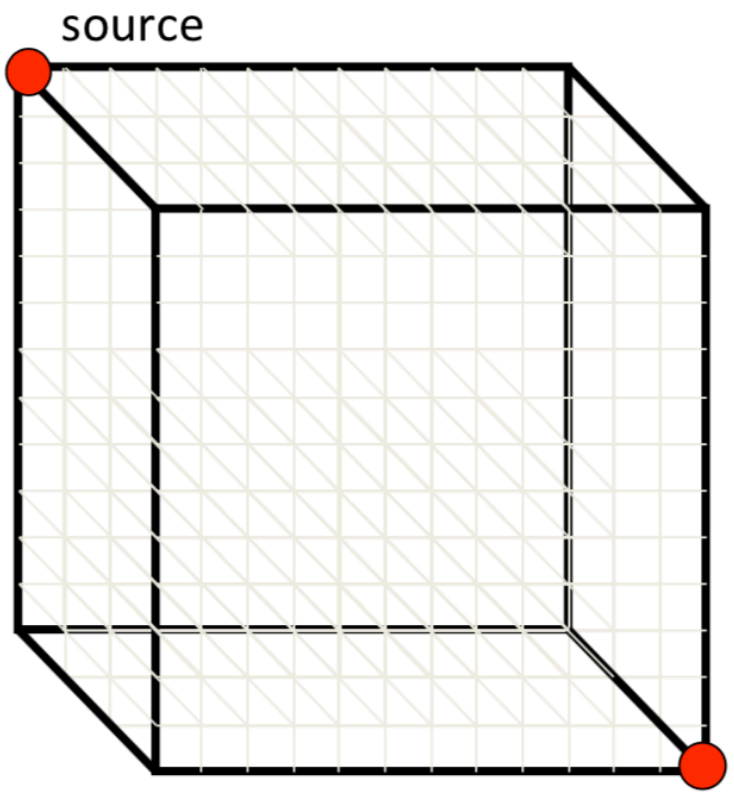
- Jones and Pevzner. Chapter 6.10

Aligning Three Sequences

- Same strategy as pairwise edit distance
- Use 3-D cube, with each axis representing an input sequence
- Alignment is a path from source to sink

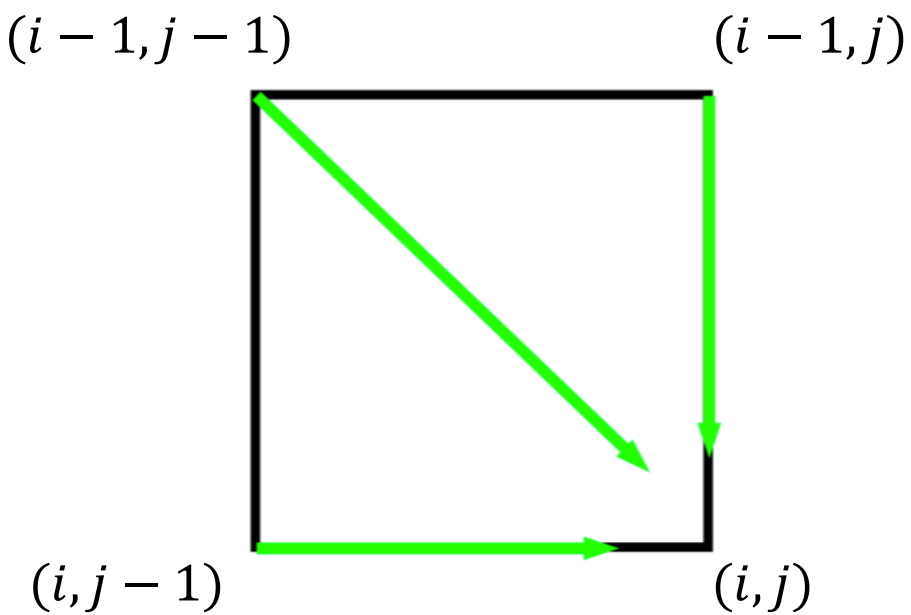


2-D Edit Graph

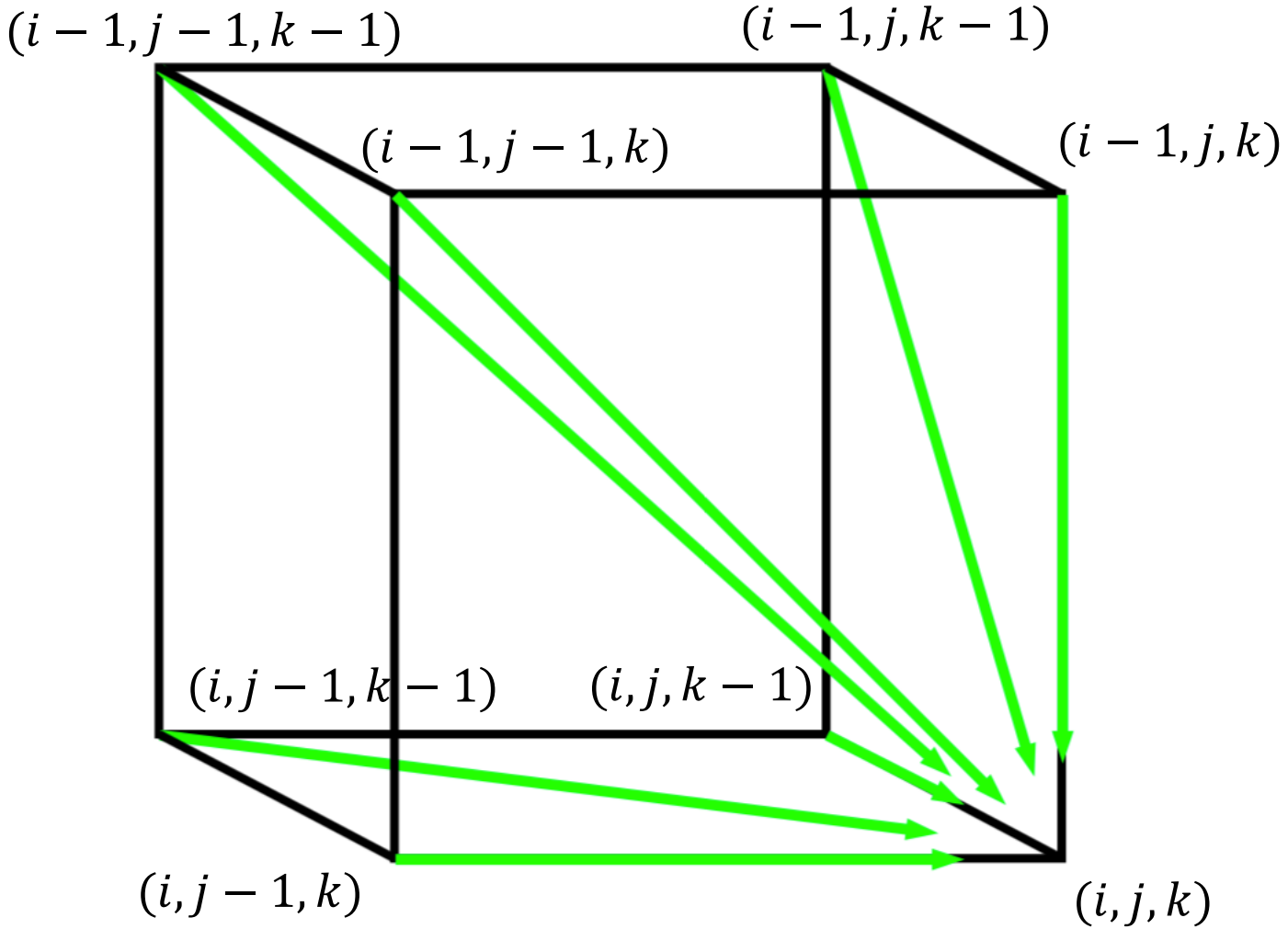


3-D Edit Graph

2-D vs 3-D Vertex Neighborhood



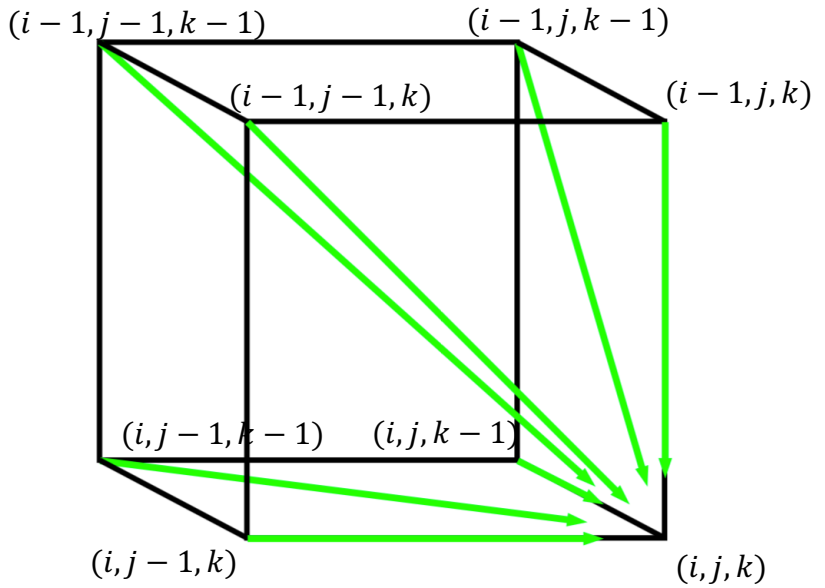
2-D Neighborhood
(3 edges)



3-D Neighborhood
(7 edges)

3-D Sequence Alignment

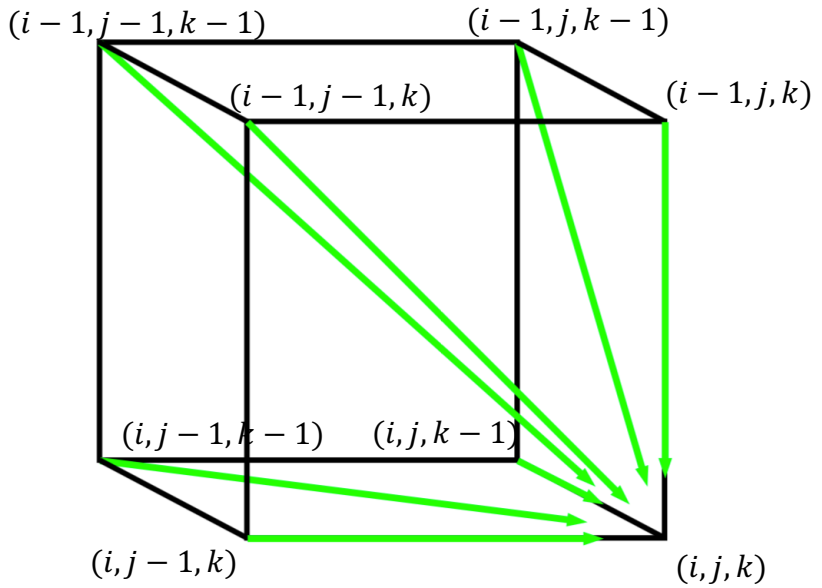
$\delta(x, y, z)$ is an entry in 3-D scoring matrix



3-D Sequence Alignment

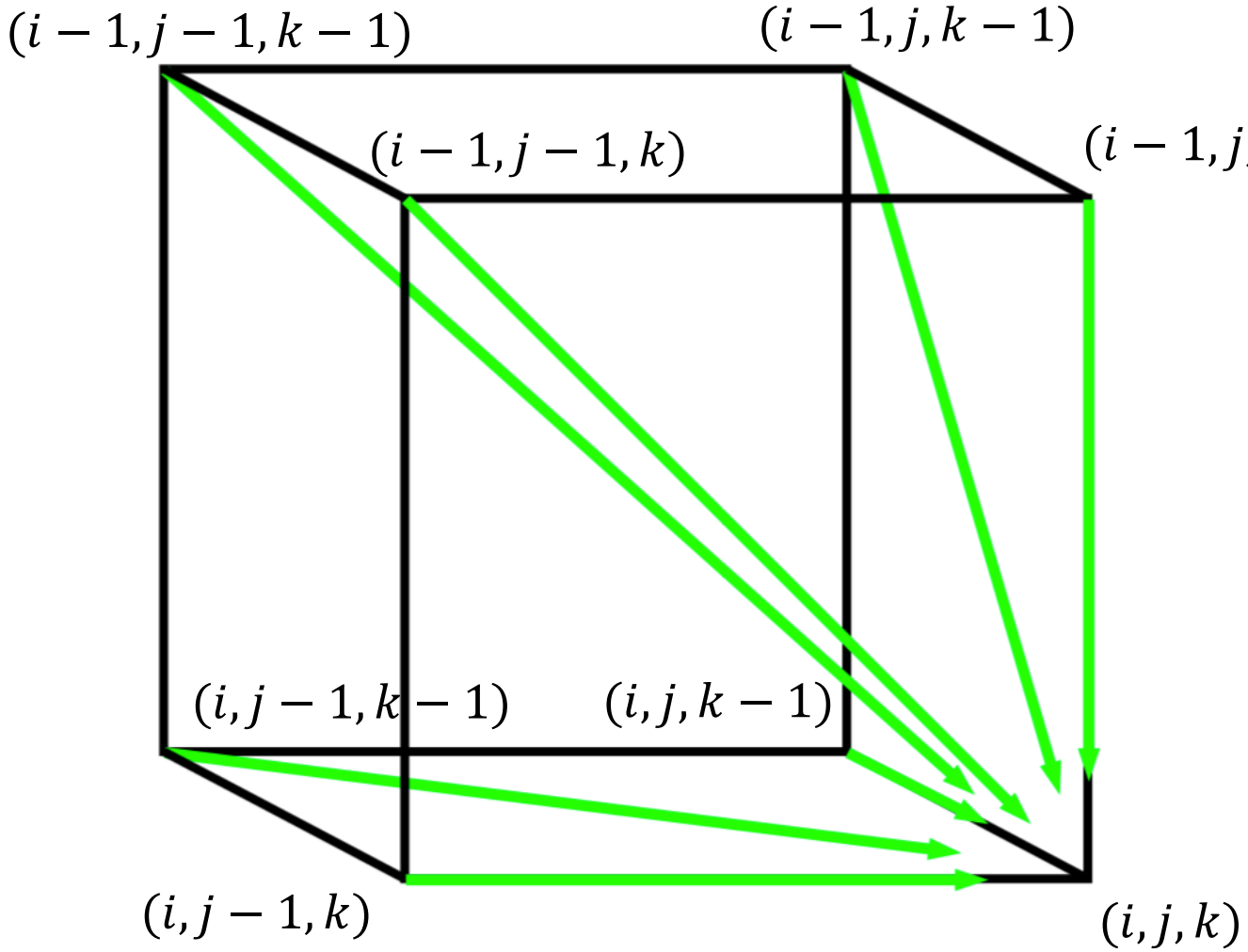
$\delta(x, y, z)$ is an entry in 3-D scoring matrix

Given three sequences each of length n, running time: $O(n^3)$



$$s[i,j,k] = \max \begin{cases} s[i-1,j-1,k-1] + \delta(v_i, w_j, u_k), & \text{no gaps} \\ s[i-1,j-1,k] + \delta(v_i, w_j, -), & \\ s[i-1,j,k-1] + \delta(v_i, -, u_k), & \text{one gap} \\ s[i,j-1,k-1] + \delta(-, w_j, u_k), & \\ s[i-1,j,k] + \delta(v_i, -, -), & \\ s[i,j-1,k] + \delta(-, w_j, -), & \text{two gaps} \\ s[i,j,k-1] + \delta(-, -, u_k), & \end{cases}$$

3-D vs k-D Vertex Neighborhood



3-D Neighborhood
(7 edges)

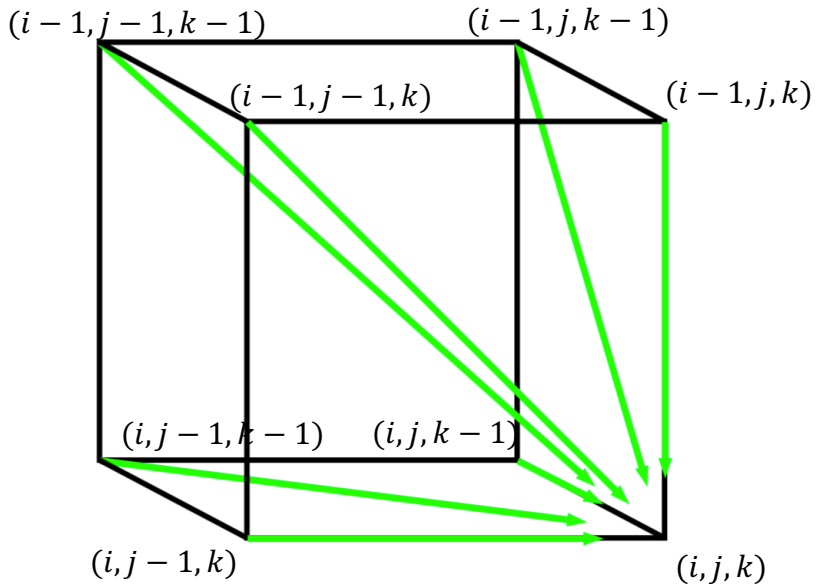
$$(i_1 - 1, i_2 - 1, \dots, i_{k-1} - 1, i_k - 1)$$
$$(i_1 - 1, i_2 - 1, \dots, i_{k-1} - 1, i_k)$$
$$\dots$$
$$(i_1, i_2 - 1, \dots, i_{k-1} - 1, i_k - 1)$$
$$\dots$$
$$(i_1 - 1, i_2, \dots, i_{k-1}, i_k)$$
$$\dots$$
$$(i_1, i_2, \dots, i_{k-1}, i_k - 1)$$

k-D Neighborhood
($2^k - 1$ edges)

k-D Sequence Alignment

$\delta(x_1, \ldots, x_k)$ is an entry in k-D scoring matrix

Given k sequences each of length n,
running time: $O(2^k n^k)$



$$s[i_1, i_2, \ldots, i_{k-1}, i_k] = \max \begin{cases} s[i_1-1, i_2-1, \ldots, i_{k-1}-1, i_k-1] + \delta(\mathbf{v_1}[i_1], \mathbf{v_2}[i_2], \ldots, \mathbf{v_{k-1}}[i_{k-1}], \mathbf{v_k}[i_k]) \\ s[i_1-1, i_2-1, \ldots, i_{k-1}-1, i_k] + \delta(\mathbf{v_1}[i_1], \mathbf{v_2}[i_2], \ldots, \mathbf{v_{k-1}}[i_{k-1}], -) \\ \quad \vdots \\ s[i_1, i_2-1, \ldots, i_{k-1}-1, i_k-1] + \delta(-, \mathbf{v_2}[i_2], \ldots, \mathbf{v_{k-1}}[i_{k-1}], \mathbf{v_k}[i_k]) \\ \quad \vdots \\ s[i_1-1, i_2, \ldots, i_{k-1}, i_k] + \delta(\mathbf{v_1}[i_1], -, \ldots, -, -) \\ \quad \vdots \\ s[i_1, i_2, \ldots, i_{k-1}, i_k-1] + \delta(-, -, \ldots, -, \mathbf{v_k}[i_k]) \end{cases}$$

no gaps

one gap

$k-1$ gaps

Multiple Sequence Alignment – Running Time

Given 2 sequences each of length n, running time: $O(n^2)$

Given 3 sequences each of length n, running time: $O(n^3)$

Given k sequences each of length n, running time: $O(2^k n^k)$

Multiple Sequence Alignment – Running Time

Given 2 sequences each of length n, running time: $O(n^2)$

Given 3 sequences each of length n, running time: $O(n^3)$

Given k sequences each of length n, running time: $O(2^k n^k)$

Question: Can we align k sequences each of length n in time $O(\text{poly}(n))$?

Multiple Sequence Alignment – Running Time

Given 2 sequences each of length n, running time: $O(n^2)$

Given 3 sequences each of length n, running time: $O(n^3)$

Given k sequences each of length n, running time: $O(2^k n^k)$

Question: Can we align k sequences each of length n in time $O(\text{poly}(n))$?

Let's look at a wieldier scoring function

Outline

- Multiple sequence alignment
- Exact algorithm
- Sum-of-pairs (SP) score

Reading:

- Jones and Pevzner. Chapter 6.10

Multiple Alignment Induces Pairwise Alignments

v_1	A	T	–	G	C	G	–
v_2	A	–	C	G	T	–	C
v_3	A	T	C	A	C	–	A

v_1	A	T	–	G	C	G	–
v_2	A	–	C	G	T	–	C

v_1	A	T	–	G	C	G	–
v_3	A	T	C	A	C	–	A

v_2	A	–	C	G	T	–	C
v_3	A	T	C	A	C	A	

Resulting columns with -/- are removed

Sum-of-Pairs (SP) Score

$\mathbf{v}_1$	A	T	–	G	C	G	–
$\mathbf{v}_2$	A	–	C	G	T	–	C
$\mathbf{v}_3$	A	T	C	A	C	–	A

Multiple sequence alignment $\mathcal{M}$

$S(\mathbf{v}_i, \mathbf{v}_j)$ is score of induced pairwise alignment of sequences $(\mathbf{v}_i, \mathbf{v}_j)$

$\mathbf{v}_1$	A	T	–	G	C	G	–
$\mathbf{v}_2$	A	–	C	G	T	–	C

$\mathbf{v}_1$	A	T	–	G	C	G	–
$\mathbf{v}_3$	A	T	C	A	C	–	A

$\mathbf{v}_2$	A	–	C	G	T	–	C
$\mathbf{v}_3$	A	T	C	A	C	–	A

$$\text{SP-score}(\mathcal{M}) = \sum_{i=1}^{k} \sum_{j=i+1}^{k} S(\mathbf{v}_i, \mathbf{v}_j)$$

Sum-of-Pairs (SP) Score – Example

$\mathbf{v}_1$	A	T	G	–	C
$\mathbf{v}_2$	A	–	G	–	C
$\mathbf{v}_3$	A	T	C	C	C

Multiple sequence alignment $\mathcal{M}$

Match score: 3
Mismatch score: 1
Gap score: -1

Question: Calculate
SP-score$(\mathcal{M}) =$
$\sum_{i=1}^{k} \sum_{j=i+1}^{k} S(\mathbf{v}_i, \mathbf{v}_j)$

Sum-of-Pairs (SP) Score – Example

$\mathbf{v}_1$	A	T	G	–	C
$\mathbf{v}_2$	A	–	G	–	C
$\mathbf{v}_3$	A	T	C	C	C

Multiple sequence alignment $\mathcal{M}$

Match score: 3
Mismatch score: 1
Gap score: -1

Question: Calculate
SP-score$(\mathcal{M}) =$
$\sum_{i=1}^{k} \sum_{j=i+1}^{k} S(\mathbf{v}_i, \mathbf{v}_j)$

We can sum over scores for the columns, ignoring -/-

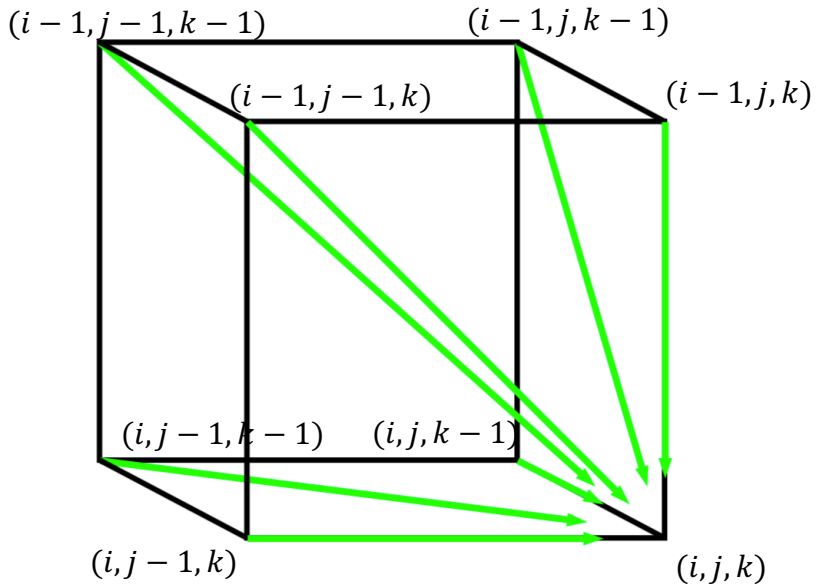
Multiple Sequence Alignment Problem w/ SP-Score

A **multiple sequence alignment** $\mathcal{M}$ between k strings $\mathbf{v}_1, \ldots, \mathbf{v}_k$ is a $k \times q$ matrix, where $q = \{\max\{|\mathbf{v}_i| : i \in [k]\}, \ldots, \sum_{i=1}^{k} |\mathbf{v}_i|\}$ such that the i-th row contains the characters of $\mathbf{v}_i$ in order with spaces '-' interspersed and no column contains k spaces

MSA-SP problem: Given strings strings $\mathbf{v}_1, \ldots, \mathbf{v}_k$ find multiple sequence alignment $\mathcal{M}^*$ with **minimum** value of SP-score$(\mathcal{M}^*) = \sum_{i=1}^{k} \sum_{j=i+1}^{k} S(\mathbf{v}_i, \mathbf{v}_j)$ where $S(\mathbf{v}_i, \mathbf{v}_j)$ is the score of the induced pairwise alignment of $(\mathbf{v}_i, \mathbf{v}_j)$ in $\mathcal{M}^*$

3-D MSA-SP

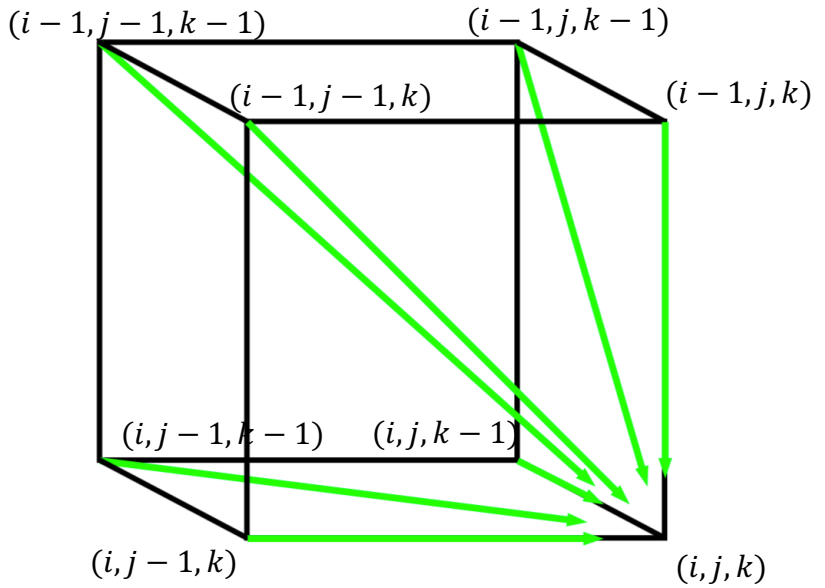
$\delta(x, y, z)$ is an entry in 3-D scoring matrix



3-D MSA-SP

$\delta(x, y, z)$ is an entry in 3-D scoring matrix

Given three sequences each of length n, running time: $O(n^3)$

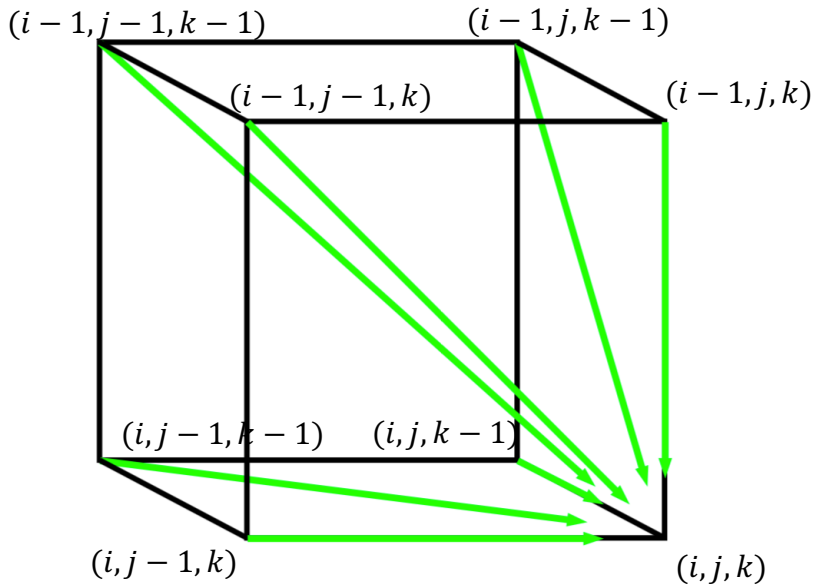


$$d[i_1, i_2, i_3] = \min \begin{cases} d[i_1-1, i_2-1, i_3-1] + \delta(\mathbf{v_1}[i_1], \mathbf{v_2}[i_2]) + \delta(\mathbf{v_1}[i_1], \mathbf{v_3}[i_3]) + \delta(\mathbf{v_2}[i_2], \mathbf{v_3}[i_3]) & \text{no gaps} \\ d[i_1-1, i_2-1, i_3] + \delta(\mathbf{v_1}[i_1], \mathbf{v_2}[i_2]) + 2\sigma \\ d[i_1-1, i_2, i_3-1] + \delta(\mathbf{v_1}[i_1], \mathbf{v_3}[i_3]) + 2\sigma \\ d[i_1, i_2-1, i_3-1] + \delta(\mathbf{v_2}[i_2], \mathbf{v_3}[i_3]) + 2\sigma & \text{one gap} \\ d[i_1-1, i_2, i_3] + 2\sigma \\ d[i_1, i_2-1, i_3] + 2\sigma \\ d[i_1, i_2, i_3-1] + 2\sigma & \text{two gaps} \end{cases}$$

k-D MSA-SP

Computing SP-score in each case: $O(k^2)$ time

Given k sequences each of length n, running time: $O(k^2 2^k n^k)$

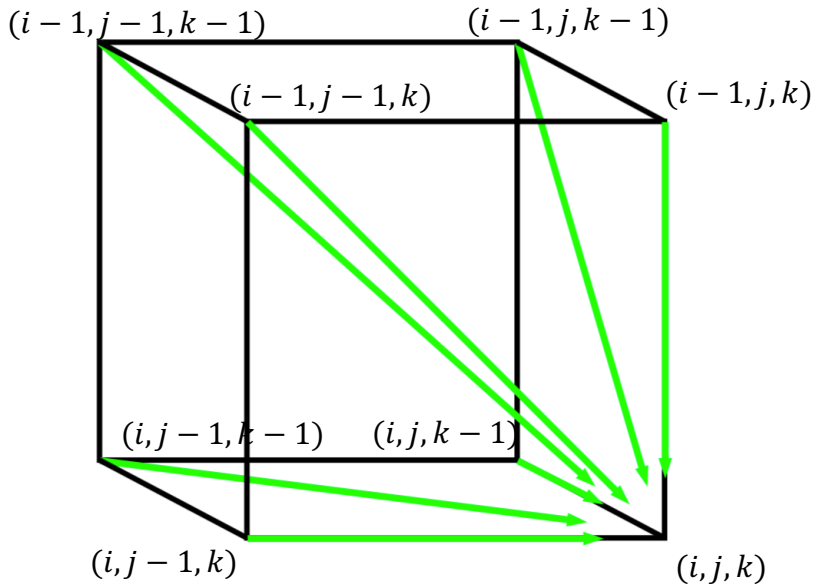


$$d[i_1, i_2, \ldots, i_{k-1}, i_k] = \min \begin{cases} s[i_1 - 1, i_2 - 1, \ldots, i_{k-1} - 1, i_k - 1] + \sum_{p=1}^{k} \sum_{q=p+1}^{k} \delta(\mathbf{v_p}[i_p], \mathbf{v_q}[i_q]) & \text{no gaps} \\ s[i_1 - 1, i_2 - 1, \ldots, i_{k-1} - 1, i_k] + (k-1)\sigma + \sum_{p=1}^{k-1} \sum_{q=p+1}^{k-1} \delta(\mathbf{v_p}[i_p], \mathbf{v_q}[i_q]) & \\ \quad \vdots & \\ s[i_1, i_2 - 1, \ldots, i_{k-1} - 1, i_k - 1] + (k-1)\sigma + \sum_{p=2}^{k} \sum_{q=p+1}^{k} \delta(\mathbf{v_p}[i_p], \mathbf{v_q}[i_q]) & \text{one gap} \\ \quad \vdots & \\ s[i_1 - 1, i_2, \ldots, i_{k-1}, i_k] + (k-1)\sigma & \\ \quad \vdots & k-1 \text{ gaps} \\ s[i_1, i_2, \ldots, i_{k-1}, i_k - 1] + (k-1)\sigma & \end{cases}$$

k-D MSA-SP

Computing SP-score in each case: $O(k^2)$ time

Given k sequences each of length n, running time: $O(k^2 2^k n^k)$



$$d[i_1, i_2, \ldots, i_{k-1}, i_k] = \min \begin{cases} s[i_1-1, i_2-1, \ldots, i_{k-1}-1, i_k-1] + \sum_{p=1}^{k} \sum_{q=p+1}^{k} \delta(\mathbf{v_p}[i_p], \mathbf{v_q}[i_q]) & \text{no gaps} \\ s[i_1-1, i_2-1, \ldots, i_{k-1}-1, i_k] + (k-1)\sigma + \sum_{p=1}^{k-1} \sum_{q=p+1}^{k-1} \delta(\mathbf{v_p}[i_p], \mathbf{v_q}[i_q]) & \\ \quad \vdots & \\ s[i_1, i_2-1, \ldots, i_{k-1}-1, i_k-1] + (k-1)\sigma + \sum_{p=2}^{k} \sum_{q=p+1}^{k} \delta(\mathbf{v_p}[i_p], \mathbf{v_q}[i_q]) & \text{one gap} \\ \quad \vdots & \\ s[i_1-1, i_2, \ldots, i_{k-1}, i_k] + (k-1)\sigma & \\ \quad \vdots & \\ s[i_1, i_2, \ldots, i_{k-1}, i_k-1] + (k-1)\sigma & k-1 \text{ gaps} \end{cases}$$

Question: How many cases have 2 gaps?

Multiple Sequence Alignment Problem w/ SP-Score

MSA-SP problem: Given strings strings $\mathbf{v}_1, \ldots, \mathbf{v}_k$ find multiple sequence alignment $\mathcal{M}^*$ with **minimum** value of SP-score$(\mathcal{M}^*) = \sum_{i=1}^{k} \sum_{j=i+1}^{k} S(\mathbf{v}_i, \mathbf{v}_j)$ where $S(\mathbf{v}_i, \mathbf{v}_j)$ is the score of the induced pairwise alignment of $(\mathbf{v}_i, \mathbf{v}_j)$ in $\mathcal{M}^*$

Question: Can we align k sequences each of length n in time $O(\text{poly}(n))$?

Multiple Sequence Alignment Problem w/ SP-Score

MSA-SP problem: Given strings strings $\mathbf{v}_1, \ldots, \mathbf{v}_k$ find multiple sequence alignment $\mathcal{M}^*$ with **minimum** value of SP-score$(\mathcal{M}^*) = \sum_{i=1}^{k} \sum_{j=i+1}^{k} S(\mathbf{v}_i, \mathbf{v}_j)$ where $S(\mathbf{v}_i, \mathbf{v}_j)$ is the score of the induced pairwise alignment of $(\mathbf{v}_i, \mathbf{v}_j)$ in $\mathcal{M}^*$

Question: Can we align k sequences each of length n in time $O(\text{poly}(n))$?

No, MSA-SP is NP-hard.

[WANG, L., & JIANG, T. (1994). On the Complexity of Multiple Sequence Alignment. *Journal of Computational Biology*, *1*(4), 337–348. http://doi.org/10.1089/cmb.1994.1.337]

Inverse Problem: From Pairwise to Multiple Alignment

v_1	A	C	G	C	T	G	G	-	C
v_2	A	C	G	C	-	-	G	A	G

v_1	A	C	-	G	C	T	G	G	-	C
v_3	G	C	C	G	C	A	-	G	A	G

v_2	A	C	-	G	C	-	G	A	G
v_3	G	C	C	G	C	A	G	A	G

Question: Can we construct a multiple alignment that induces the above three pairwise alignments?

Inverse Problem: From Pairwise to Multiple Alignment

v_1	A	C	G	C	T	G	G	-	C
v_2	A	C	G	C	-	-	G	A	G

v_1	A	C	-	G	C	T	G	G	-	C
v_3	G	C	C	G	C	A	-	G	A	G

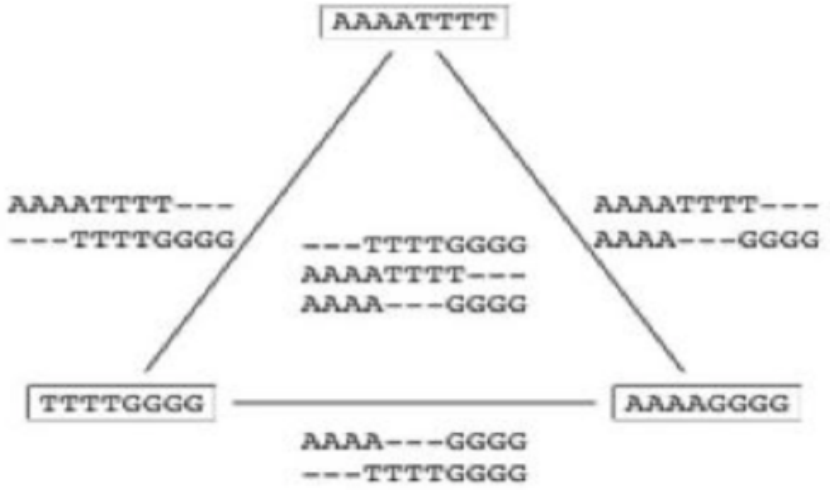
v_2	A	C	-	G	C	-	G	A	G
v_3	G	C	C	G	C	A	G	A	G

Question: Can we construct a multiple alignment that induces the above three pairwise alignments?

Not always!

Compatibility

Compatible: Pairwise alignments can be combined into multiple alignment



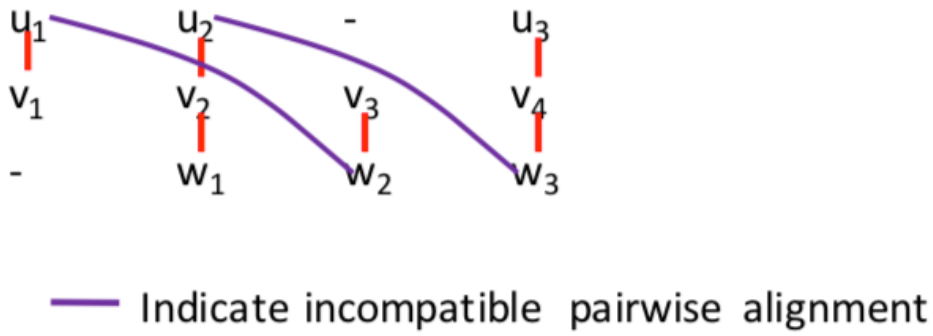
(a) Compatible pairwise alignments

Incompatible: Pairwise alignments *cannot* be combined into multiple alignment

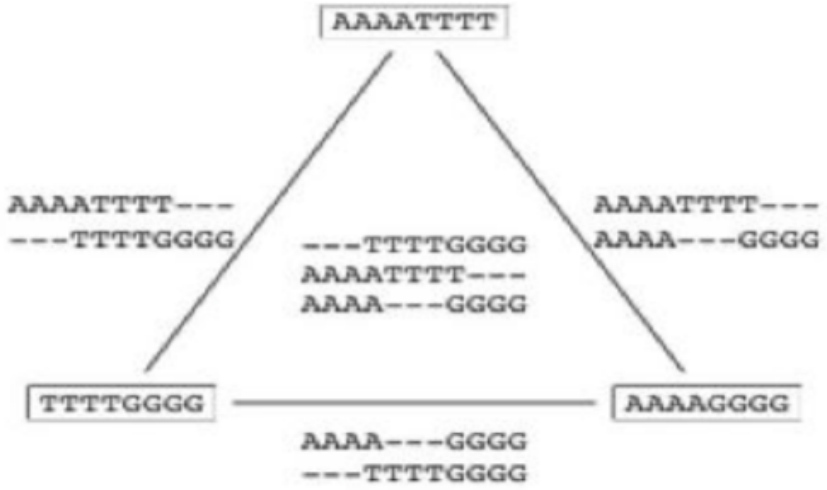
(b) Incompatible pairwise alignments

Compatibility

Compatible: Pairwise alignments can be combined into multiple alignment



Incompatible: Pairwise alignments *cannot* be combined into multiple alignment



(a) Compatible pairwise alignments

(b) Incompatible pairwise alignments

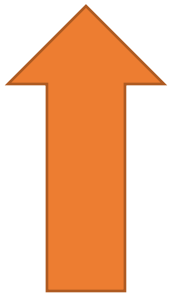
From Compatible Pairwise to Multiple Alignment

Optimal multiple alignment

Easy

Pairwise alignments between *all* pairs of sequences, but they are *not* necessarily optimal

(Sub)optimal multiple alignment



Challenging

Good (or optimal) *compatible* pairwise alignments between all sequences

Summary

- Multiple sequence alignment
- Exact algorithm
- Sum-of-pairs (SP) score

Reading:

- Jones and Pevzner. Chapter 6.10