Pairwise Sequence Alignments

Mark Stenglein, MIP 280A4

“Nothing in Biology Makes Sense Except in the Light of Evolution”

- Theodosius Dobzhansky

Image credit: Steve McCluskey CC BY-SA 3.0 Link

Biological sequences really only make sense when you compare them to each other

Sequence alignment is at the heart of a lot of bioinformatics and sequence analysis

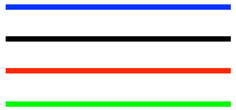
Major categories of sequence alignment

Alignment type	Purpose	Commonly-used software
Pairwise alignment	Identify the similarities or differences between two sequences	Needle (global alignment) Water (local alignment)
Multiple sequence alignment	Identify the similarities or differences between >2 sequences. Input to tree building.	MAFFT
Alignment-based search	Find the most closely related sequence in a database of sequences	BLAST
Mapping (alignment to reference)	Determine the most likely location in a reference sequence from which a shorter sequence (a read) derives	BWA Bowtie2
Assembly	Create a new reference sequence using overlapping reads	SPAdes

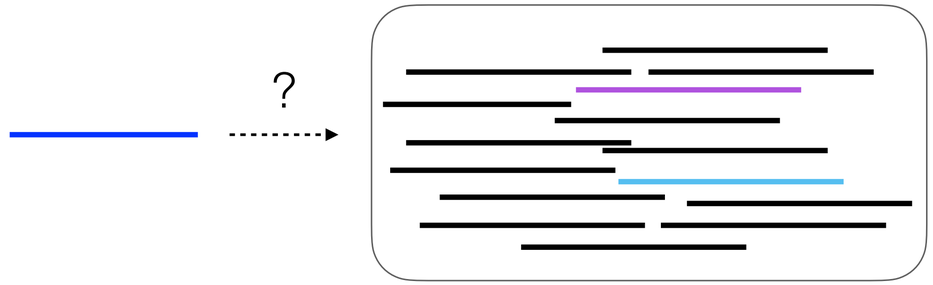
Major categories of sequence alignment

Pairwise alignment

Multiple sequence alignment

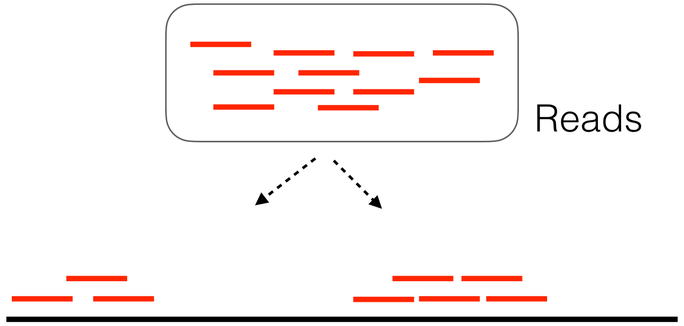


Alignment-based database searches



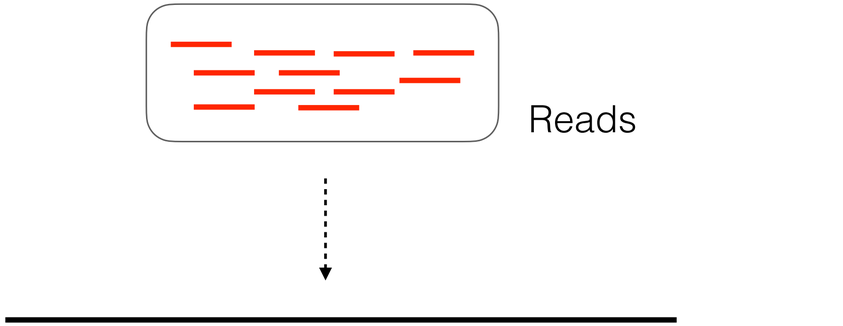
Database

Alignment to reference (mapping)



Existing Reference sequence

Assembly

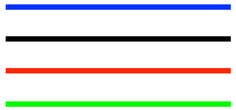


New Reference sequence

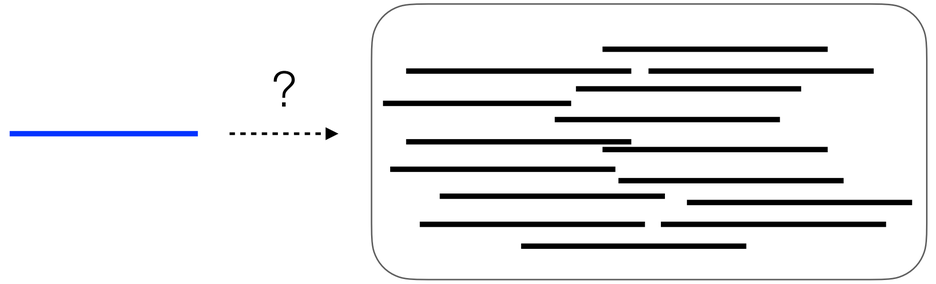
Today we'll be talking about pairwise alignments

Pairwise alignment

Multiple sequence alignment

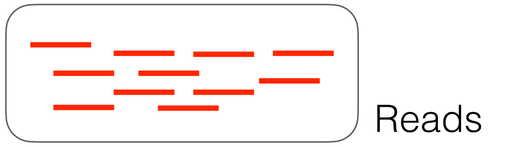


Alignment-based database searches

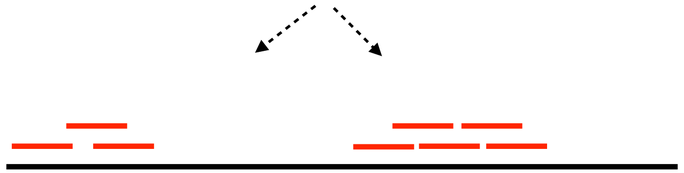


Database

Alignment to reference (mapping)

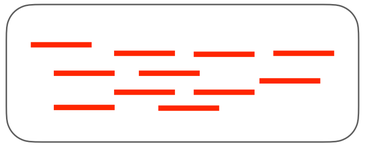


Reads

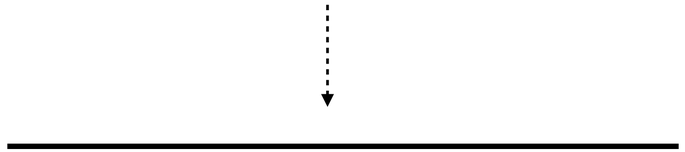


Existing Reference sequence

Assembly



Reads



New Reference sequence

Sequence alignment exercise

image: Stephen Altschul

There is not necessarily one definitively best way to align two sequences

```
groan      colo-r          theatre      theatre
|||:|      |||| |          |||||::      |||||x
grown      colour          theater      theater
```

```
elephant          vermiform     vermiform-----
|||: |||          ::||:::::          ||||
eleg-ant          formation     -----formation
```

```
disestablishment       disestablishment
|||         |  |||      |||          :|||
dis-------s--ent       dis---------sent
```

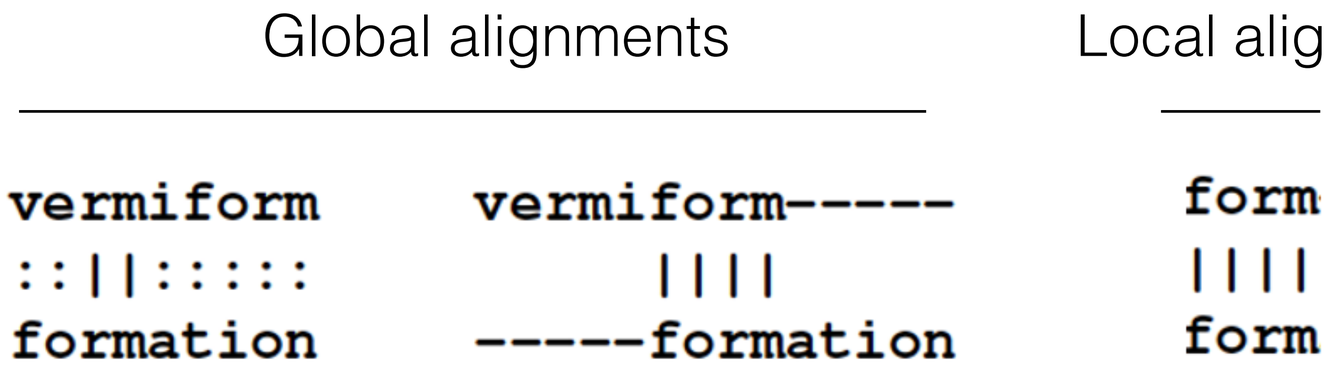
image: Stephen Altschul

The “optimal” pairwise alignment is a function of:

- The algorithm used
- Whether it’s a **global** or **local** alignment
- The **scoring system** used
- How **gaps** are handled

Global alignments force entire sequences to be aligned

Local alignments don’t force entire sequences to be aligned.



The “best” alignment is the alignment with the best score
given a particular scoring scheme

Which of these alignments should have a higher score?

```
vermiform
::||:::::
formation
```

Score = ?

```
vermiform-----
     ||||
-----formation
```

Score = ?

Exercise: Invent an alignment scoring system, and, *given that scoring system*, try to find the highest scoring alignment of these two sequences

```
Sequence 1        A G C A A C T T

Sequence 2        A G G C A A C T
```

Alignments are typically scored using rewards or penalties for matches, mismatches, and gaps

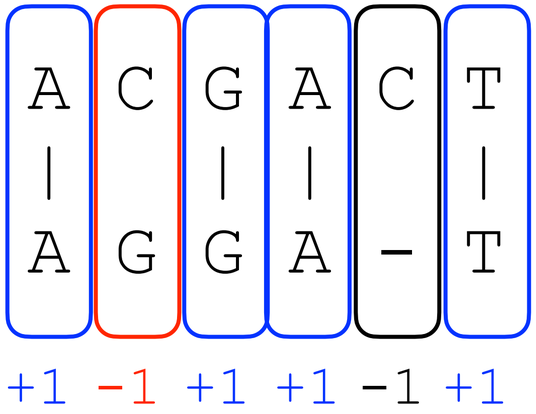
```
Sequence 1      A C G A C T
                |   | |   |
Sequence 2      A G G A - T
```

Match: +1 reward
Mismatch: -1 penalty
Gap: -1 penalty

What is the score of this alignment, given this scoring system?

Alignment algorithms score alignments using rewards and penalties for matches, mismatches, and gaps

```
Sequence 1   A C G A C T
             | |   | |
Sequence 2   A G   G A - T
```



Match: +1 reward
Mismatch: -1 penalty
Gap: -1 penalty

Score: +4 -1 -1 = 2

In local alignment, partial alignments can have higher scores than full-length alignments

Global alignments

```
vermiform
::||::::
formation
```

```
vermiform-----
     ||||
-----formation
```

Local alignment

```
form
||||
form
```

This local alignment might have the highest score of all these alignments (4 matches, 0 gaps, 0 mismatches)

In some global alignment scoring schemes, “end gaps” aren’t penalized

Global alignments

```
vermiform
::||::::
formation
```

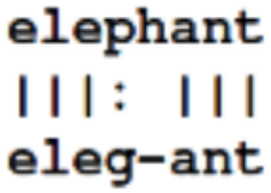
```
vermiform-----
     ||||
-----formation
```

Local alignment

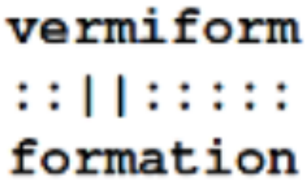
```
form
||||
form
```

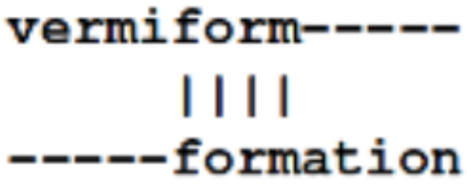
In some global alignment scoring schemes, end gaps aren’t penalized

An assumption of alignments is that sequences share ancestry (are homologous)



Good alignment, but not legitimate homology





Legitimate shared ancestry involving root “form”

What happens if you align random sequences?

Exercises:

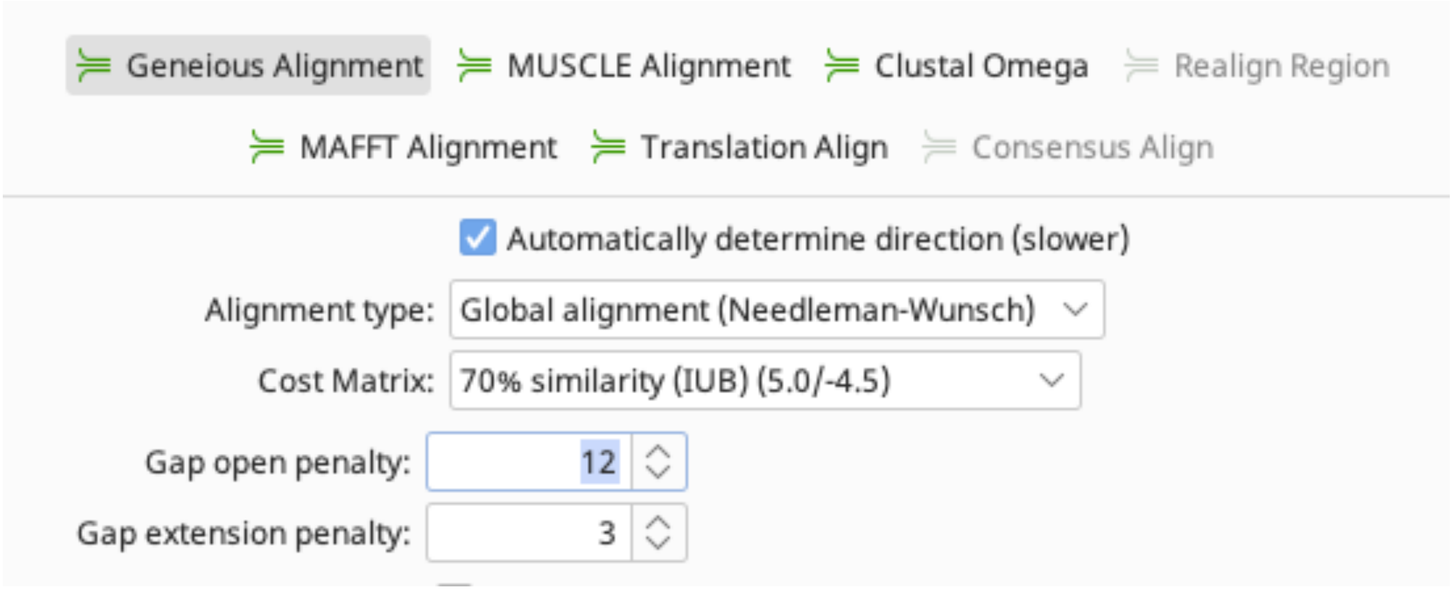
- Perform a local alignment of two randomly generated 200 bp sequences
- Perform a global alignment of two randomly generated 200 bp sequences

Generate random sequences here:

http://www.faculty.ucr.edu/~mmaduro/random.htm

Do the alignments in Geneious using the Geneious aligner.
Make sure “Automatically determine direction (slower)” is checked on.

There can be different “gap open” and “gap extension” penalties

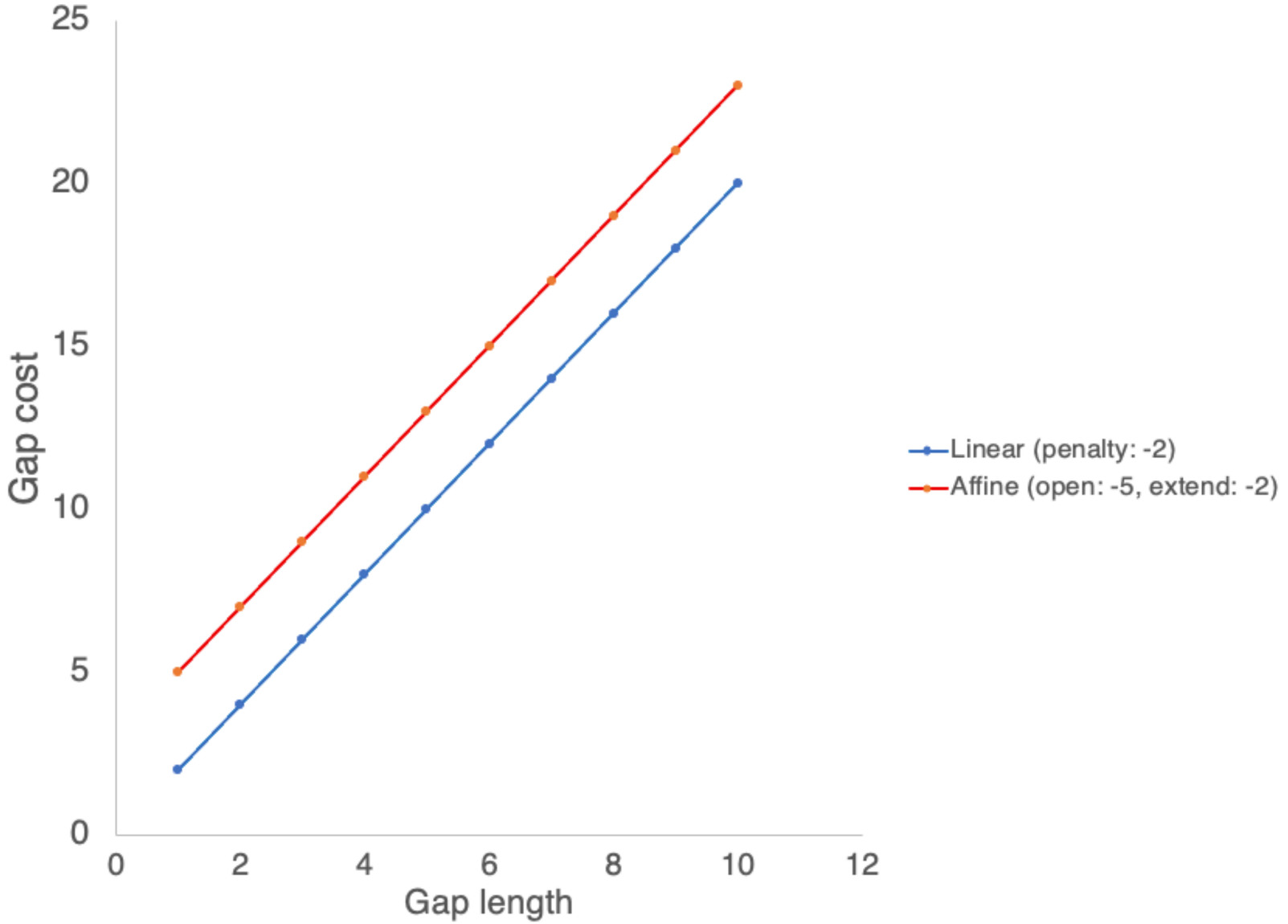


Linear gap penalties vs. “affine” gap penalties

Linear gap cost = gap length x gap penalty

Affine gap cost = gap open penalty + ((gap length-1) x gap extension penalty)

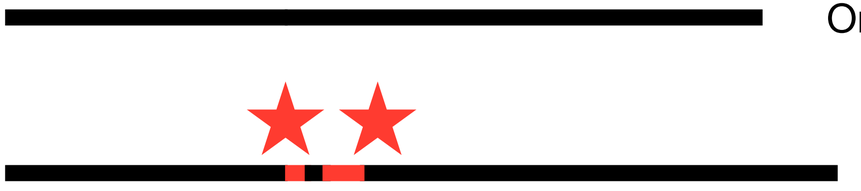
Affine gap scoring systems favor fewer, longer gaps



Affine gap scoring systems favor fewer, longer gaps

Original sequence

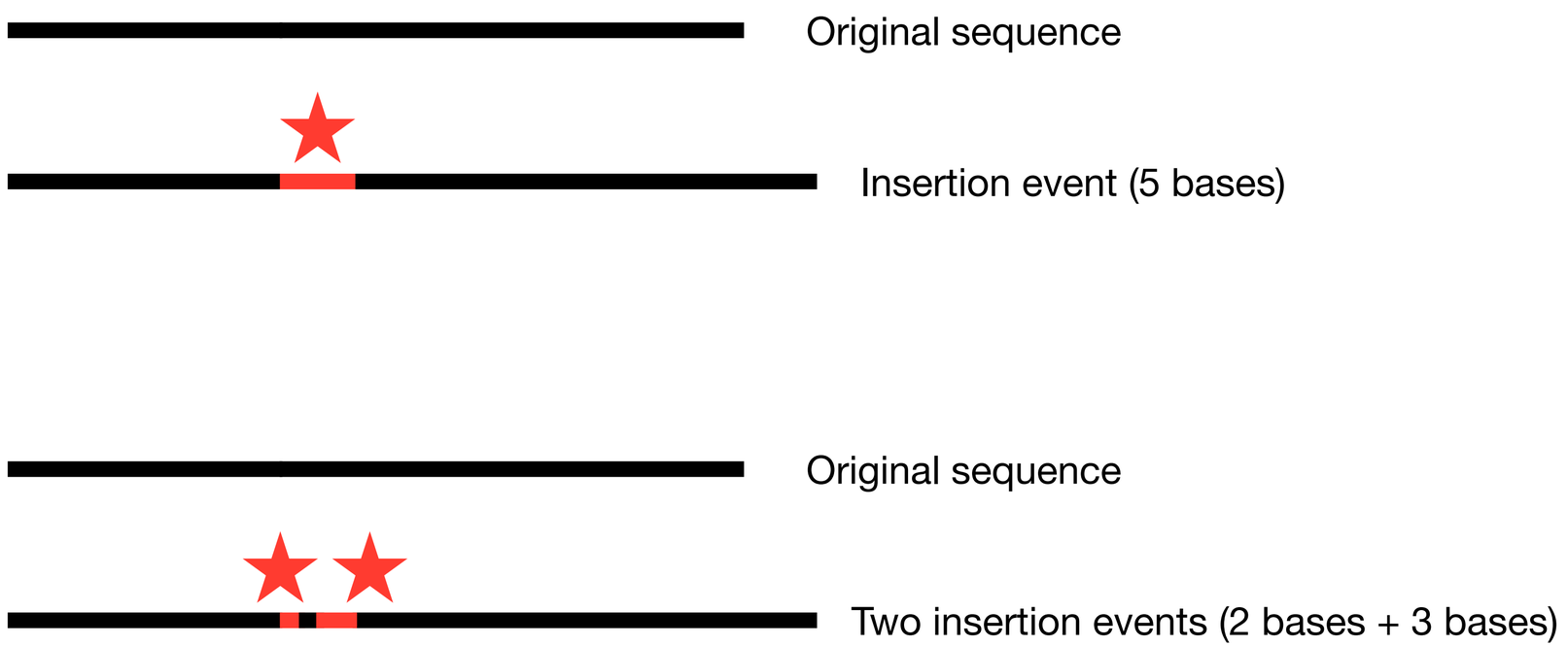
Insertion event (5 bases): penalty = 5 + (4x2) = 13



Original sequence

Two insertion events (2 bases + 3 bases): penalty = 5 + (1x2) + 5 + (2x2) = 16

The rationale for having separate gap open penalties is that
One insertion or deletion event of larger size is more likely than multiple insertions and deletions of smaller size near each other



Zika virus exercise