Wed, Oct 7

Lecture 4B:

Sequence alignment

Hibiscus lasiocarpus
© Matilda Adams/
Missouri Botanical Garden

Practical Bioinformatics (Biol 4220)
Instructor: Michael Landis
Email: michael.landis@wustl.edu

Lecture 4B outline

1. Sequence variation
2. Sequence alignment
3. Lab 4B overview

Sequence variation

Many questions in genome biology are
comparative

- *what amino acid differences cause two proteins to differ in function?*
- *where in the reference genome does this anonymous sequence belong?*
- *how are these genes related to each other?*

Sequence variation

- If sequences only changed by ***character substitutions*** then all sequences would be equal in length
- Length varies because ***insertions*** introduce new characters into a sequence, and ***deletions*** remove old characters from the sequence

```
CCAAGCGTTATC
TCAGTGGTATC
TAGTGGTATC
CTCAGTGGATC
```

Sequence variation

Sequence alignment

- However, any two molecular sequences are expected to differ in content and in length
- ***Sequence alignment*** defines which parts of the sequence are evolutionarily or functionally comparable (***homologous***)

```
  **    *  *
CCAAG-CGTTATC
-TCAGTGGT-ATC
-T-AGTGGT-ATC
CTCAGTGGA--TC
```

Sequence alignment

Sequence alignment is the task of constructing a data matrix from a set of sequences, where columns are *homologous* characters

Two dominant method families:

- ***Heuristic methods*** focus on maximizing "match" scores
- ***Generative methods*** focus on modeling how sequence variation arises, evolutionarily

Heuristic alignment

These methods minimize the ***cost*** of a proposed alignnment matrix

These methods insert ***gaps*** into sequences to ensure they all have the same length

Gaps are called ***indels*** when it cannot be determined whether sequence length variation was caused by an insertion on one sequence or a deletion on the other sequence

Pairwise example

```
TCAA--GGTATCGACCT
TCATGCGGTATT-ACCT
```

+1 Match (12)
−1 Mismatch (2)
−2 Gap open (2)
−1 Gap extension (1)

Needlemen-Wunsch example

		T	**G**	**A**	**C**
	0				
T					
C					
A					
T					
C					

Type	**Score**
Match	+1
Mismatch	-1
Gap	-2

S1: ?
S2: ?

Needlemen-Wunsch example

		T	G	A	C
	0 →	-2			
T	-2				
C					
A					
T					
C					

Type	Score
Match	+1
Mismatch	-1
Gap	-2

```
S1: ?
S2: ?
```

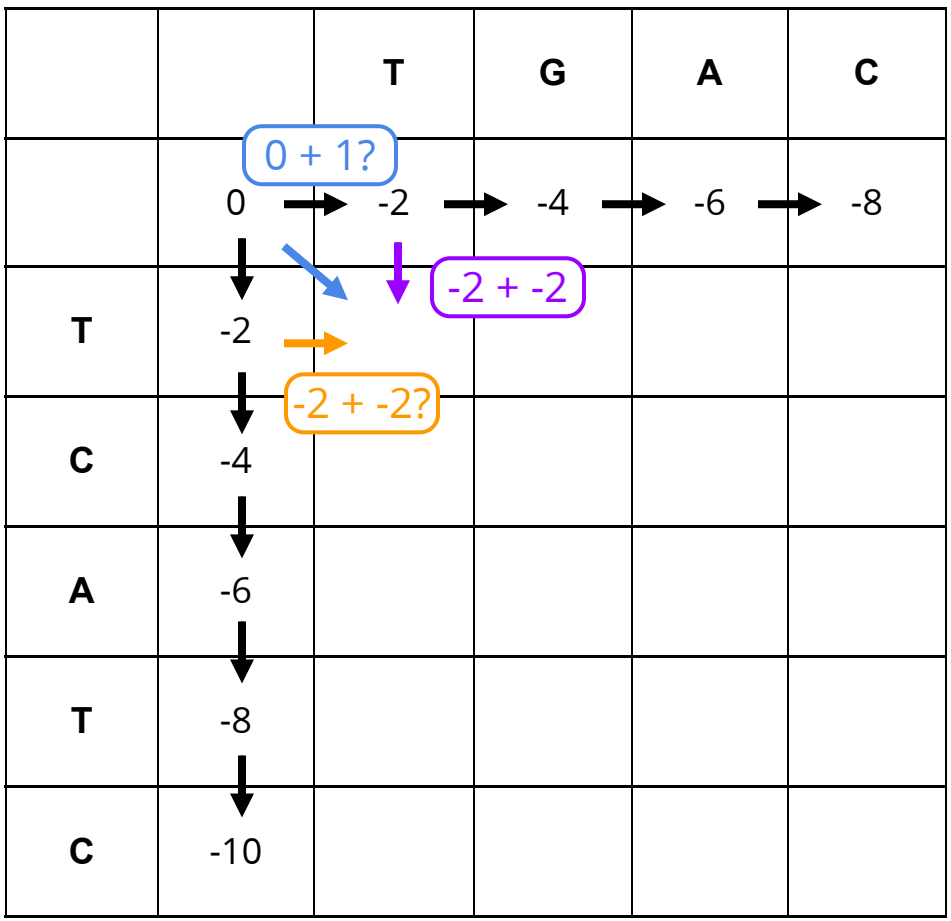
Needlemen-Wunsch example

		T	G	A	C
	0 →	-2 →	-4 →	-6 →	-8
T	-2				
C	-4				
A	-6				
T	-8				
C	-10				

Type	Score
Match	+1
Mismatch	-1
Gap	-2

```
S1: ?
S2: ?
```

Needlemen-Wunsch example



Type	Score
Match	+1
Mismatch	-1
Gap	-2

```
S1: ?
S2: ?
```

Needlemen-Wunsch example

		T	G	A	C
	0	-2	-4	-6	-8
T	-2	+1			
C	-4				
A	-6				
T	-8				
C	-10				

Type	Score
Match	+1
Mismatch	-1
Gap	-2

S1: ?
S2: ?

Needlemen-Wunsch example

		T	G	A	C
	0	-2	-4	-6	-8
T	-2	+1	-1	-3	-5
C	-4				
A	-6				
T	-8				
C	-10				

Type	Score
Match	+1
Mismatch	-1
Gap	-2

```
S1: ?
S2: ?
```

Needlemen-Wunsch example

		T	G	A	C
	0	-2	-4	-6	-8
T	-2	+1	-1	-3	-5
C	-4	-1	0	-2	-2
A	-6				
T	-8				
C	-10				

Type	Score
Match	+1
Mismatch	-1
Gap	-2

S1: ?
S2: ?

Needlemen-Wunsch example

		T	G	A	C
	0	-2	-4	-6	-8
T	-2	+1	-1	-3	-5
C	-4	-1	0	-2	-2
A	-6	-3	-2	+1	-1
T	-8				
C	-10				

Type	Score
Match	+1
Mismatch	-1
Gap	-2

```
S1: ?
S2: ?
```

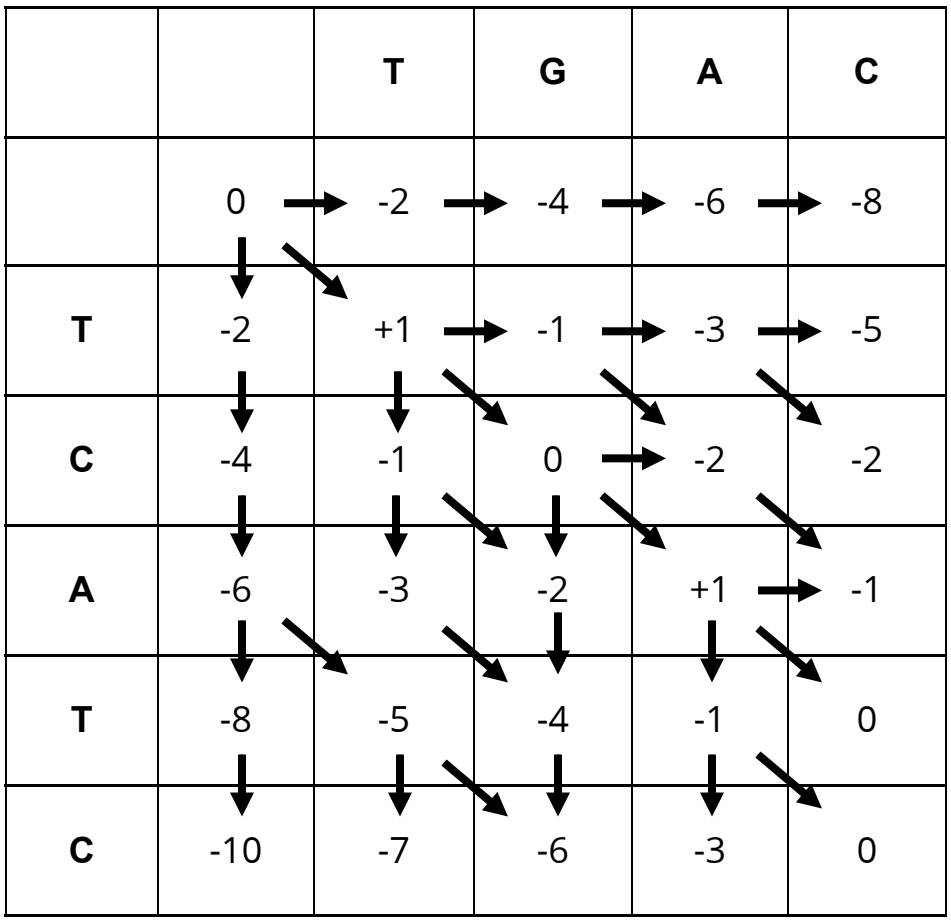
Needlemen-Wunsch example

		T	G	A	C
	0	-2	-4	-6	-8
T	-2	+1	-1	-3	-5
C	-4	-1	0	-2	-2
A	-6	-3	-2	+1	-1
T	-8	-5	-4	-1	0
C	-10				

Type	Score
Match	+1
Mismatch	-1
Gap	-2

S1: ?
S2: ?

Needlemen-Wunsch example



Type	Score
Match	+1
Mismatch	-1
Gap	-2

```
S1: ?
S2: ?
```

Needlemen-Wunsch example

		T	G	A	C
	0	-2	-4	-6	-8
T	-2	+1	-1	-3	-5
C	-4	-1	0	-2	-2
A	-6	-3	-2	+1	-1
T	-8	-5	-4	-1	0
C	-10	-7	-6	-3	0

Type	Score
Match	+1
Mismatch	-1
Gap	-2

```
S1: C
S2: C
```

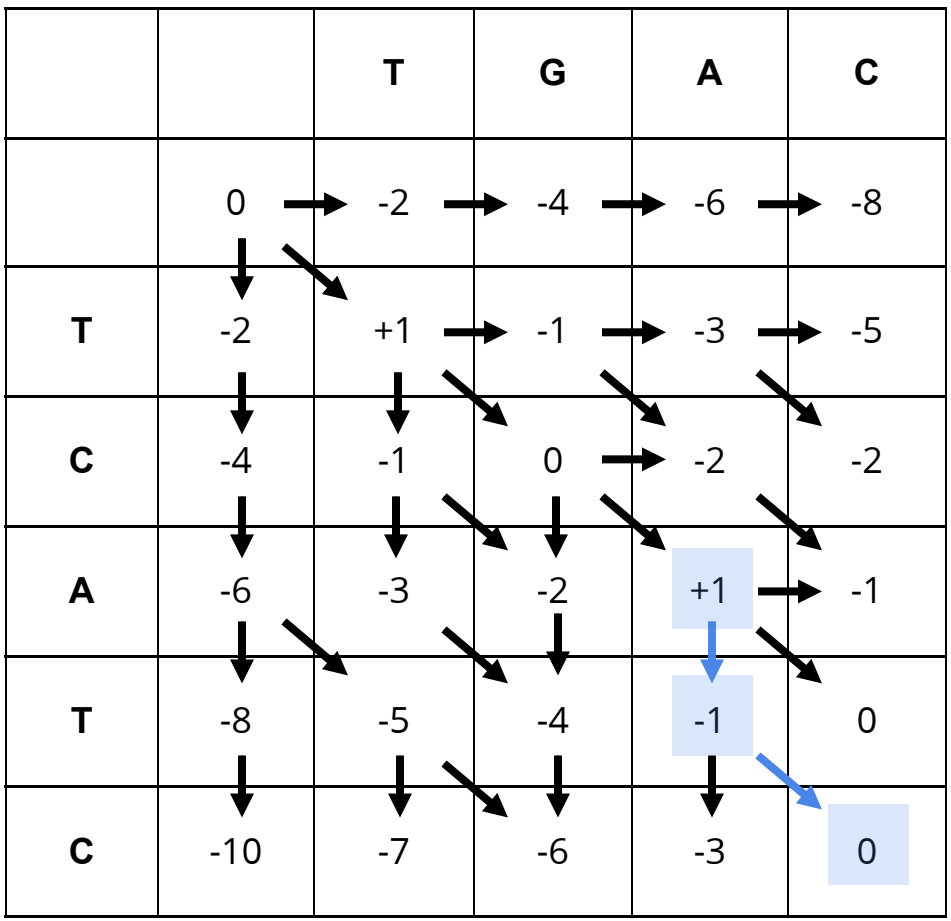
Needlemen-Wunsch example

		T	G	A	C
	0	-2	-4	-6	-8
T	-2	+1	-1	-3	-5
C	-4	-1	0	-2	-2
A	-6	-3	-2	+1	-1
T	-8	-5	-4	-1	0
C	-10	-7	-6	-3	0

Type	Score
Match	+1
Mismatch	-1
Gap	-2

```
S1: AC
S2: TC
```

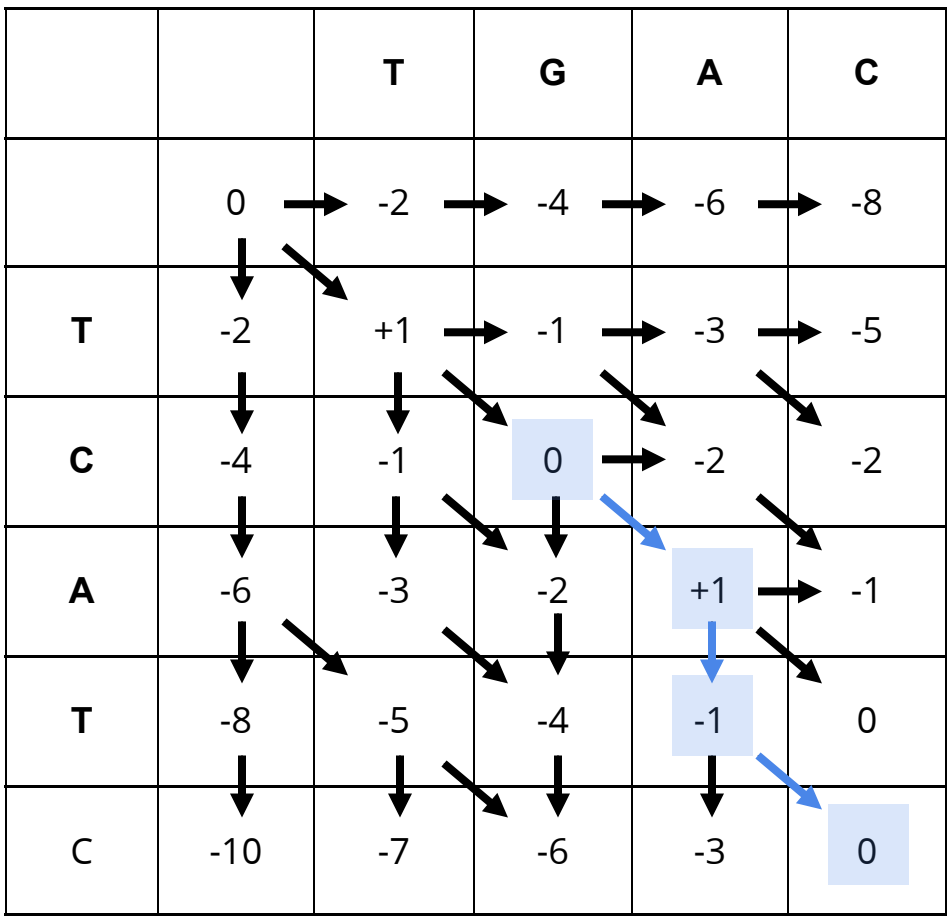
Needlemen-Wunsch example



Type	Score
Match	+1
Mismatch	-1
Gap	-2

```
S1: A-C
S2: ATC
```

Needlemen-Wunsch example



Type	Score
Match	+1
Mismatch	-1
Gap	-2

```
S1: GA-C
S2: CATC
```

Needlemen-Wunsch example

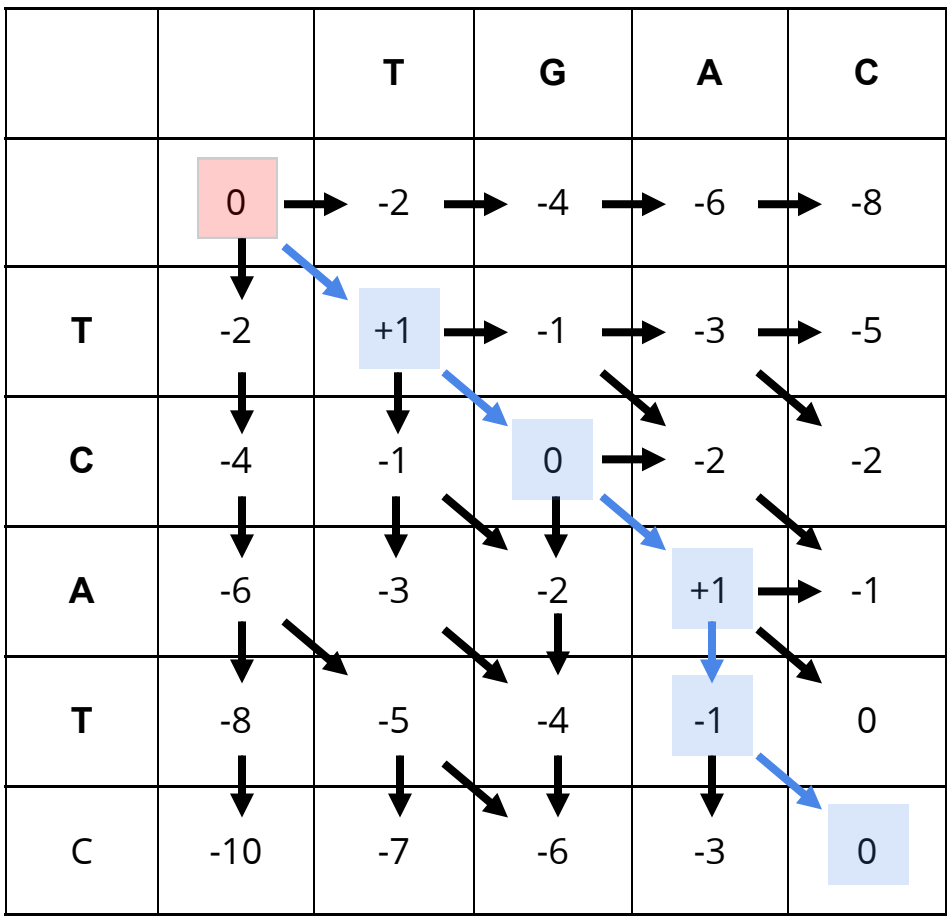
		T	G	A	C
	0	-2	-4	-6	-8
T	-2	+1	-1	-3	-5
C	-4	-1	0	-2	-2
A	-6	-3	-2	+1	-1
T	-8	-5	-4	-1	0
C	-10	-7	-6	-3	0

Type	Score
Match	+1
Mismatch	-1
Gap	-2

```
S1: TGA-C
S2: TCATC
```

Needlemen-Wunsch example

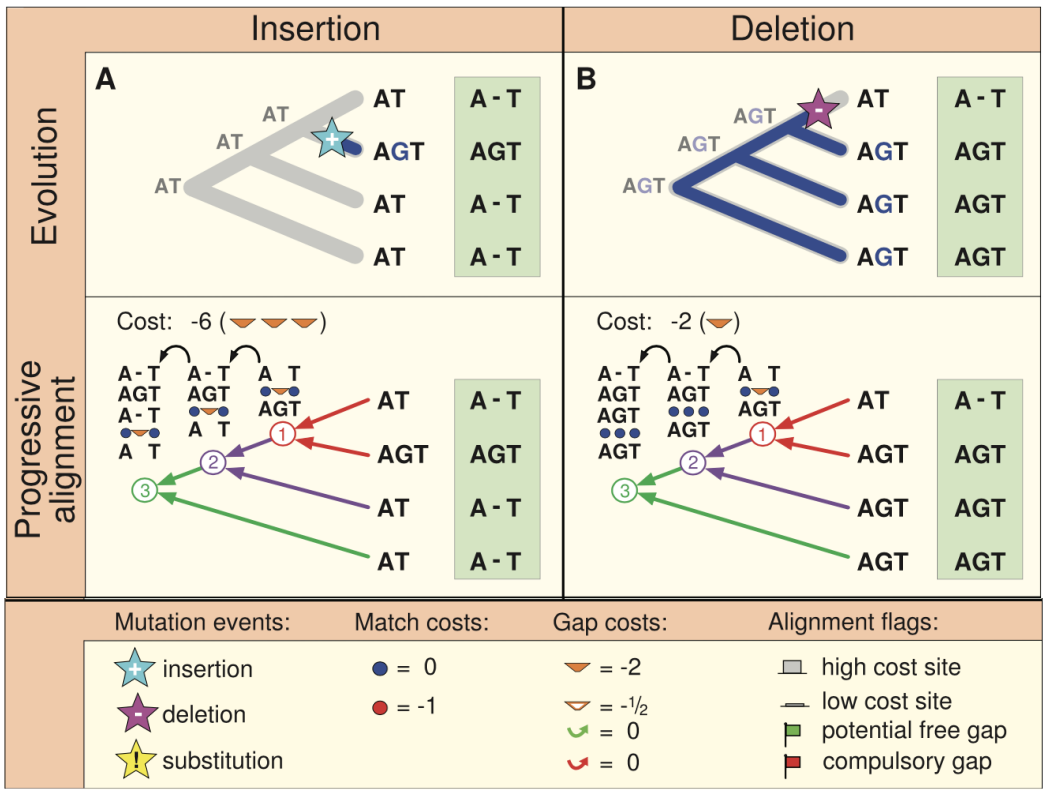
...done!



Type	Score
Match	+1
Mismatch	-1
Gap	-2

```
S1: TGA-C
S2: TCATC
```

Progressive alignment

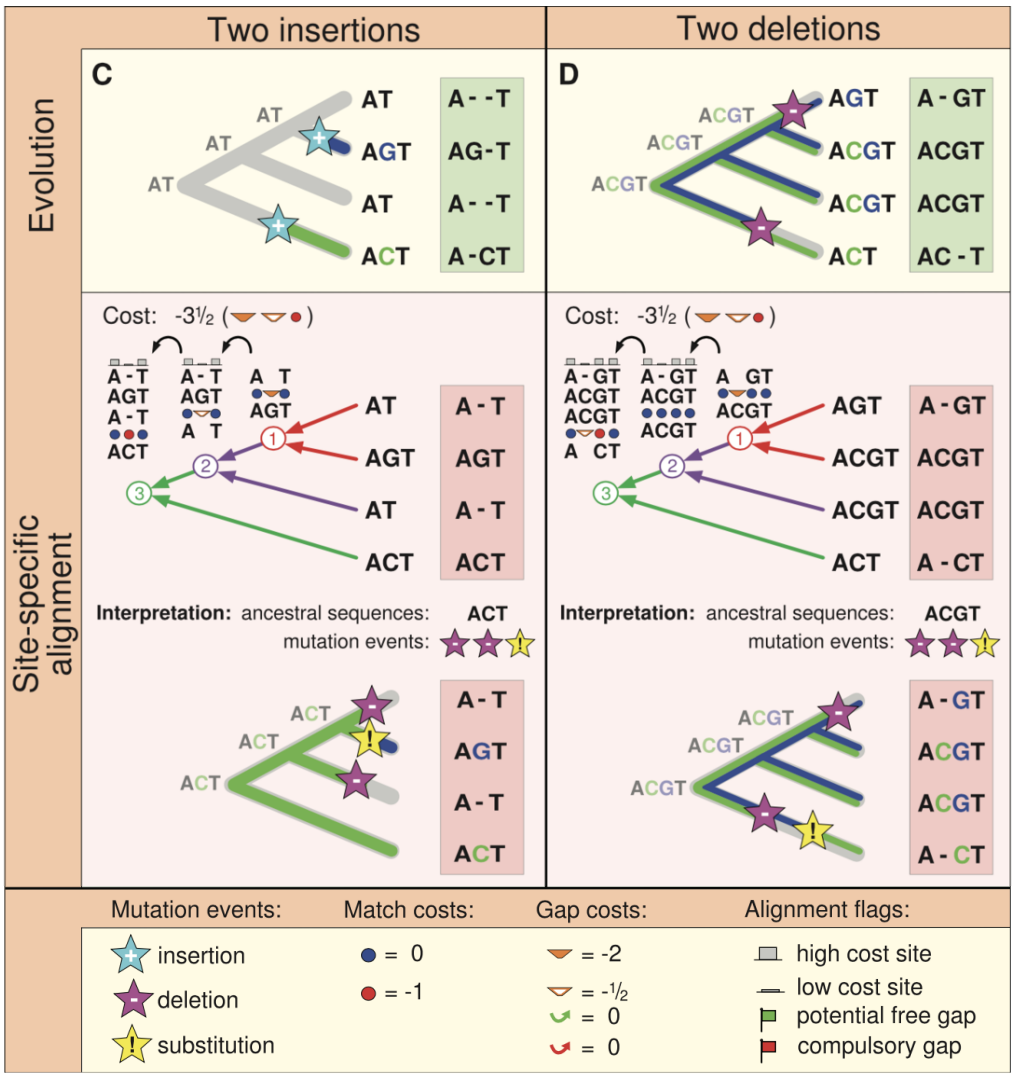


Multiple sequences aligned by adding new sequences in *progressive* manner based on ***guide tree***

Events costs may be "double-counted"

from Löytynoja and Goldman (2008 Science)

Progressive alignment



Mildly complex evolutionary scenarios may cause progressive alignments to produce inaccurate homology statements; see two-insertion scenario (left)

from Löytynoja and Goldman (2008 Science)

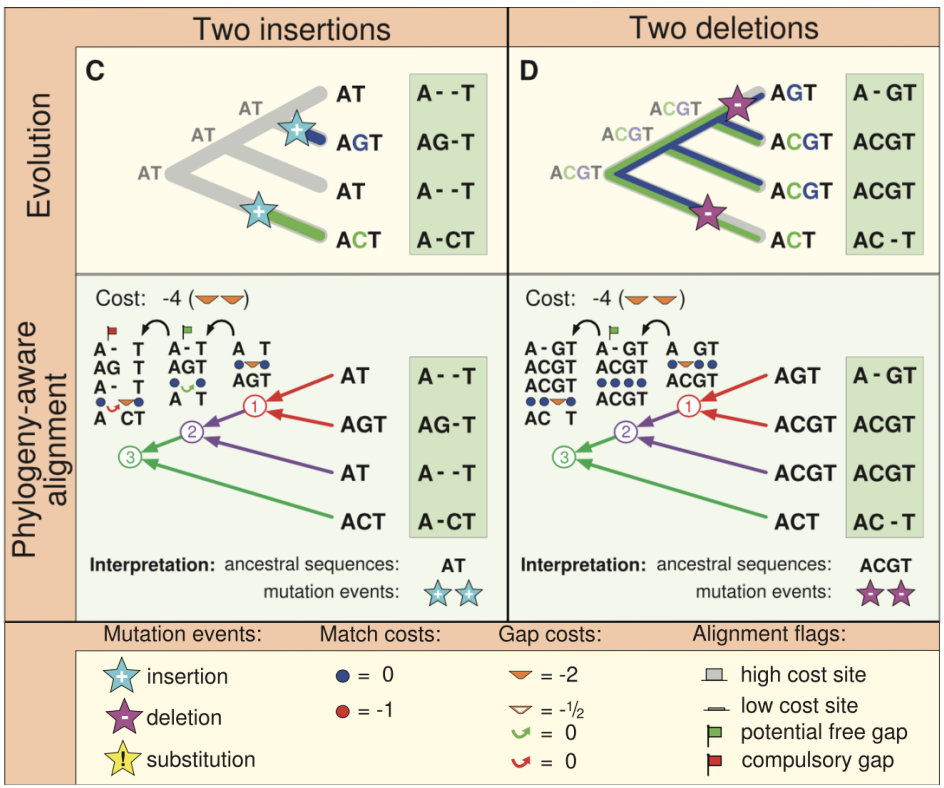
Generative alignment

Generative methods reconstruct the series of substitution, insertion, and deletion events that most likely generated a sequence alignment

Evolutionary events

- substitutions change a character state
- deletions remove a character from a sequence
- insertions add a character from a sequence

Generative alignment



Generative alignments tend to be "gappier"; more evolutionarily accurate on shorter timescales

This software (PRANK; left) improves accuracy by "flagging" evolutionary insertion events as scored, so they are not double-counted

from Löytynoja and Goldman (2008 Science)

Why alignment matters?

```
  **     * *
CCAAG-CGTTATC
-TCAGTGGT-ATC
-T-AGTGGT-ATC
CTCAGTGGA--TC
```

Why alignment matters?

```
      *
C-CAA-CGGTTATC
-TCA-GTGGT-ATC
-T-A-GTGGT-ATC
CTCA-GTGG--ATC
```

Lab 4A

github.com/WUSTL-Biol4220/home/labs/lab_04B.md