Genome 373 4/17/17

Homework recap

Extra review and practice

A bit of programming (maybe)

Reminder:
Office hours Foege S-040, Mondays 4:30-5:30

Don’t confuse an algorithm with its output

- Homework 2 question 2: Can you **use dynamic programming** to find the **optimal** alignment with an **affine gap penalty**?
  - What’s optimal?
  - What’s dynamic programming?

Don’t confuse an algorithm with its output

- Can you **use dynamic programming** to find the **optimal** alignment with an **affine gap penalty**?
  - **Optimal/best:** whatever we define it to be. How does using an affine gap change our definition of best?

ACCCTCCG
AC–CAC–G

vs

ACCCTCCG
ACC- - ACG

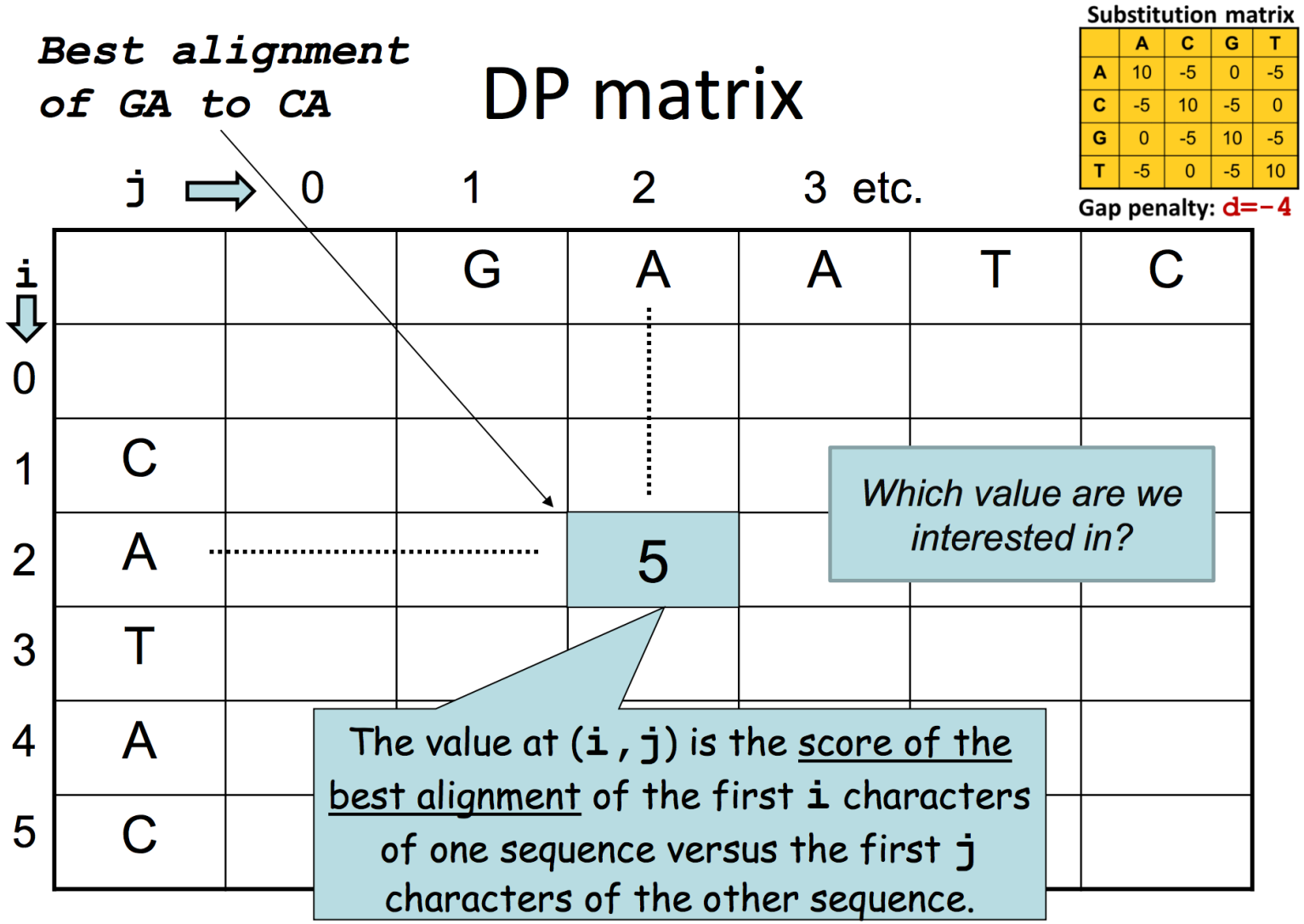
Don’t confuse an algorithm with its output

- Can you **use dynamic programming** to find the **optimal** alignment with an **affine gap penalty**?
  - **Optimal/best:** whatever we define it to be. **Affine gap penalty** defines fewer larger gaps as better than many smaller gaps.
  - What’s **dynamic programming**? (The name is not helpful)

Don’t confuse an algorithm with its output

- Can you **use dynamic programming** to find the **optimal** alignment with an **affine gap penalty**?
  - **Optimal/best:** whatever we define it to be. **Affine gap penalty** defines fewer larger gaps as better than many smaller gaps.
  - **Dynamic programming:** solve an instance of a problem by taking advantage of solutions for subparts of the problem

Dynamic programming stores **partial** best solutions to avoid spending time calculating many full solutions



Other uses of dynamic programming

- Solving Hidden Markov Models to identify genes
- Finding the *longest common subsequence* of 2 text files with *diff*
- Signal processing algorithms
- Many more!

For affine gap we need to use a more complicated version of the DP algorithm

- to do in $O(n^2)$ time, need 3 matrices instead of 1

$M(i, j)$ best score given that $x[i]$ is aligned to $y[j]$

$I_x(i, j)$ best score given that $x[i]$ is aligned to a gap

$I_y(i, j)$ best score given that $y[j]$ is aligned to a gap

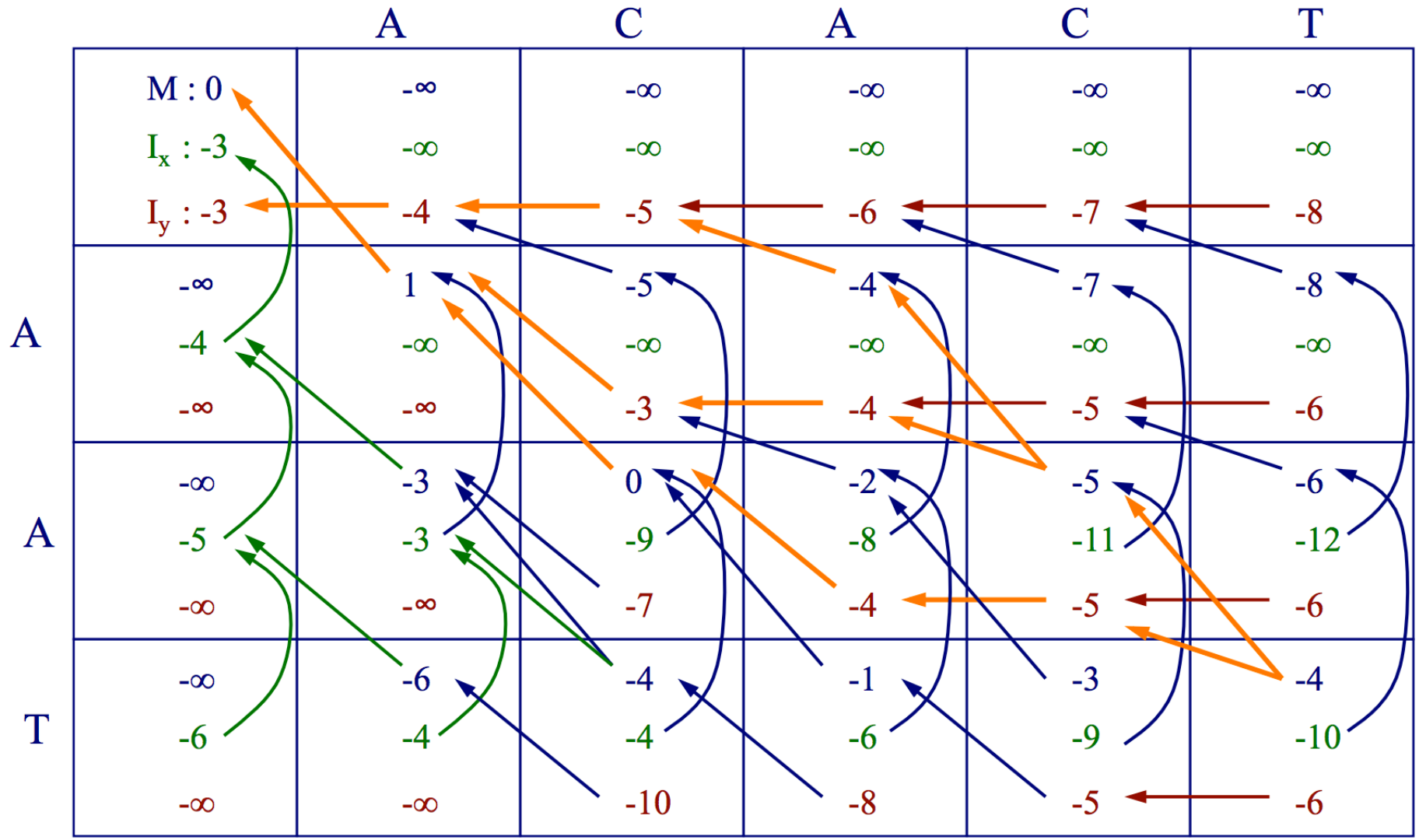
For affine gap we need to use a more complicated version of the DP algorithm

$$M(i,j) = \max \begin{cases} M(i-1,j-1) + s(x_i, y_j) & \text{match } x_i \text{ with } y_j \\ I_x(i-1,j-1) + s(x_i, y_j) & \text{insertion in } x \\ I_y(i-1,j-1) + s(x_i, y_j) & \text{insertion in } y \end{cases}$$

$$I_x(i,j) = \max \begin{cases} M(i-1,j) + h + g & \text{open gap in } x \\ I_x(i-1,j) + g & \text{extend gap in } x \end{cases}$$

$$I_y(i,j) = \max \begin{cases} M(i,j-1) + h + g & \text{open gap in } y \\ I_y(i,j-1) + g & \text{extend gap in } y \end{cases}$$

For affine gap we need to use a more complicated version of the DP algorithm



HW question 5b

- Suppose your database is random sequences with 25% of each nucleotide
  - You have a sequence that is 80% A-T
    - Probability of a match between any given base in your sequence w/ one in the database?

HW question 5b

- Suppose the database is random sequences with 25% of each nucleotide
  - You have a sequence that is 80% A-T
    - Probability of a match between any given base in your sequence w/ one in the database = still 0.25

- Suppose the database is random sequences with the same nucleotide content as your sequence: 80% A-T.
  - Now what?

HW question 6c: Probability calculations review

- What is the probability that at least one of these scores will have a low p-value by chance? = What's the probability that seq 1 or seq 2 or …seq 1000 will have a p-value that low?

  - *Any/one or more = OR*
  - *All = AND*

HW question 6c: Probability calculations review

- What is the probability that at least one of these scores will have a low p-value by chance? = What's the probability that seq 1 or seq 2 or …seq 1000 will have a p-value that low?

  - *Any/one or more = OR:* Use 1 – p(All of them will NOT)

  = 1- (0.99995)^1000 = 0.0488

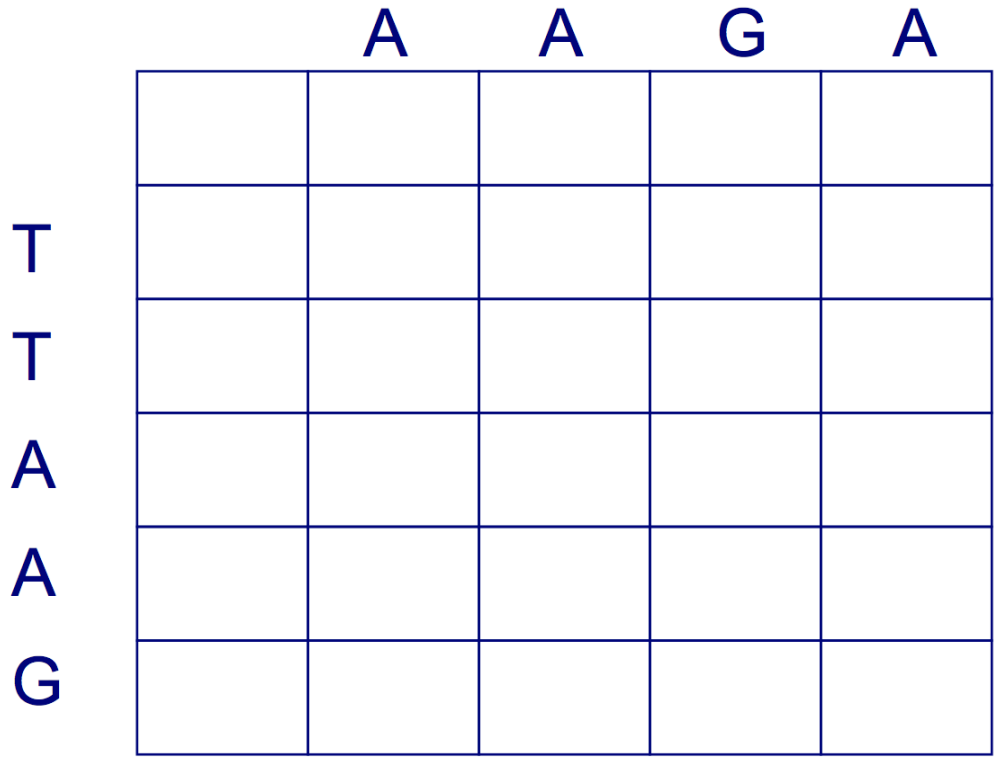
  - *All = AND*

Local alignment tracebacks: Start from the highest score!

- Don’t include anything from either end that worsens the alignment

Substitution matrix

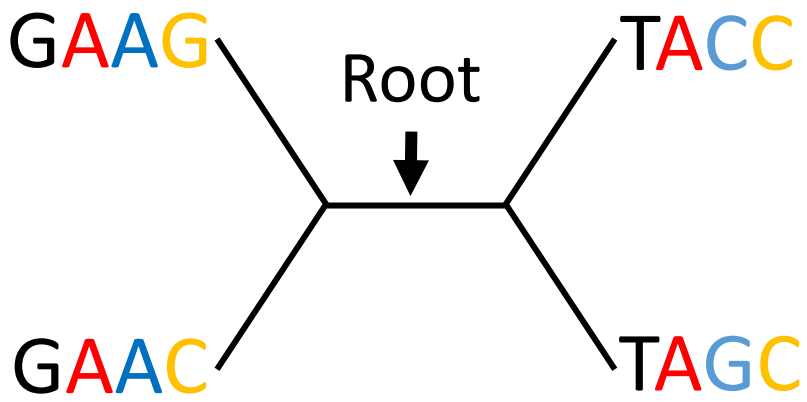
	A	C	G	T
A	10	-5	0	-5
C	-5	10	-5	0
G	0	-5	10	-5
T	-5	0	-5	10



Note: Problem-solving intuition

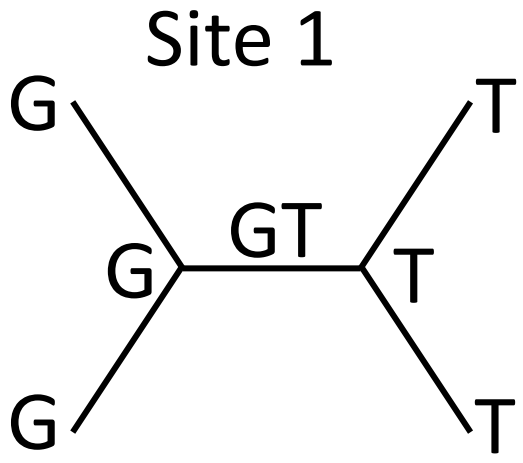
- Look at your final answer and ask yourself if it makes sense
- E.g. All cells with positive numbers should have arrows pointing in
- Does my local alignment have gaps on the end?
- Does my global alignment include all bases from the original sequences?

Fitch algorithm, bottom up phase: what is the parsimony score?



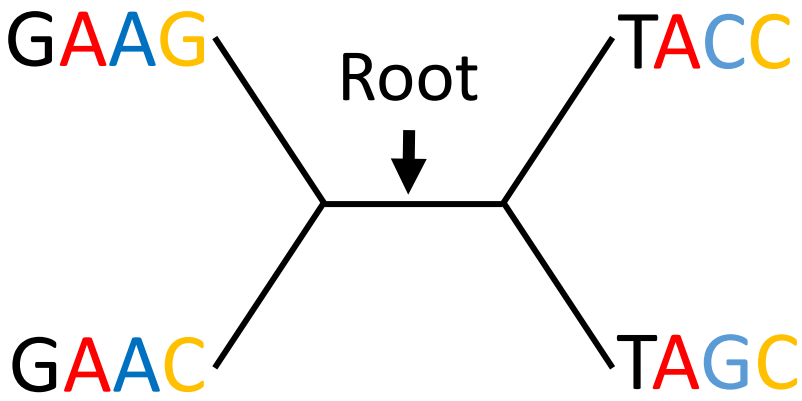
1. Initialization: $R_i = \{s_i\}$
2. Traverse the tree from leaves to root ("post-order")
3. Determine R_i of internal node i with children j, k:

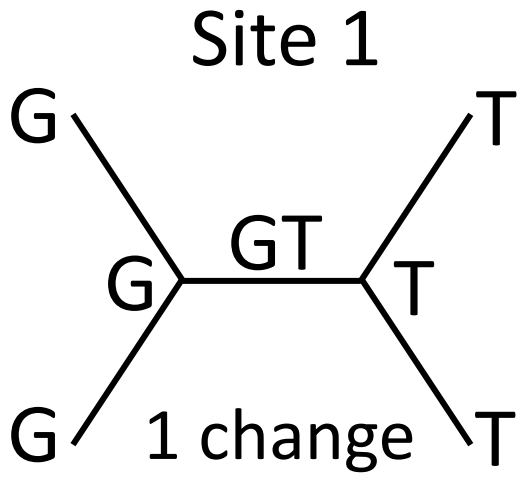
$$R_i = \begin{cases} if \;\; R_j \cap R_k \neq \phi \rightarrow R_j \cap R_k \\ otherwise \rightarrow R_j \cup R_k \end{cases}$$



1 change
parsimony score of 1

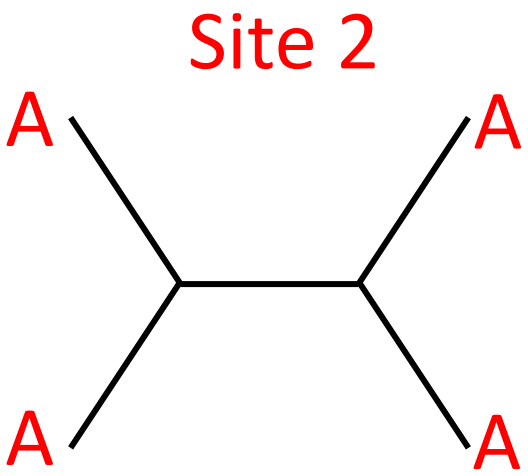
What is a possible pair of sequences consistent with the score?

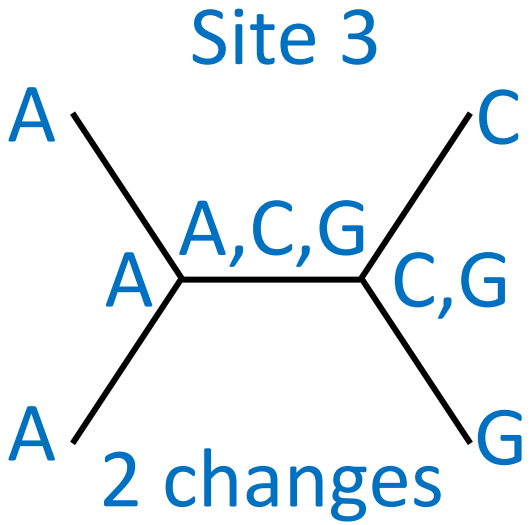


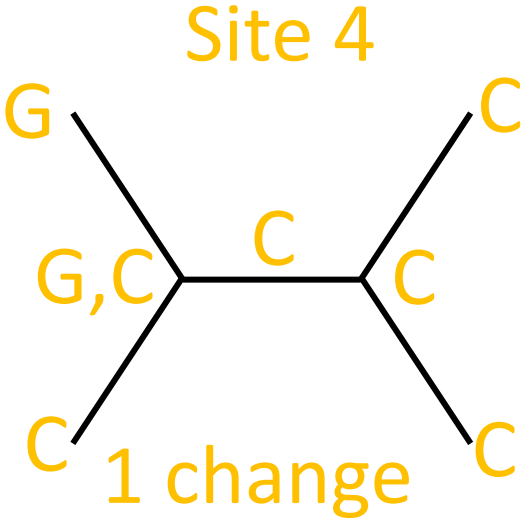


Total:

1 + 2 + 1 = 4





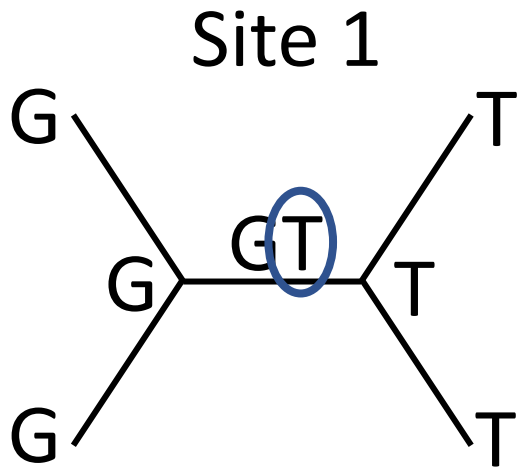


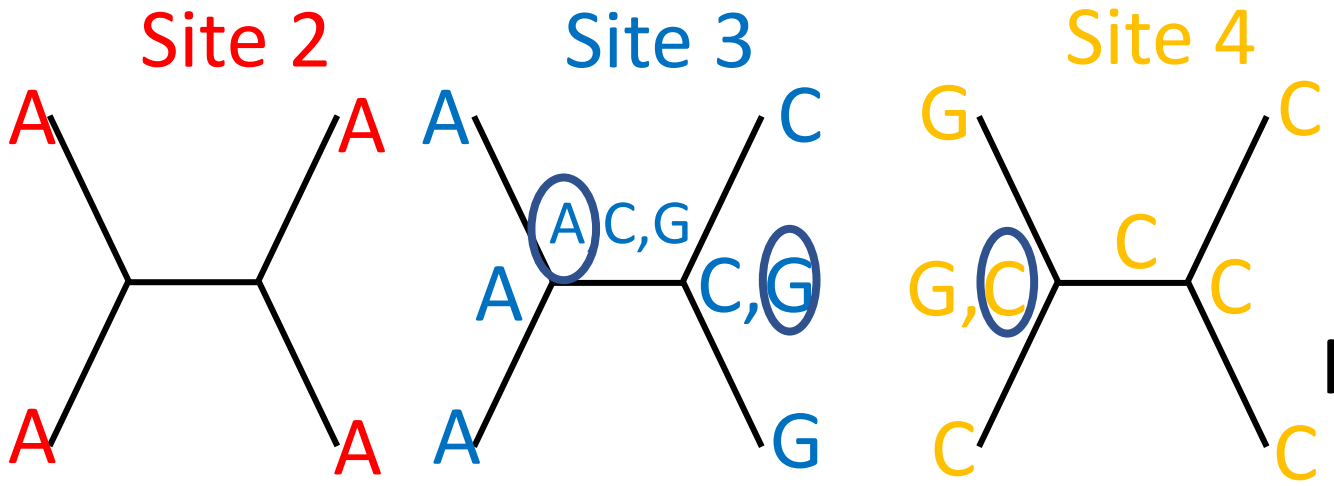
2. Fitch's algorithm: Top-down phase

(Pick a state for each internal node)

1. Pick arbitrary state in R_{root} to be the state of the root ,s_{root}
2. Traverse the tree from root to leaves ("pre-order")
3. Determine s_i of internal node i with parent j:

$$s_i = \begin{cases} if \quad s_j \in R_i \rightarrow s_j \\ otherwise \rightarrow arbitrary \quad state \in R_i \end{cases}$$





Root	TAAC
Left internal node	GAAC
Right internal node	TAGC

Fitch algorithm practice