

# Approximate matching

Read

CTCAAACCTCTGACCTTTGGTGATCCACCCGCCTAGGCCTTC

Reference

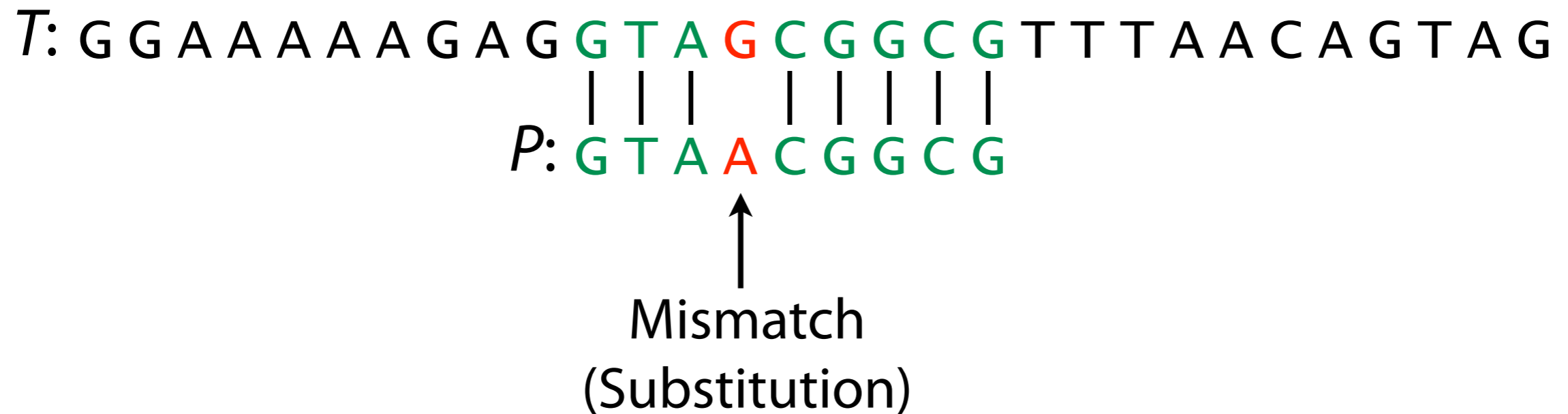
GATCACAGGTCTATCACCTATTAACCACTCACGGGAGCTCTCCATGCATTTGGTATTTT  
CGTCTGGGGGGTATGCACGCGATAGCATTGCGAGACGCTGGAGCCGGAGCACCTATGTC  
GCAGTATCTGTCTTTGATTCTGCCTCATCCTATTATTTATCGCACCTACGTTCAATATT  
ACAGGCGAACATACTTACTAAAGTGTGTTAATTAATTAATGCTTGTAGGACATAATAATA  
ACAATTGAATGTCTGCACAGCCACTTTCCACACAGACATCATAACAAAAAATTTCCACCA  
AACCCCCCTCCCCGCTTCTGGCCACAGCACTCTGCCAAACCCCAAAA  
ACAAAGAACCCTAACACCAGCCTAACCAATTTCAAATTTTATCTTTGGCGGTATGCAC  
TTTTAACAGTCACCCCCCACTAACCAATTTTCCCCTCCCCTCCATACTACTAAT  
CTCATCAATACAACCCCCGCCATCTACCCAGCACACACACACCCCTCTAACCCCAT  
CCCCGAACCAACCAACCCCAAAAGCACCCCCACAGTTTATGTAGCTTCCCTCTCAA  
GCAATACACTGACCCGCTCAAACCTCTGGATTTTGGATCCACCCAGCGCTTGGCCTAAA  
CTAGCCTTTCTATTAGCTCTTAGAAGATTACACATGCAAGCATCCCCCTCCAGTGAGT  
TCACCCTCTAAATCACCACGATCAAGGAACAAGCATCAAGCACGCAGCAATGCAGCTC  
AAAACGCTTAGCCTAGCCACACCCTCACGGGAAACAGCAGTGATTAACCTTAGCAATAA  
ACGAAAGTTTAACTAAGCTATACTAACCCAGGGTTGGTCAATTTCTGTCACAGCCACCGC  
GGTCACACGATTAACCCAAGTCAATGAAGCCGGCGTAAAGAGTGTCTAGATCACCCCC  
TCCCCAATAAAGCTAAAACCTCACCTGATTTGTAAAAAACTCCAGTTACAAAAATAGAC  
TACGAAAGTGGCTTTAACATATCTGAACAACAATAGCTAAGACCTGGGATTAGA  
TACCCCACTATGCTTAGCCCTAAACCTCAACAGCAACAACCTGGCCAGAA  
CACTACGAGCCACAGCTTAAAACCTCAAAGGACCTGGCGGTGCTTCATCTAGAGG  
AGCCTGTTCTGTAATCGATAAACCCCGATCAACCTCACCACTCTTGCTCTATA  
CCGCCATCTTCAGCAAACCCTGATGAAGGCTACAAAGTAAGCGCAAGTACCTAG  
ACGTTAGGTCAAGGTGTAGCCCATGAGGTGGCAAGAAATGGGCTACATTTTC  
AAAACCTACGATAGCCCTTATGAACTTAAGGGTCAAGGTGGATTTAGCAGTAA  
AGTAGAGTGCTTAGTTGAACAGGGCCCTGAAGCGGTACACACCCGCCCGTCACCCT  
AAGTATACTTCAAAGGACATTTAACTAAAACCCCTACGCATTTATATAGAGGAGACA  
CGTAACCTCAAACCTCTGCCTTTGGTGATCCACCCGCCTTGGCCTACCTGCATAATGAAG  
AAGCACCCAACCTTACTTAGGAGATTTCAACTTAACTTGACCGCTCTGAGCTAAACCTA  
GCCCAAACCCACTCCACCTTACTACCAGACAACCTTAGCCAAACCATTTACCCAAATAA  
AGTATAGGCGATAGAAATTGAAACCTGGCGCAATAGATATAGTACCGCAAGGGAAAGATG  
AAAAATTATAACCAAGCATAATATAGCAAGGACTAACCCCTATACCTTCTGCATAATGAA  
TTAACTAGAAATAACTTTGCAAGGAGAGCCAAAGCTAAGACCCCCGAAACCAGACGAGCT  
ACCTAAGAACAGCTAAAAGAGCACACCCGTCTATGTAGCAAAATAGTGGGAAGATTTATA  
GGTAGAGGCGACAAACCTACCGAGCCTGGTGATAGCTGGTTGTCCAAGATAGAATCTTAG  
TTCAACTTTAAATTTGCCACAGAACCCTCTAAATCCCCTTGTAATTTAACTGTTAGTC  
CAAAGAGGAACAGCTCTTTGGACACTAGGAAAAAAACCTTGTAGAGAGAGTAAAAAATTTA

Differences between read and reference occur because of...

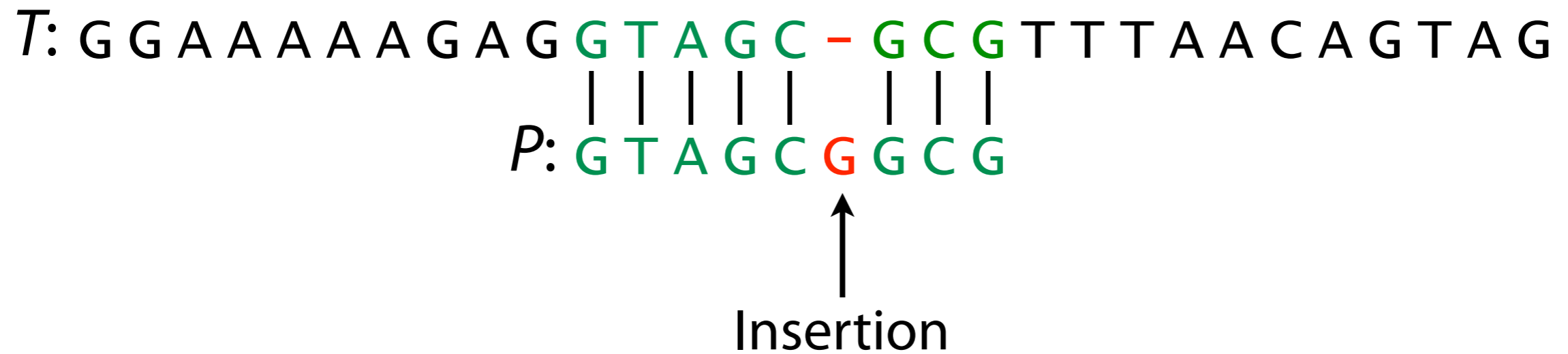
1. Sequencing error
2. Natural variation



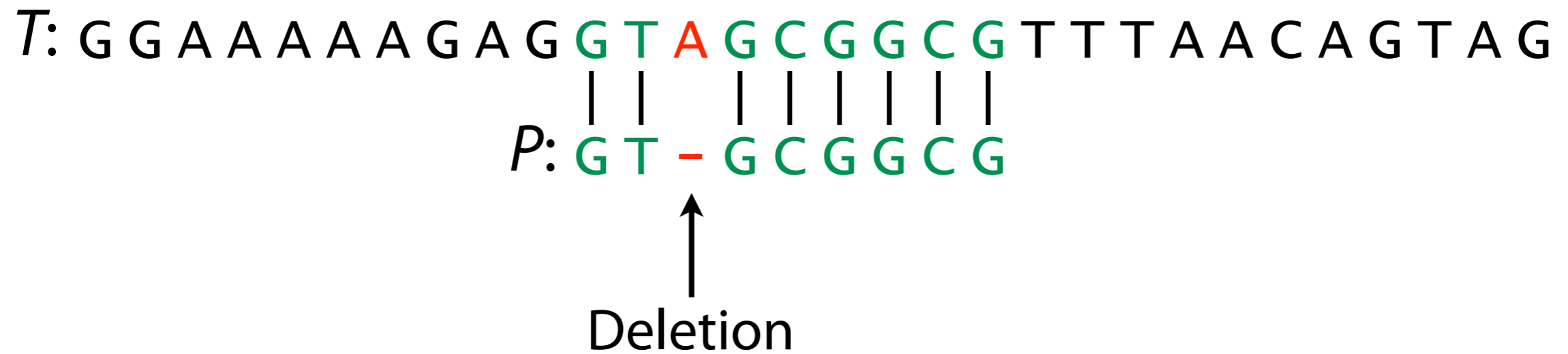
# Approximate matching



# Approximate matching



# Approximate matching



# Hamming distance

For  $X$  &  $Y$  where  $|X| = |Y|$ , *hamming distance* =  
minimum # substitutions needed to turn one into the other

$X:$	G	A	G	G	T	A	G	C	G	G	C	G	T	T	T	A	A	C
$Y:$	G	T	G	G	T	A	A	C	G	G	G	G	T	T	T	A	A	C

*Hamming distance = 3*

# Edit distance

(AKA Levenshtein distance)

For  $X$  &  $Y$ , *edit distance* = minimum # edits (substitutions, insertions, deletions) needed to turn one into the other

X: T G G C C G C G C A A A A A C A G C  
| | | | | | | | | | | | | | | | |  
Y: T G A C C G C G C A A A A - C A G C

*Edit distance = 2*

X: G C G T A T G C G G C T A - A C G C  
| | | | | | | | | | | | | | | | |  
Y: G C - T A T G C G G C T A T A C G C

*Edit distance = 2*

# Edit distance

For  $X, Y$  where  $|X| = |Y|$ , *hamming distance* = minimum # substitutions needed to turn one into the other

For  $X, Y$ , *edit distance* = minimum # edits (substitutions, insertions, deletions) needed to turn one into the other

# Edit distance

If  $|X| = |Y|$  what can we say about the relationship between **editDistance**( $X, Y$ ) and **hammingDistance**( $X, Y$ )?

$$\text{editDistance}(X, Y) \leq \text{hammingDistance}(X, Y)$$

X:	G	C	G	T	A	T	G	C	G	G	C	T	A	-	A	C	G	C
Y:	G	C	-	T	A	T	G	C	G	G	C	T	A	T	A	C	G	C



# Edit distance

If  $x$  and  $y$  are different lengths, what can we say about **editDistance**( $X, Y$ )?

$$\mathbf{editDistance}(X, Y) \geq ||X| - |Y||$$

$X: ? ?$

$Y: ? ? ? ?$

*X*

G G C C G C G C A A A A A C A G C

*Y*

A T G C C G C G A A A A A C A T A

**editDistance( X[:-1], Y[:-1] ) = 147**

G G C C G C G C A A A A A C A G C

$\alpha$

A T G C C G C G A A A A A C A T A

$\beta$

$\alpha$  C

$\beta$  A

$$\text{edist}(\alpha C, \beta A) = \min \begin{cases} \text{edist}(\alpha, \beta) + 1 \\ \text{edist}(\alpha C, \beta) + 1 \\ \text{edist}(\alpha, \beta A) + 1 \end{cases}$$

$\alpha$  C

$\beta$  A

$$\text{edist}(\alpha C, \beta A) = \min \begin{cases} \text{edist}(\alpha, \beta) + 1 \\ \text{edist}(\alpha C, \beta) + 1 \\ \text{edist}(\alpha, \beta A) + 1 \end{cases}$$

$\alpha C$

$\beta A$

$$\text{edist}(\alpha C, \beta A) = \min \begin{cases} \text{edist}(\alpha, \beta) + 1 \\ \text{edist}(\alpha C, \beta) + 1 \\ \text{edist}(\alpha, \beta A) + 1 \end{cases}$$

$\alpha C$

$\beta A$

$$\text{edist}(\alpha C, \beta A) = \min \begin{cases} \text{edist}(\alpha, \beta) + 1 \\ \text{edist}(\alpha C, \beta) + 1 \\ \text{edist}(\alpha, \beta A) + 1 \end{cases}$$

$\alpha$   $x$

$\beta$   $y$

$$\text{edist}(\alpha x, \beta y) = \min \begin{cases} \text{edist}(\alpha, \beta) + \delta(x, y) \\ \text{edist}(\alpha x, \beta) + 1 \\ \text{edist}(\alpha, \beta y) + 1 \end{cases}$$

$\delta(x, y) = 0$  if  $x = y$ , or 1 otherwise



```

delt = 1 if a[-1] != b[-1] else 0
return min(edDistRecursive(a[:-1], b[:-1]) + delt,
           edDistRecursive(a, b[:-1]) + 1,
           edDistRecursive(a[:-1], b) + 1)

```

$$\text{edist}(\alpha x, \beta y) = \min \begin{cases} \text{edist}(\alpha, \beta) + \delta(x, y) \\ \text{edist}(\alpha x, \beta) + 1 \\ \text{edist}(\alpha, \beta y) + 1 \end{cases}$$

$\delta(x, y) = 0$  if  $x = y$ , or 1 otherwise

**edDistRecursive**("ABC", "BBC")

("ABC", "BB") ("AB", "BB") ("AB", "BBC")

("ABC", "B") ("AB", "B") ("AB", "BB")

**edDistRecursive**("ABC", "BBC")

("ABC", "BB")

("AB", "BB")

("AB", "BBC")

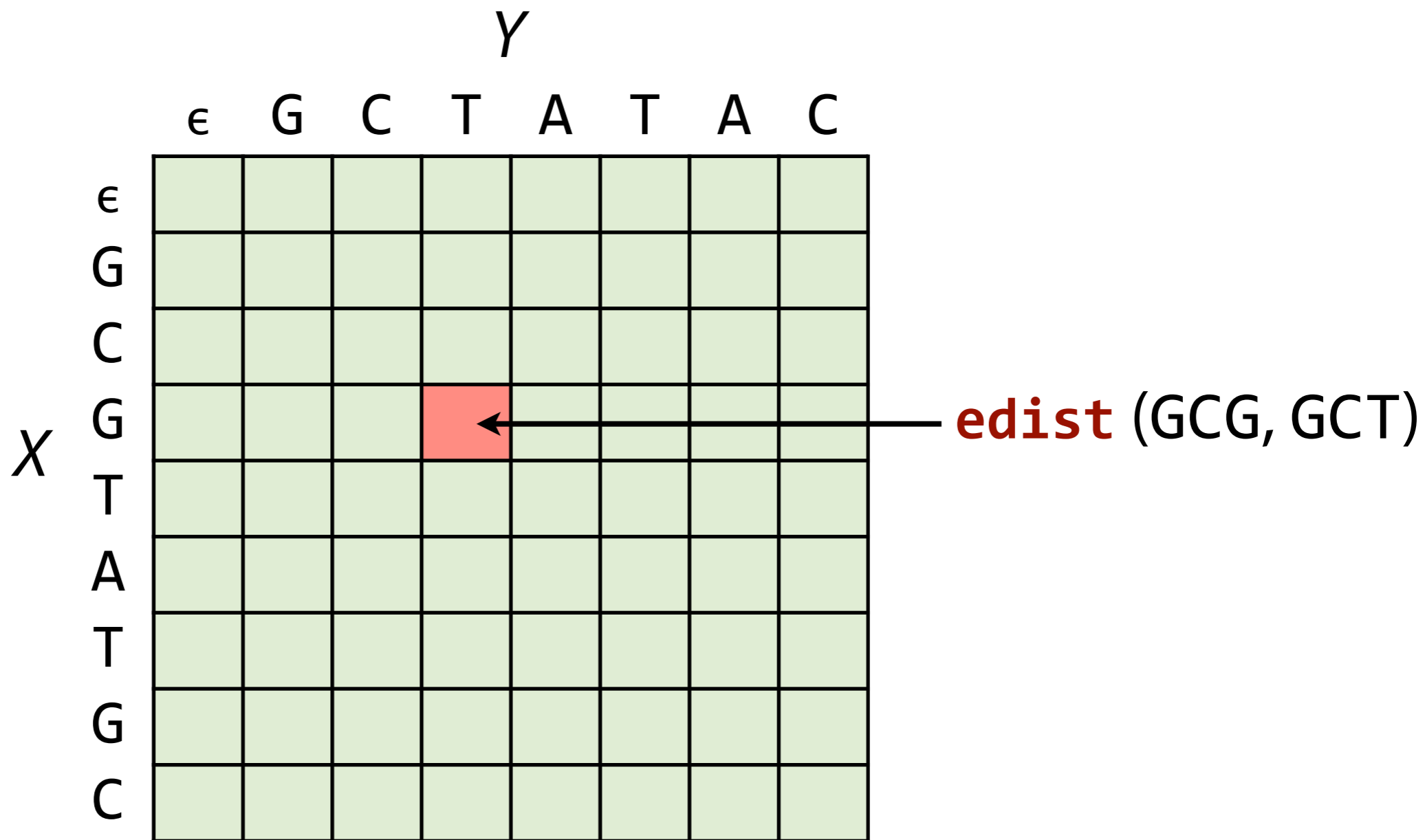
("ABC", "B")

("AB", "B")

("AB", "BB")

*Y*

	$\epsilon$	G	C	T	A	T	A	C
<i>X</i> $\epsilon$								
G								
C								
G								
T								
A								
T								
G								
C								



*Y*

	$\epsilon$	G	C	T	A	T	A	C
<i>X</i>	$\epsilon$							
G								
C								
G								
T								
A								
T								
G								
C								

**edist** (GCGTATGC, GCTATAC)

		Y							
		ε	G	C	T	A	T	A	C
X	ε								
	G								
	C								
	G								
	T								
	A								
	T								
	G								
	C								

$$\boxed{\text{edist}(\alpha x, \beta y)} = \min \begin{cases}
 \boxed{\text{edist}(\alpha, \beta) + \delta(x, y)} \\
 \boxed{\text{edist}(\alpha x, \beta) + 1} \\
 \boxed{\text{edist}(\alpha, \beta y) + 1}
 \end{cases}$$

		Y							
		ε	G	C	T	A	T	A	C
X	ε	0	1	2	3				
	G	1	0	1	2				
	C	2	1	0	1				
	G	3	2	1					
	T								
	A								
	T								
	G								
	C								

$$\text{edist}(\alpha x, \beta y) = \min \begin{cases}
 \text{edist}(\alpha, \beta) + \delta(x, y) & = 0 + 1 = 1 \\
 \text{edist}(\alpha x, \beta) + 1 & = 1 + 1 = 2 \\
 \text{edist}(\alpha, \beta y) + 1 & = 1 + 1 = 2
 \end{cases}$$



		Y							
		ε	G	C	T	A	T	A	C
X	ε	0	1	2	3				
	G	1	0	1	2				
	C	2	1	0	1				
	G	3	2	1	1				
	T								
	A								
	T								
	G								
	C								

$$\text{edist}(\alpha x, \beta y) = \min \begin{cases}
 \text{edist}(\alpha, \beta) + \delta(x, y) & = 0 + 1 = 1 \\
 \text{edist}(\alpha x, \beta) + 1 & = 1 + 1 = 2 \\
 \text{edist}(\alpha, \beta y) + 1 & = 1 + 1 = 2
 \end{cases}$$

		Y							
		ε	G	C	T	A	T	A	C
X	ε	0	1	2	3	4	5	6	7
	G	1	0	1	2	3	4	5	6
	C	2	1	0	1	2	3	4	5
	G	3	2	1	1	2	3	4	5
	T	4	3	2	1	2	2	3	4
	A	5	4	3	2	1	2	2	3
	T	6	5	4	3	2	1	2	3
	G	7	6	5	4	3	2	2	3
	C	8	7	6	5	4	3	3	2

← Final result

$$\text{edist}(\alpha x, \beta y) = \min \begin{cases} \text{edist}(\alpha, \beta) + \delta(x, y) \\ \text{edist}(\alpha x, \beta) + 1 \\ \text{edist}(\alpha, \beta y) + 1 \end{cases}$$

		Y							
		ε	G	C	T	A	T	A	C
X	ε	0	1	2	3	4	5	6	7
	G	1	0	1	2	3	4	5	6
	C	2	1	0	1	2	3	4	5
	G	3	2	1	1	2	3	4	5
	T	4	3	2	1	2	2	3	4
	A	5	4	3	2	1	2	2	3
	T	6	5	4	3	2	1	2	3
	G	7	6	5	4	3	2	2	3
	C	8	7	6	5	4	3	3	2

For any pair of prefixes from  $X$  &  $Y$ , edit distance is calculated *once*

*Dynamic programming*

# Edit distance

$Y$

	$\epsilon$	G	C	T	A	T	A	C
$X$ $\epsilon$								
G								
C								
G								
T								
A								
T								
G								
C								



		Y							
		ε	G	C	T	A	T	A	C
X	ε	0	1	2	3	4	5	6	7
	G	1							
	C	2							
	G	3							
	T	4							
	A	5							
	T	6							
	G	7							
	C	8							

*T*

ε T A T T G G C T A T A C G G T T

*offset* →

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

*P*

*T*

ε T A T T G G C T A T A C G G T T

*P*

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



		<i>T</i>																
		$\epsilon$	T	A	T	T	G	G	C	T	A	T	A	C	G	G	T	T
<i>P</i>	G	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	C	1	1	1	1	1	0	0	1	1	1	1	1	1	0	0	1	1
	G	2	2	2	2	2	1	1	0	1	2	2	2	1	1	1	1	2
	T	3	3	3	3	3	2	1	1	1	2	3	3	2	1	1	2	2
	A	4	3	4	3	3	3	2	2	1	2	2	3	3	2	2	1	2
	T	5	4	3	4	4	4	3	3	2	1	2	2	3	3	3	2	2
	G	6	5	4	3	4	5	4	4	3	2	1	2	3	4	4	3	2
	C	7	6	5	4	4	4	5	5	4	3	2	2	3	3	4	4	3
	8	7	6	5	5	5	5	5	5	4	3	3	2	3	4	5	4	

*P* occurs in *T* with 2 edits

		<i>T</i>																
		$\epsilon$	T	A	T	T	G	G	C	T	A	T	A	C	G	G	T	T
<i>P</i>	G	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	C	1	1	1	1	1	0	0	1	1	1	1	1	1	0	0	1	1
	G	2	2	2	2	2	1	1	0	1	2	2	2	1	1	1	1	2
	T	3	3	3	3	3	2	1	1	1	2	3	3	2	1	1	2	2
	A	4	3	4	3	3	3	2	2	1	2	2	3	3	2	2	1	2
	T	5	4	3	4	4	4	3	3	2	1	2	2	3	3	3	2	2
	G	6	5	4	3	4	5	4	4	3	2	1	2	3	4	4	3	2
	G	7	6	5	4	4	4	5	5	4	3	2	2	3	3	4	4	3
	C	8	7	6	5	5	5	5	5	5	4	3	3	2	3	4	5	4



How did I get here?

		<i>T</i>																
		$\epsilon$	T	A	T	T	G	G	C	T	A	T	A	C	G	G	T	T
<i>P</i>	G	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	C	1	1	1	1	1	0	0	1	1	1	1	1	1	0	0	1	1
	G	2	2	2	2	2	1	1	0	1	2	2	2	1	1	1	1	2
	T	3	3	3	3	3	2	1	1	1	2	3	3	2	1	1	2	2
	A	4	3	4	3	3	3	2	2	1	2	2	3	3	2	2	1	2
	T	5	4	3	4	4	4	3	3	2	1	2	2	3	3	3	2	2
	G	6	5	4	3	4	5	4	4	3	2	1	2	3	4	4	3	2
	T	7	6	5	4	4	4	5	5	4	3	2	2	3	3	4	4	3
	C	8	7	6	5	5	5	5	5	5	4	3	3	2	3	4	5	4

*T*

		$\epsilon$	T	A	T	T	G	G	C	T	A	T	A	C	G	G	T	T	
<i>P</i>	G	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	C	1	1	1	1	1	0	0	1	1	1	1	1	1	1	0	0	1	1
	G	2	2	2	2	2	1	1	0	1	2	2	2	2	1	1	1	1	2
	T	3	3	3	3	3	2	1	1	1	2	3	3	2	1	1	1	2	2
	A	4	3	4	3	3	3	2	2	1	2	2	3	3	2	2	2	1	2
	T	5	4	3	4	4	4	3	3	2	1	2	2	3	3	3	3	2	2
	G	6	5	4	3	4	5	4	4	3	2	1	2	3	4	4	4	3	2
	T	7	6	5	4	4	4	5	5	4	3	2	2	3	3	4	4	4	3
	C	8	7	6	5	5	5	5	5	5	4	3	2	2	3	3	4	5	4

How did I get here?

		<i>T</i>																
		$\epsilon$	T	A	T	T	G	G	C	T	A	T	A	C	G	G	T	T
<i>P</i>	G	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	C	1	1	1	1	1	0	0	1	1	1	1	1	1	0	0	1	1
	G	2	2	2	2	2	1	1	0	1	2	2	2	1	1	1	1	2
	T	3	3	3	3	3	2	1	1	1	2	3	3	2	1	1	2	2
	A	4	3	4	3	3	3	2	2	1	2	2	3	3	2	2	1	2
	T	5	4	3	4	4	4	3	3	2	1	2	2	3	3	3	2	2
	G	6	5	4	3	4	5	4	4	3	2	1	2	3	4	4	3	2
	T	7	6	5	4	4	4	5	5	4	3	2	2	3	3	4	4	3
	C	8	7	6	5	5	5	5	5	5	4	3	3	2	3	4	5	4

		<i>T</i>																
		$\epsilon$	T	A	T	T	G	G	C	T	A	T	A	C	G	G	T	T
<i>P</i>	G	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	C	1	1	1	1	1	0	0	1	1	1	1	1	1	0	0	1	1
	G	2	2	2	2	2	1	1	0	1	2	2	2	1	1	1	1	2
	T	3	3	3	3	3	2	1	1	1	2	3	3	2	1	1	2	2
	A	4	3	4	3	3	3	2	2	1	2	2	3	3	2	2	1	2
	T	5	4	3	4	4	4	3	3	2	1	2	2	3	3	3	2	2
	G	6	5	4	3	4	5	4	4	3	2	1	2	3	4	4	3	2
	C	7	6	5	4	4	4	5	5	4	3	2	2	3	3	4	4	3
	8	7	6	5	5	5	5	5	5	4	3	3	2	3	4	5	4	

		<i>T</i>																
		$\epsilon$	T	A	T	T	G	G	C	T	A	T	A	C	G	G	T	T
<i>P</i>	G	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	C	1	1	1	1	1	0	0	1	1	1	1	1	1	0	0	1	1
	G	2	2	2	2	2	1	1	0	1	2	2	2	1	1	1	1	2
	T	3	3	3	3	3	2	1	1	1	2	3	3	2	1	1	2	2
	A	4	3	4	3	3	3	2	2	1	2	2	3	3	2	2	1	2
	T	5	4	3	4	4	4	3	3	2	2	2	3	3	3	3	2	2
	G	6	5	4	3	4	5	4	4	3	2	1	2	3	4	4	3	2
	C	7	6	5	4	4	4	5	5	4	3	2	2	3	3	4	4	3
	8	7	6	5	5	5	5	5	5	4	3	3	2	3	4	5	4	

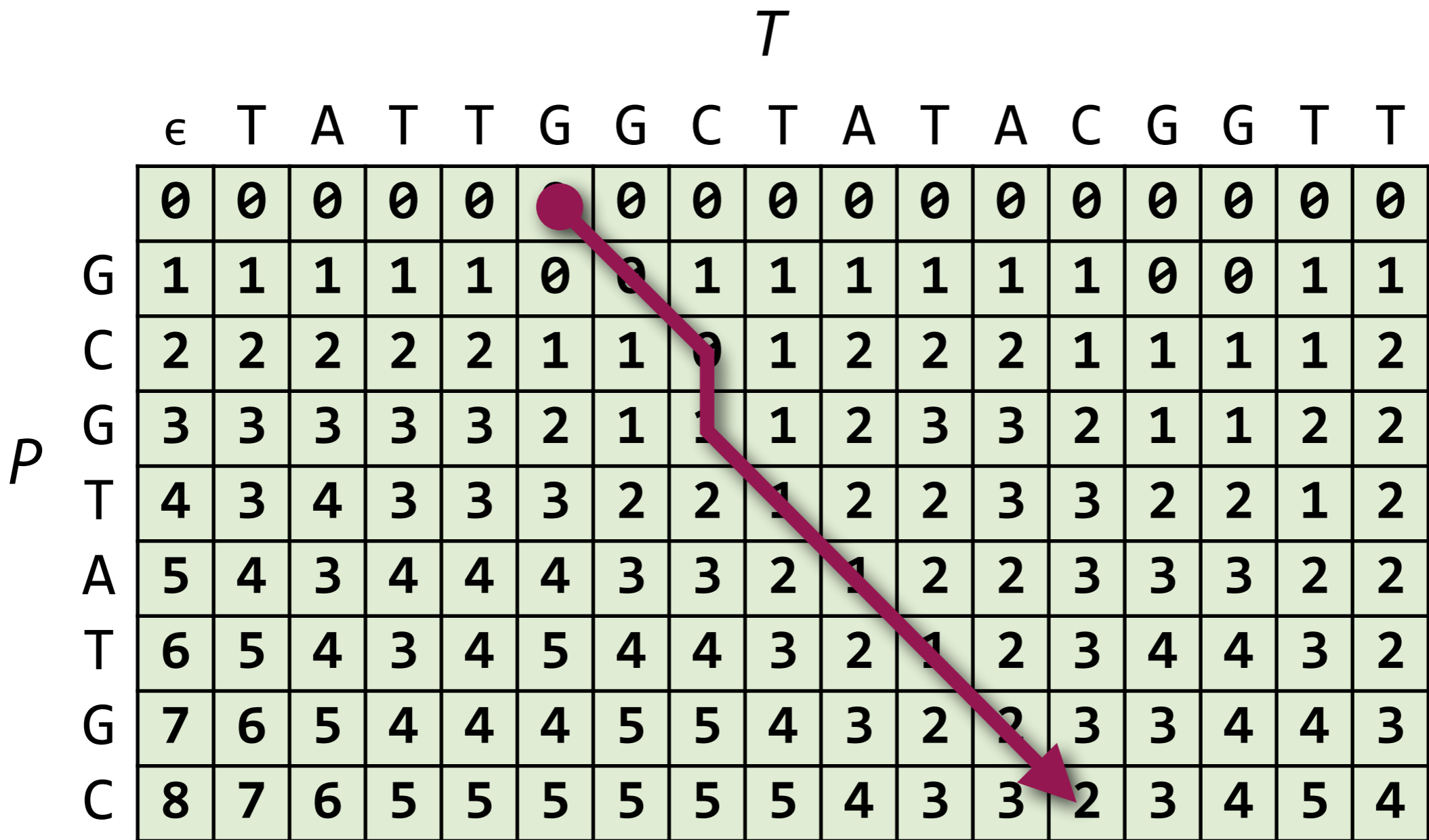
		<i>T</i>																
		$\epsilon$	T	A	T	T	G	G	C	T	A	T	A	C	G	G	T	T
<i>P</i>	G	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	C	1	1	1	1	1	0	0	1	1	1	1	1	1	0	0	1	1
	G	2	2	2	2	2	1	1	0	1	2	2	2	1	1	1	1	2
	T	3	3	3	3	3	2	1	1	1	2	3	3	2	1	1	2	2
	A	4	3	4	3	3	3	2	2	1	2	2	3	3	2	2	1	2
	T	5	4	3	4	4	4	3	3	2	1	2	2	3	3	3	2	2
	G	6	5	4	3	4	5	4	4	3	2	1	2	3	4	4	3	2
	C	7	6	5	4	4	4	5	5	4	3	2	2	3	3	4	4	3
C	8	7	6	5	5	5	5	5	5	4	3	3	2	3	4	5	4	



		<i>T</i>																
		$\epsilon$	T	A	T	T	G	G	C	T	A	T	A	C	G	G	T	T
<i>P</i>	G	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	C	1	1	1	1	1	0	0	1	1	1	1	1	1	0	0	1	1
	G	2	2	2	2	2	1	1	0	1	2	2	2	1	1	1	1	2
	T	3	3	3	3	3	2	1	1	1	2	3	3	2	1	1	2	2
	A	4	3	4	3	3	3	2	2	1	2	2	3	3	2	2	1	2
	T	5	4	3	4	4	4	3	3	2	1	2	2	3	3	3	2	2
	G	6	5	4	3	4	5	4	4	3	2	1	2	3	4	4	3	2
	C	7	6	5	4	4	4	5	5	4	3	2	2	3	3	4	4	3
C	8	7	6	5	5	5	5	5	5	4	3	3	2	3	4	5	4	

		<i>T</i>																
		$\epsilon$	T	A	T	T	G	G	C	T	A	T	A	C	G	G	T	T
<i>P</i>	G	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	C	1	1	1	1	1	0	0	1	1	1	1	1	1	0	0	1	1
	G	2	2	2	2	2	1	1	0	1	2	2	2	1	1	1	1	2
	T	3	3	3	3	3	2	1	1	1	2	3	3	2	1	1	2	2
	A	4	3	4	3	3	3	2	2	1	2	2	3	3	2	2	1	2
	T	5	4	3	4	4	4	3	3	2	1	2	2	3	3	3	2	2
	G	6	5	4	3	4	5	4	4	3	2	1	2	3	4	4	3	2
	C	7	6	5	4	4	4	5	5	4	3	2	2	3	3	4	4	3
C	8	7	6	5	5	5	5	5	5	4	3	3	2	3	4	5	4	

		<i>T</i>																
		$\epsilon$	T	A	T	T	G	G	C	T	A	T	A	C	G	G	T	T
<i>P</i>	G	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	G	1	1	1	1	1	0	0	1	1	1	1	1	1	0	0	1	1
	C	2	2	2	2	2	1	1	0	1	2	2	2	1	1	1	1	2
	G	3	3	3	3	3	2	1	1	1	2	3	3	2	1	1	2	2
	T	4	3	4	3	3	3	2	2	1	2	2	3	3	2	2	1	2
	A	5	4	3	4	4	4	3	3	2	1	2	2	3	3	3	2	2
	T	6	5	4	3	4	5	4	4	3	2	1	2	3	4	4	3	2
	C	7	6	5	4	4	4	5	5	4	3	2	2	3	3	4	4	3
C	8	7	6	5	5	5	5	5	5	4	3	3	2	3	4	5	4	

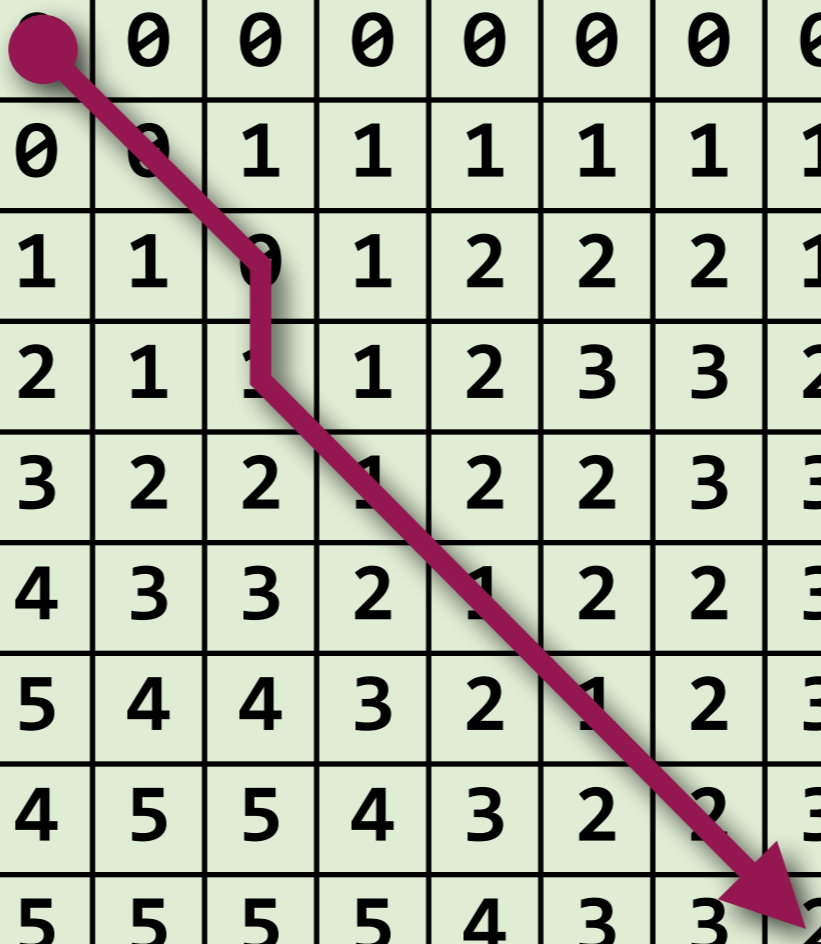
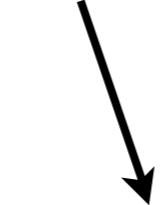


Match occurs at offset 5

*T*

	$\epsilon$	T	A	T	T	G	G	C	T	A	T	A	C	G	G	T	T
<i>offset</i> →	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
G	1	1	1	1	1	0	0	1	1	1	1	1	1	0	0	1	1
C	2	2	2	2	2	1	1	0	1	2	2	2	1	1	1	1	2
G	3	3	3	3	3	2	1	1	1	2	3	3	2	1	1	2	2
T	4	3	4	3	3	3	2	2	1	2	2	3	3	2	2	1	2
A	5	4	3	4	4	4	3	3	2	1	2	2	3	3	3	2	2
T	6	5	4	3	4	5	4	4	3	2	1	2	3	4	4	3	2
G	7	6	5	4	4	4	5	5	4	3	2	2	3	3	4	4	3
C	8	7	6	5	5	5	5	5	5	4	3	3	2	3	4	5	4

*P*

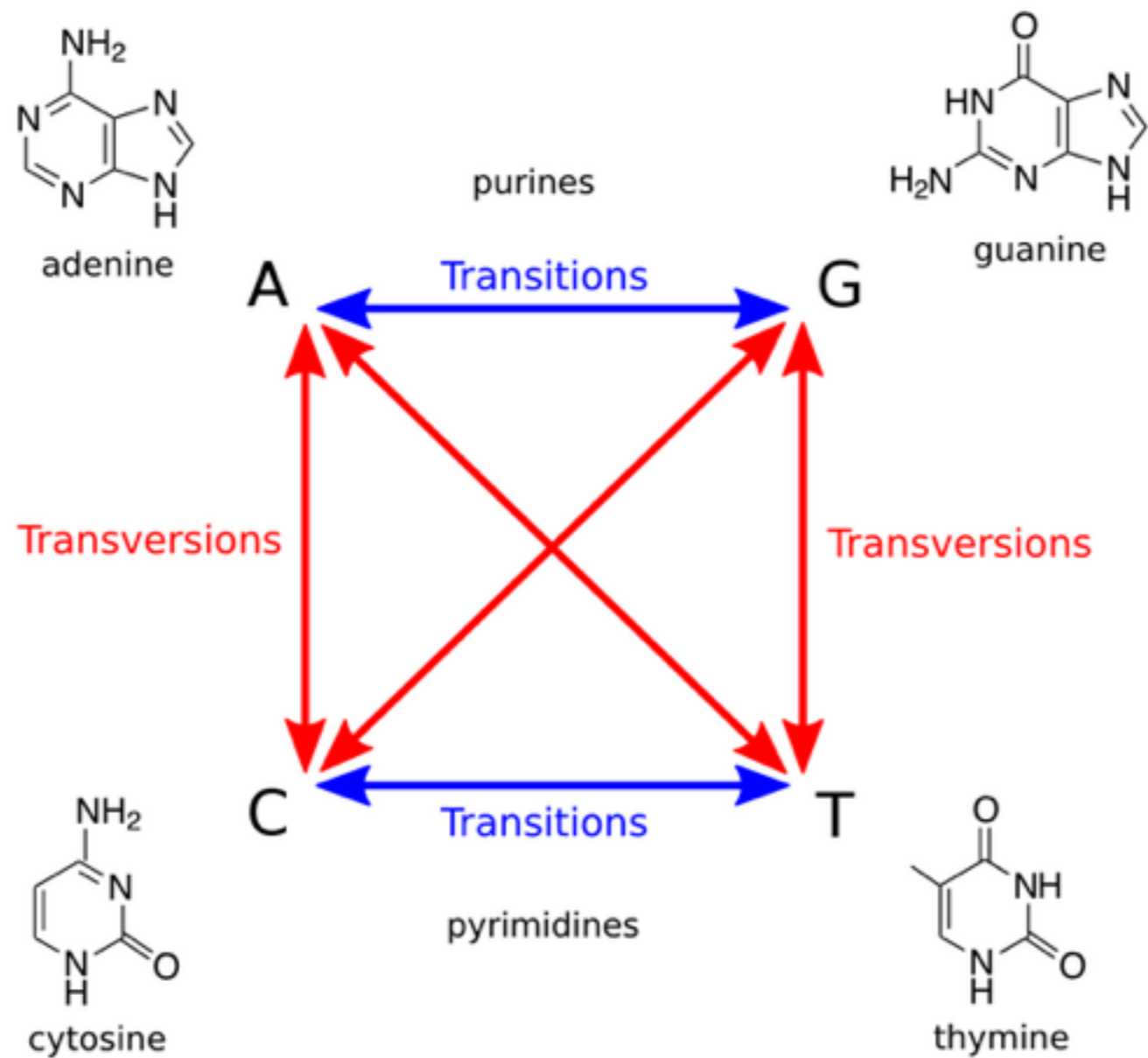


*T*: G C - T A T A C  
 | | | | | | |  
*P*: G C G T A T G C

*T*

		ε	T	A	T	T	G	G	C	T	A	T	A	C	G	G	T	T	
<i>P</i>	G	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
	C	1	1	1	1	1	0	0	1	1	1	1	1	1	1	0	0	1	1
	G	2	2	2	2	2	1	1	0	1	2	2	2	2	1	1	1	1	2
	T	3	3	3	3	3	2	1	1	1	2	3	3	3	2	1	1	2	2
	A	4	3	4	3	3	3	2	2	1	2	2	3	3	2	2	1	2	2
	T	5	4	3	4	4	4	3	3	2	1	2	2	3	3	3	2	2	2
	G	6	5	4	3	4	5	4	4	3	2	1	2	3	4	4	3	2	2
	C	7	6	5	4	4	4	5	5	4	3	2	2	3	3	4	4	3	3
	G	8	7	6	5	5	5	5	5	5	4	3	3	2	3	4	5	4	4

		Y							
		ε	G	C	T	A	T	A	C
X	ε	0	1	2	3	4	5	6	7
	G	1	0	1	2	3	4	5	6
	C	2	1	0	1	2	3	4	5
	G	3	2	1	1	2	3	4	5
	T	4	3	2	1	2	2	3	4
	A	5	4	3	2	1	2	2	3
	T	6	5	4	3	2	1	2	3
	G	7	6	5	4	3	2	2	3
	C	8	7	6	5	4	3	3	2



Human *transition to transversion ratio* (AKA *ti/tv*) is ~2.1



G G G T A G C G G G T T T A A C  
| | | | | | | | | | | | | |  
G G G T A A C G G G T T T A A C

Human substitution rate  $\approx$  1 in 1,000

G G G T A G C G G G T T T A A C  
| | | | | | | | | | | | | |  
G G G T A - - G G G T T T A A C

Small-gap rate is  $\approx$  1 in 3,000

# Penalty matrix

	A	C	G	T	-
A	0	4	2	4	8
C	4	0	4	2	8
G	2	4	0	4	8
T	4	2	4	0	8
-	8	8	8	8	

2 *Transitions (A ↔ G, C ↔ T)*

4 *Transversions*

8 *Gaps*

$$\mathbf{edist}(\alpha x, \beta y) = \min \begin{cases} \mathbf{edist}(\alpha, \beta) + \delta(x, y) \\ \mathbf{edist}(\alpha x, \beta) + 1 \\ \mathbf{edist}(\alpha, \beta y) + 1 \end{cases}$$

$$\mathbf{galign}(\alpha x, \beta y) = \min \begin{cases} \mathbf{galign}(\alpha, \beta) + p(x, y) \\ \mathbf{galign}(\alpha x, \beta) + p(x, -) \\ \mathbf{galign}(\alpha, \beta y) + p(-, y) \end{cases}$$

↑  
Use penalty matrix

# Global alignment

	ε	T	A	T	G	T	C	A	T	G	C
ε	0	8	16	24	32	40	48	56	64	72	80
T	8	0	8	16	24	32	40	48	56	64	72
A	16	8	0	8	16	24	32	40	48	56	64
C	24	16	8	2	10	18	24	32	40	48	56
G	32	24	16	10	2	10	18	26	34	40	48
T	40	32	24	16	10	2	10	18	26	34	42
C	48	40	32	24	18	10	2	10	18	26	34
A	56	48	40	32	26	18	10	2	10	18	26
G	64	56	48	40	32	26	18	10	6	10	18
C	72	64	56	48	40	34	26	18	12	10	10

	A	C	G	T	-
A	0	4	2	4	8
C	4	0	4	2	8
G	2	4	0	4	8
T	4	2	4	0	8
-	8	8	8	8	

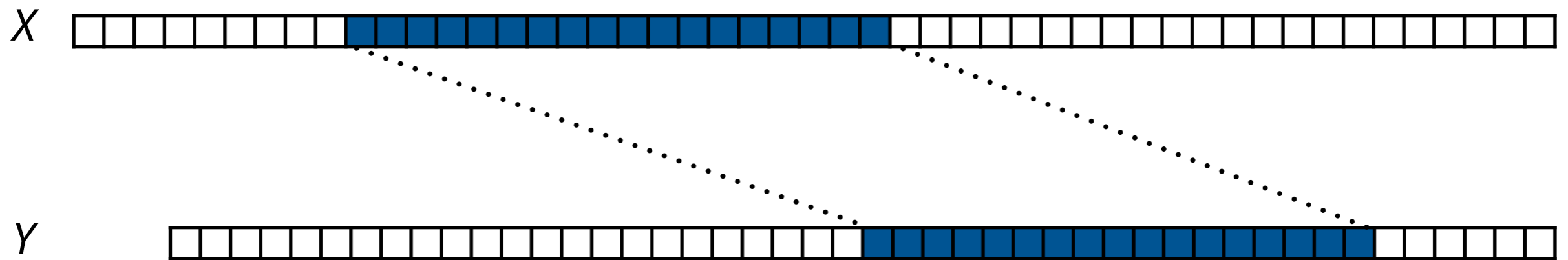
# Global alignment

	ε	T	A	T	G	T	C	A	T	G	C
ε	0	8	16	24	32	40	48	56	64	72	80
T	8	0	8	16	24	32	40	48	56	64	72
A	16	8	0	8	16	24	32	40	48	56	64
C	24	16	8	2	10	18	24	32	40	48	56
G	32	24	16	10	2	10	18	26	34	40	48
T	40	32	24	16	10	2	10	18	26	34	42
C	48	40	32	24	18	10	2	10	18	26	34
A	56	48	40	32	26	18	10	2	10	18	26
G	64	56	48	40	32	26	18	10	6	10	18
C	72	64	56	48	40	34	26	18	12	10	0

	A	C	G	T	-
A	0	4	2	4	8
C	4	0	4	2	8
G	2	4	0	4	8
T	4	2	4	0	8
-	8	8	8	8	

# Local alignment

Find the most similar *pair of substrings* from X and Y



# Local alignment

Find the most similar *pair of substrings* from X and Y

X he\_will\_after\_his\_sour\_fashion\_tell\_you

Y struts\_and\_frets\_his\_hour\_upon\_the\_stage

his\_sour\_  
| | | | | | | | | |  
his\_hour\_

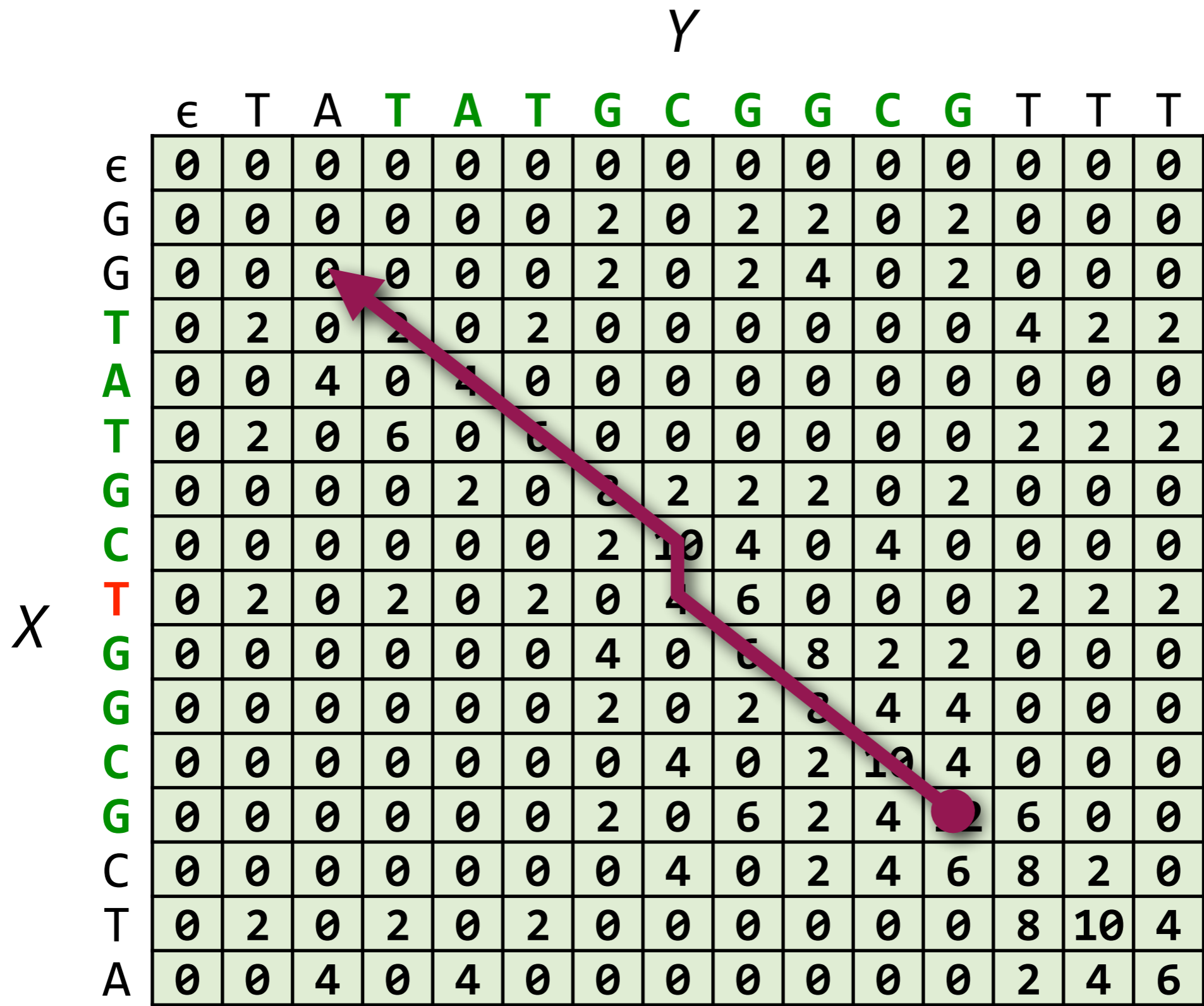
$$\text{lalign}(\alpha x, \beta y) = \max \left\{ \begin{array}{l} \text{lalign}(\alpha, \beta) + s(x, y) \\ \text{lalign}(\alpha x, \beta) + s(x, -) \\ \text{lalign}(\alpha, \beta y) + s(-, y) \\ 0 \end{array} \right.$$

Scoring matrix: matches are positive, differences negative

	A	C	G	T	-
A	2	-4	-4	-4	-6
C	-4	2	-4	-4	-6
G	-4	-4	2	-4	-6
T	-4	-4	-4	2	-6
-	-6	-6	-6	-6	



		Y														
		ε	T	A	T	A	T	G	C	G	G	C	G	T	T	T
X	ε	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	G	0	0	0	0	0	0	2	0	2	2	0	2	0	0	0
	G	0	0	0	0	0	0	2	0	2	4	0	2	0	0	0
	T	0	2	0	2	0	2	0	0	0	0	0	0	4	2	2
	A	0	0	4	0	4	0	0	0	0	0	0	0	0	0	0
	T	0	2	0	6	0	6	0	0	0	0	0	0	2	2	2
	G	0	0	0	0	2	0	8	2	2	2	0	2	0	0	0
	C	0	0	0	0	0	0	2	10	4	0	4	0	0	0	0
	T	0	2	0	2	0	2	0	4	6	0	0	0	2	2	2
	G	0	0	0	0	0	0	4	0	6	8	2	2	0	0	0
	G	0	0	0	0	0	0	2	0	2	8	4	4	0	0	0
	C	0	0	0	0	0	0	0	4	0	2	10	4	0	0	0
	G	0	0	0	0	0	0	2	0	6	2	4	12	6	0	0
	C	0	0	0	0	0	0	0	4	0	2	4	6	8	2	0
	T	0	2	0	2	0	2	0	0	0	0	0	0	8	10	4
A	0	0	4	0	4	0	0	0	0	0	0	0	2	4	6	



Y T A T G C - G G C G  
 | | | | | | | | |  
 X T A T G C T G G C G