String Algorithms and Data Structures
Boyer-Moore

CS 199-225
Brad Solomon

February 20, 2023
Exact Pattern Matching \textit{w/} Z-algorithm

Pattern, $P$ \quad Text, $T$

Naive $\approx \theta(|P| + |T|)$ \quad Z-Algorithm $\approx \theta(|P| + |T|)$

Find instances of $P$ in $T$

‘instances’: An exact, full length copy
Why continue?

The Z-algorithm is:

The Z-algorithm is: $O(|P| + |T|)$ time

An alphabet-independent solution

The Z-algorithm is less good at:

Searching for a **set** of patterns (Aho-Corasick)

Running in *sub-linear* time (Boyer-Moore)

* — in practice, not theory
Exact pattern matching \textit{w/} Boyer-Moore

Boyer Moore \textit{preprocesses} the pattern

\[ \text{Preprocess} \approx O(|P|) \]

\[ \text{Boyer-Moore} \approx O(|P| + |T|) \]

Find instances of \( P \) in \( T \)

‘instances’: An exact, full length copy
Boyer-Moore

**Intuition:** Learn from alignments to avoid others

\[ P: \text{cat} \]
\[ T: \text{carl carried the cat} \]
\[ cat \]

\[ 0 \quad 1 \quad 2 \quad 3 \quad 4 \quad 5 \quad 6 \quad 7 \quad 8 \quad 9 \quad \ldots \]

What does this alignment tell us?
Boyer-Moore

**Intuition:** Learn from alignments to avoid others

\[ P: \text{cat} \]

\[ T: \text{carl carried the cat} \]

\[ \text{cat} \]

\[ \theta 1 2 3 4 5 6 7 8 9 \ldots \]

What does this alignment tell us?

1) Our pattern doesn’t match at this alignment

There is no ‘r’ in ‘cat’!
Intuition: Learn from alignments to avoid others

P: cat

T: carl carried the cat
cat

What does this alignment tell us?

2) Our pattern doesn’t match at later alignments

There is no ‘r’ in ‘cat’!
Boyer-Moore

Intuition: Learn from alignments to avoid others

\[ P: \text{cat} \]

\[ T: \text{carl carried the cat} \]

\[ \text{cat} \]

\[ 0 \ 1 \ 2 \ 3 \ 4 \ 5 \ 6 \ 7 \ 8 \ 9 \ldots \]

What does this alignment tell us?

2) Our pattern doesn’t match at later alignments

There is no ‘r’ in ‘cat’!
Boyer-Moore

**Intuition:** Learn from alignments to avoid others

\[ P: \text{cat} \]

\[ T: \text{carried the cat} \]

\[ \text{car} \quad \text{skip!} \]

\[ \text{cat} \quad \text{skip!} \]

What does this alignment tell us?

2) Our pattern doesn’t match at *later* alignments

\[ \text{car} \quad \text{There is no ‘r’ in ‘cat’!} \]

\[ \text{cat} \]
Intuition: Learn from alignments to avoid others

\[ P: \text{word} \]

\[ T: \text{There would have been a } \ldots \]

\[ \text{----word---- } \]

\[ 0123456789\ldots \]
Boyer-Moore

**Intuition:** Learn from alignments to avoid others

\[
P: \text{word}
\]
\[
T: \text{There would have been a ...}
\]
\[
\text{----------------word-----------------}
\]
\[
0123456789...
\]

1) Our pattern doesn’t match at this alignment

\[
T: \text{woul}
\]
\[
P: \text{word}
\]
Boyer-Moore

**Intuition:** Learn from alignments to avoid others

\[ P: \text{word} \]

\[ T: \text{There would have been a } \ldots \]

\[ \text{word} \]

\[ 0 \ 1 \ 2 \ 3 \ 4 \ 5 \ 6 \ 7 \ 8 \ 9 \ldots \]

How many alignments can we skip?

2) Our pattern doesn’t match at later alignments

\[ T: \text{woul} \]

\[ P: \text{word} \]

There is no ‘u’ in ‘word’!
Intuition: Learn from alignments to avoid others

\[ P: \text{word} \]
\[ T: \text{There would have been a } \ldots \]
\[ \underline{\text{word}} \]
\[ 0 1 2 3 4 5 6 7 8 9 \ldots \]

How many alignments can we skip? 2

2) Our pattern doesn’t match at later alignments

\[ T: \text{woul} \]
\[ P: \text{word} \]

There is no ‘u’ in ‘word’!
Boyer-Moore

**Intuition:** Learn from alignments to avoid others

\[ P: \text{word} \]

\[ T: \text{There would have been a …} \]

\[ \text{-------------------------word} \]

\[ \text{word} \text{ skip!} \]

\[ \text{word} \text{ skip!} \]

\[ \text{word} \]

How many alignments can we skip? 2

2) Our pattern doesn’t match at *later* alignments

\[ T: \text{woul} \]

\[ P: \text{word} \]

There is no ‘u’ in ‘word’!
Boyer-Moore

**Intuition:** Learn from alignments to avoid others

\[ P: \text{T A G A C} \]

\[ T: \text{G T A G A T G G C T G A T C G A G T A G C G G C G} \]

There IS a T in ‘TAGAC’!

How many alignments can we skip? 3
Boyer-Moore

**Intuition:** Learn from alignments to avoid others

\[ P: T A G A C \]

\[ T: G T A G A T G G C T G A T C G A G T A G C G G C G \]

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

How many alignments can we skip? 3

TAGAT

TAGAC

There IS a T in ‘TAGAC’!
Boyer-Moore

**Intuition:** Learn from alignments to avoid others

\[ P: \text{A A B B B} \]

\[ T: \text{A A A B A B A A A A A A A A A A A A A A A A A A A A} \]

\[ \overset{\cdot}{\text{A A B B B}} \]

---

How many alignments can we skip? 1

AABAB

There IS an A in ‘AAABB’!
Boyer-Moore

**Intuition:** Learn from alignments to avoid others

\[ P: \ A \ A \ B \ B \ B \]

\[ T: \ A \ A \ A \ B \ A \ B \ A \ A \ A \ A \ A \ A \ A \ A \ A \ A \ A \ A \ A \ A \ A \]

\[ \cdots \cdot \ \ A \ A \ B \ B \ B \]

\[ \ \ A \ A \ B \ B \ B \quad \text{skip!} \]

\[ \ \ A \ A \ B \ B \ B \quad \text{the first match we encounter!} \]

How many alignments can we skip? 1

\[ \text{AABAB} \quad \text{There IS an A in ‘AAABB’!} \]

\[ \text{AABBAB} \]

\[ \text{AABBB} \]
Boyer-Moore: Bad Character rule

Upon mismatch, skip alignments until (a) mismatch becomes a match, or (b) \( P \) moves past mismatched character. (c) If there was no mismatch, don't skip
Boyer-Moore: Bad Character rule

Step 1:

\[ T: \text{CCTTCTGCTACTTTTGCGCGCGCGCGGAA} \]
\[ P: \text{CCTTTTGCG} \]
\[ \text{skip!} \]

Step 2:

\[ T: \text{CCTTCTGCTACCTTTTGCGCGCGCGCGGAA} \]
\[ P: \text{CCTTTTGCG} \]

Step 3:

\[ T: \text{CCTTCTGCTACCTTTTGCGCGCGCGCGGAA} \]
\[ P: \text{CCTTTTGCG} \]
\[ \text{skip!} \]

We skipped three alignments

Can we do anything to make this better?
Boyer-Moore: Bad Character rule

Which of the following alignments skips the most?

A) $T$: TATAT...
   $P$: TAGAC

B) $T$: TGAT...
   $P$: TAGAC

C) $T$: TAGAT...
   $P$: TAGAC

D) $T$: TAGTT...
   $P$: TAGAC
Boyer-Moore: Bad Character rule improvement

Continue to test alignment from left-to-right

… but compare characters from right to left.

$P$: T A G A C

$T$: G T A G A T G G C T G A T C G A G T A G C G G C G

- T A G A C
There would have been a …

There is no ‘l’ in ‘word’!

How many alignments do we skip?
Right-to-left-scanning w/ BC Rule

$P: \text{word}$

$T: \text{There would have been a ...}$

\[\text{word} \quad \text{word} \quad \text{word} \quad \text{word} \quad \text{word}\]

How many alignments do we skip? 3
Right-to-left-scanning w/ BC Rule

Upon mismatch, skip alignments until (a) mismatch becomes a match, or (b) $P$ moves past mismatched character. (c) If there was no mismatch, don't skip

<table>
<thead>
<tr>
<th>Step 1: $T$: C C T T C T G C T A C C T T T T G C G C G C G C G C G G A A</th>
</tr>
</thead>
<tbody>
<tr>
<td>$P$: C C T T T T G C</td>
</tr>
<tr>
<td><strong>Case (a)</strong></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Step 2: $T$: C C T T C T G C T A C C T T T T G C G C G C G C G C G C G G A A</th>
</tr>
</thead>
<tbody>
<tr>
<td>$P$: C C T T T T G C</td>
</tr>
<tr>
<td><strong>Case (b)</strong></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Step 3: $T$: C C T T C T G C T A C C T T T T G C G C G C G C G C G C G G A A</th>
</tr>
</thead>
<tbody>
<tr>
<td>$P$: C C T T T T G C</td>
</tr>
<tr>
<td><strong>Case (c)</strong></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Step 4: $T$: C C T T C T G C T A C C T T T T G C G C G C G C G C G G A A</th>
</tr>
</thead>
<tbody>
<tr>
<td>$P$: C C T T T T G C</td>
</tr>
<tr>
<td><strong>Case (a)</strong></td>
</tr>
</tbody>
</table>

(etc)
Right-to-left-scanning w/ BC Rule

Step 1:

T: CCTTTCTGCTACCTTTTGCGCGCGCGGAA
P: CCTTTTGGC

Step 2:

T: CCTTTCTGCTACTCTTTTGCGCGCGCGGAA
P: CCTTTTGGC

Step 3:

T: CCTTTCTGCTACTCTTTTGCGCGCGCGGAA
P: CCTTTTGGC

Up to step 3, we skipped 8 alignments

5 characters in T were never looked at
Right-to-left-scanning w/ BC Rule

Learn from character comparisons to skip pointless alignments

1. When we hit a mismatch $c$, move $P$ along until $c$ becomes a match (or $P$ moves past $c$) "Bad character rule"

2. Try alignments in one direction, but do character comparisons in opposite direction "Right-to-left scanning"

How do we put the first two rules in practice?

Exact pattern matching w/ Boyer-Moore

Boyer Moore preprocesses the pattern

\[ \text{Preprocess} \approx O( |P| ) \]

Boyer-Moore \( \approx O( |P| + |T| ) \)

Find instances of \( P \) in \( T \)

‘instances’: An exact, full length copy
Boyer-Moore: BC rule preprocessing

Preprocessing requires two args: \( P: \text{T C G C} \) \( \Sigma: \text{A C G T} \)

The goal is to produce a table which tracks skips
Boyer-Moore: BC rule preprocessing

Preprocessing requires two args: \( P: T C G C \quad \Sigma: A C G T \)

The goal is to produce a table which tracks *skips*
Boyer-Moore: BC rule preprocessing

Preprocessing requires two args: \( P: \text{T C G C} \) \( \Sigma: \text{A C G T} \)

The goal is to produce a table which tracks *skips*

\[
\begin{array}{cccc}
  & T & C & G & C \\
\hline
A & \ & \ & \ & \ \\
C & \ & \ & \ & \ \\
G & \ & \ & \ & \ \\
T & \ & \ & \ & 2 \\
\end{array}
\]
Boyer-Moore: BC rule preprocessing

Preprocessing requires two args: \( P: T \ C \ G \ C \), \( \Sigma: A \ C \ G \ T \)

The goal is to produce a table which tracks *skips*
Boyer-Moore: BC rule preprocessing

Preprocessing requires two args: \( P: T\ C\ G\ C \) \( \Sigma: A\ C\ G\ T \)

The goal is to produce a table which tracks *skips*

<table>
<thead>
<tr>
<th>( \Sigma )</th>
<th>T</th>
<th>C</th>
<th>G</th>
<th>C</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td></td>
<td></td>
<td></td>
<td>3</td>
</tr>
<tr>
<td>C</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>G</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>T</td>
<td></td>
<td></td>
<td>2</td>
<td></td>
</tr>
</tbody>
</table>

\( T: ?\ ?\ ?\ A ?\ ?\ ?\ ?\ ?\ ?\ ? \)

\( P: T\ C\ G\ C \)
Boyer-Moore: BC rule preprocessing

Preprocessing requires two args: \( P: T C G C \) \( \Sigma: A C G T \)

The goal is to produce a table which tracks *skips*

\[
\begin{array}{|c|c|c|c|}
\hline
 & T & C & G & C \\
\hline
A & 0 & 1 & 2 & 3 \\
C & 0 & - & 0 & - \\
G & 0 & 1 & - & 0 \\
T & - & 0 & 1 & 2 \\
\hline
\end{array}
\]

\( P \): T C G C

\( P \): T C G C

\( P \): T C G C

\( P \): T C G C
Boyer-Moore: BC rule preprocessing

Preprocessing requires two args: \( P: B A B A A A B \) \( \Sigma: A B \)
Boyer-Moore: BC rule preprocessing

Preprocessing requires two args: \( P: \text{BABAAAB} \) \( \Sigma: \text{AB} \)

For each character \( p \) in pattern \( P \)

For each character \( c \) in alphabet \( \Sigma \)

Find the closest previous instance of \( p \) (to the left of \( c \)).

<table>
<thead>
<tr>
<th>( \Sigma )</th>
<th>B</th>
<th>A</th>
<th>B</th>
<th>A</th>
<th>A</th>
<th>A</th>
<th>B</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>0</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>B</td>
<td>0</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Boyer-Moore: BC rule preprocessing

Preprocessing requires two args: \( P: \text{B A B A A A B} \quad \Sigma: \text{A B} \)

For each character \( p \) in pattern \( P \)

For each character \( c \) in alphabet \( \Sigma \)

Find the closest previous instance of \( p \) (to the left of \( c \)).

<table>
<thead>
<tr>
<th>Pattern</th>
<th>( P )</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>B A B A A A B</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>( \Sigma )</th>
<th>A</th>
<th>B</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>B</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>
Boyer-Moore: BC rule preprocessing

Preprocessing requires two args: \( P: B A B A A A B \) \( \Sigma: A B \)

For each character \( p \) in pattern \( P \)

For each character \( c \) in alphabet \( \Sigma \)

Find the closest previous instance of \( p \) (to the left of \( c \)).
Learning Objective:

**Implement preprocessing of patterns with Boyer-Moore**

Observe Boyer-Moore* efficiency *as a heuristic*

Consider: Optimal preprocessing is $\theta(|P| |\Sigma|)$. Can you code it?
Boyer-Moore: Using the BC Table

Try alignments from left-to-right and match characters from right-to-left

When we encounter a mismatch, skip the calculated number of alignments

<table>
<thead>
<tr>
<th></th>
<th>T</th>
<th>C</th>
<th>G</th>
<th>C</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>C</td>
<td>0</td>
<td>-</td>
<td>0</td>
<td>-</td>
</tr>
<tr>
<td>G</td>
<td>0</td>
<td>1</td>
<td>-</td>
<td>0</td>
</tr>
<tr>
<td>T</td>
<td>-</td>
<td>0</td>
<td>1</td>
<td>2</td>
</tr>
</tbody>
</table>

\[ T: \text{TTTTTT} \]
\[ P: \text{TCGC} \]
\[ \Sigma \]
Boyer-Moore: Using the BC Table

Try alignments from left-to-right and match characters from right-to-left

When we encounter a mismatch, skip the calculated number of alignments

$$
\begin{array}{c|cccc}
\Sigma & T & C & G & C \\
A & 0 & 1 & 2 & 3 \\
C & 0 & - & 0 & - \\
G & 0 & 1 & - & 0 \\
T & - & 0 & 1 & 2 \\
\end{array}
$$

$$
P: T \ C \ G \ C
$$

$$
T: G G G G G G G G G G
$$
Boyer-Moore: Using the BC Table

Try alignments from left-to-right and match characters from right-to-left

When we encounter a mismatch, skip the calculated number of alignments

\[ \Sigma \]

\[ \begin{array}{c|cccc} \ & T & C & G & C \\
A & 0 & 1 & 2 & 3 \\
C & 0 & - & 0 & - \\
G & 0 & 1 & - & 0 \\
T & - & 0 & 1 & 2 \\
\end{array} \]

\[ T: \text{AATAGC} \]

\[ P: \text{TCGCC} \]
Boyer-Moore: Tracking total skips

\[\begin{array}{c|cc}
\Sigma & A & A \\
\hline 
A & 0 & 0 \\
B & 0 & 1 \\
\end{array}\]

\[p\]

\[T: \text{B B B B B} \]

\[T: \text{B B B B B} \]

\[T: \text{B B B B B} \]

\[T: \text{B B B B B} \]
Boyer-Moore: Tracking total skips

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>A</th>
<th>A</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>B</td>
<td>0</td>
<td>1</td>
<td>2</td>
</tr>
</tbody>
</table>

$\Sigma$ $P$

$T$: B B B B
Assignment 4: a_bmoore

Learning Objective:

Implement preprocessing of patterns with Boyer-Moore*

Observe Boyer-Moore* efficiency as a heuristic

Consider: Our Boyer-Moore is theoretically slower than Z-algorithm.

But is it slower in practice? What is our total character comparisons?
A complete bonus lecture!
A better Boyer-Moore

Learn from character comparisons to skip pointless alignments

1. When we hit a mismatch $c$, move $P$ along until $c$ becomes a match (or $P$ moves past $c$) “Bad character rule”

2. Try alignments in one direction, but do character comparisons in *opposite* direction “Right-to-left scanning”

Is this $O(|P| + |T|)$?
Worst-Case Bad Character rule

Upon mismatch, skip alignments until (a) mismatch becomes a match, or (b) $P$ moves past mismatched character. (c) If there was no mismatch, don't skip.

**Step 1:**

$T$: A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A A
A better Boyer-Moore

The complete Boyer-Moore algorithm, \textit{with all refinements}, is $O(|P| + |T|)$.

Refinements include:

- "strong" good suffix rule
- Galil rule

We will be covering the ‘weak’ good suffix rule

If interested in refinements, see Gusfield textbook (syllabus) or contact me for details
“Weak” Good Suffix rule

**Intuition:** Learn from alignments to avoid others

\[
P: \ A \ C \ A \ T \ A \ C
\]
\[
T: \ T \ A \ C \ A \ G\boxed{A\ C}\ A \ T \ A \ C \ A \ T \ G \ A \ C \ A \ G \ T \ G \ A \ C \ C \ A
\]

What does this alignment tell us?
“Weak” Good Suffix rule

**Intuition:** Learn from alignments to avoid others

\[ P: \text{A C A T A C} \]
\[ T: \text{T A C A G}_\text{AC}_\text{A T A C A T G A C A G T G A C C A} \]
\[ \text{A C A T A C} \]
\[ \text{A C A T A C} \]

We only want to look at alignments that are *at least as good* as our current alignment
“Weak” Good Suffix rule

**Intuition:** Learn from alignments to avoid others

\[ P: \text{ACCATAC} \]
\[ T: \text{TACACGACATACATGAACACG} \]

What does partial match (the suffix ‘AC’) tell us?

Any alignment that overlaps this region of the text must match the suffix! So we can look for another ‘AC’ somewhere in the pattern!
“Weak” Good Suffix rule

**Intuition:** Learn from alignments to avoid others

\[ P: \text{ACATAC} \]

\[ T: \text{TACAGACATACATGACAGTGACCA} \]

Any alignment that overlaps this region of the text must match the suffix! So we can look for another ‘AC’ somewhere in the pattern!
“Weak” Good Suffix rule

**Intuition:** Learn from alignments to avoid others

\[ P: \text{A C A T A C} \]

\[ T: \text{T A C A GACATACATGACAGTGACCA} \]

How many alignments do we skip? 3

Any alignment that overlaps this region of the text must match the suffix! So we can look for another ‘AC’ somewhere in the pattern!
“Weak” Good Suffix rule

**Intuition:** Learn from alignments to avoid others

\[ P: \text{A T C} \]
\[ T: \text{A G T A G C A G C A T A G C A G C A G C T A G A} \]

Any alignment that overlaps this region of the text must match the suffix! So we can look for another _______ somewhere in the pattern!

How many alignments do we skip?
“Weak” Good Suffix rule

**Intuition:** Learn from alignments to avoid others

\[ P: \text{A T C} \]
\[ T: \text{A G T A G}^{\text{C}} \text{A G C A C A G T A G C A G C T A G A} \]

How many alignments do we skip? 2

Any alignment that overlaps this region of the text must match the suffix! So we can look for another \text{C} somewhere in the pattern!
“Weak” Good Suffix rule

**Intuition:** Learn from alignments to avoid others

\[ P: \ G\ C\ A\ G\ C \]

\[ T: \ A\ G\ T\ \boxed{A\ G\ C}\ \ A\ G\ C\ \ A\ C\ A\ G\ T\ A\ G\ C\ \ A\ G\ C\ \ T\ A\ G\ A \]

\[ \boxed{G\ C\ A\ G\ C} \]

Any alignment that overlaps this region of the text must match the suffix! So we can look for another _______ somewhere in the pattern!

How many alignments do we skip?
“Weak” Good Suffix rule

**Intuition:** Learn from alignments to avoid others

\[ P: \ G \ C \ A \ G \ C \]
\[ T: \ A \ G \ T \ AGC \ AGC \ A \ C \ A \ G \ T \ A \ G \ C \ A \ G \ C \ T \ A \ G \ A \]

This is a full length match!

How many alignments do we skip?
“Weak” Good Suffix rule

**Intuition:** Learn from alignments to avoid others

\[ P: \quad G \ C \ A \ G \ C \]
\[ T: \quad A \ G \ T \boxed{A \ G \ C} \ A \ G \ C \ A \ C \ A \ G \ T \ A \ G \ C \ A \ G \ C \ T \ A \ G \ A \]
\[ \boxed{G \ C \ A \ G \ C} \]

Any alignment that overlaps this region of the text must match the suffix … *or have a prefix-suffix partial match!*

How many alignments do we skip?
"Weak" Good Suffix rule

Let $t =$ longest suffix match at alignment; skip until (a) we find another 
**instance** of $t$ **or** (b) $P$ moves past $t$

Step 1: 
$T$: C G T G C T A C T T A C T T A C T T A C T T A C G C G A A 
$P$: C T T A C T T A C $t$ occurs *in its entirety* to the left within $P$

Step 2: 
$T$: C G T G C C T A C T T A C T T A C T T A C G C G A A 
$P$: C T T A C T T A C $t$ 

Step 3: 
$T$: C G T G C C T A C T T A C T T A C T T A C T T A C G C G A A 
$P$: C T T A C T T A C $t$

An **instance** of $t$ is either a full match to the left within $P$ 
or a **prefix** of $P$ matches a **suffix** of $t$
Boyer-Moore: Putting it together

How to combine bad character and good suffix rules?

How many characters does bad character skip? 2 characters

How many characters does good suffix skip? 7 characters

Take the maximum (7)!
Boyer-Moore: Putting it together

Use bad character or good suffix rule, *whichever skips more*

Step 1: 
*T:* GTTATAGC\(\textcolor{red}{T}\)GATCGCGGTAGCGGCGAA 
*P:* G\(\textcolor{red}{T}\)AGCGGGCG 
bc: 6, gs: 0  *bad character*

Step 2: 
*T:* GTTATAGC\(\textcolor{red}{T}\)GATCG\(\textcolor{orange}{G}\)CGGTAGCGGCGAA 
*P:* G\(\textcolor{orange}{T}\)AGCGGGCG 
bc: 0, gs: 2  *good suffix*

Step 3: 
*T:* GTTATAGC\(\textcolor{red}{T}\)GATCG\(\textcolor{orange}{G}\)CGGCGGTAGCGGCGAA 
*P:* G\(\textcolor{orange}{T}\)AGCGGGCG 
bc: 2, gs: 7  *good suffix*

Step 4: 
*T:* GTTATAGC\(\textcolor{red}{T}\)GATCG\(\textcolor{orange}{G}\)CGGCG\(\textcolor{red}{G}\)GTAGCGGCGAA 
*P:* G\(\textcolor{orange}{T}\)AGCGGGCG 
bc: 6, gs: 0  *bad character*
Boyer-Moore: Putting it together

11 characters of T ignored completely!

Step 1:

\[
T: \text{GTTATAGC} \quad P: \text{GTAGCGCGGCGTAGCGGCGAA}
\]

Step 2:

\[
T: \text{GTTATAGCTGATCGCGGCGTAGCGGCGAA} \quad P: \text{GTAGCGCG}
\]

Step 3:

\[
T: \text{GTTATAGCTGATCGCGGCGTAGCGGCGAA} \quad P: \text{GTAGCGCG}
\]

Step 4:

\[
T: \text{GTTATAGCTGATCGCGGCGTAGCGGCGAA} \quad P: \text{GTAGCGCG}
\]

Skipped 15 alignments
Boyer-Moore

Learn from character comparisons to skip pointless alignments

1. When we hit a mismatch $c$, move $P$ along until $c$ becomes a match (or $P$ moves past $c$) “Bad character rule”

2. Try alignments in one direction, but do character comparisons in opposite direction “Right-to-left scanning”

3. When we move $P$ along, make sure characters that matched in the last alignment also match in the next alignment “Good suffix rule”