String Algorithms and Data Structures

Z-values and the Z-algorithm

CS 199-225
Brad Solomon

January 31, 2022
Assignment 1: a_naive due today!

Don’t forget to submit feedback on Moodle

About how many hours did you spend in total on this assignment?

- Under 1 hour
- Between 1-2 hours
- Between 2-3 hours
- Between 3-4 hours
- Over 4 hours

The lecture was helpful for completing this assignment.

- 1 - Strongly disagree
- 2 - Disagree
- 3 - Neither agree nor disagree
- 4 - Agree
- 5 - Strongly agree

After completing this assignment, I have a good understanding of the material taught.

- 0 - I already knew the material
- 1 - Strongly disagree
- 2 - Disagree
- 3 - Neither agree nor disagree
- 4 - Agree
- 5 - Strongly agree
Exact Pattern Matching

Pattern, $P$  
Text, $T$

Find instances of $P$ in $T$

‘instances’: An exact, full length copy
What’s a simple algorithm for exact matching?

P: word
T: There would have been a time for such a word
    word word word word word word word word word word word
    word word word word word word word word word word word
    word word word word word word word word word word word
    word word word word word word word word word word word
    word word word word word word word word word word word
    word word word word word word word word word word word
    word word word word word word word word word word word
    word word word word word word word word word word word
    word word word word word word word word word word word

Try all possible alignments. For each, check if it matches. This is the naïve algorithm.
Exact Pattern Matching

What is good about the naive solution?

What is bad?
Exact Pattern Matching

What is our time complexity? \( (n = |P|, \quad m = |T|) \)

\( (\text{# of alignments}) \times (\text{cost of an alignment}) \)
Exact Pattern Matching

What is our time complexity? \((n = |P|, \quad m = |T|)\)

\((\# \text{ of alignments}) \times (\text{cost of an alignment})\)

\(P: \quad \underbrace{\leftarrow n \rightarrow}_{\text{P}}\)

\(T: \quad \underbrace{\leftarrow m \rightarrow}_{\text{T}}\)

\(P\) can fit at each `position' along \(T\) except the \text{edge}
Exact Pattern Matching

What is our time complexity? \( (n = |P|, \ m = |T|) \)

\[
(\text{___________}) \times \text{(cost of an alignment)}
\]

\(P: \text{aaaa} \)

\(T: \text{aaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaa} \)

\[
\text{aaaa aaaa aaaa aaaa aaaa aaaa aaaa aaaa aaaa aaaa aaaa}
\]

\[
\text{aaaa aaaa aaaa aaaa aaaa aaaa aaaa aaaa aaaa aaaa aaaa}
\]

\[
\text{aaaa aaaa aaaa aaaa aaaa aaaa aaaa aaaa aaaa aaaa aaaa}
\]

\[
\text{aaaa aaaa aaaa aaaa aaaa aaaa aaaa aaaa aaaa aaaa aaaa}
\]

\[
\text{aaaa aaaa aaaa aaaa aaaa aaaa aaaa aaaa aaaa aaaa aaaa}
\]

\[
\text{aaaa aaaa aaaa aaaa aaaa aaaa aaaa aaaa aaaa aaaa aaaa}
\]

There are ______ positions which extend past the edge of T
Exact Pattern Matching

What is our time complexity? \( (n = |P|, \quad m = |T|) \)

\( (m-n+1) \times \text{(cost of an alignment)} \)

\( P: \) aaaa

\( T: \) aaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaa

aaaa aaaa aaaa aaaa aaaa aaaa aaaa aaaa aaaa aaaa
aaaa aaaa aaaa aaaa aaaa aaaa aaaa aaaa aaaa
aaaa aaaa aaaa aaaa aaaa aaaa aaaa aaaa aaaa
aaaa aaaa aaaa aaaa aaaa aaaa aaaa aaaa aaaa
aaaa aaaa aaaa aaaa aaaa aaaa aaaa aaaa aaaa
aaaa aaaa aaaa aaaa aaaa aaaa aaaa aaaa aaaa
aaaa aaaa aaaa aaaa aaaa aaaa aaaa aaaa aaaa
aaaa aaaa aaaa aaaa aaaa aaaa aaaa aaaa aaaa
aaaa aaaa aaaa aaaa aaaa aaaa aaaa aaaa aaaa

Each alignment compares __________ characters.
Exact Pattern Matching

What is our time complexity?

\( n = |P|, \quad m = |T| \)

\[ \theta \left( (m - n + 1) \times n \right) \]
String Algorithms in Genomics

P: Read (n = ~50-150)

T: Reference (m = ~3 billion)
String Algorithms in Genomics
Improving exact pattern matching

How can we do better than the naïve algorithm?

… If we have infinite space?

… If I tell you the pattern ahead of time?

… If I tell you the text ahead of time?
Exact Pattern Matching w/ Z-algorithm

Pattern, $P$  
Text, $T$

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

Find instances of $P$ in $T$

‘instances’: An exact, full length copy
The Z-value $[ Z_i(S) ]$

Given a string $S$, $Z_i(S)$ is the length of the longest substring in $S$, starting at position $i$, that matches a prefix of $S$.

$S$: TTCGTTAGCG

<table>
<thead>
<tr>
<th></th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
</tr>
</thead>
<tbody>
<tr>
<td>$Z_0(S)$ =</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$Z_1(S)$ =</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$Z_2(S)$ =</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$Z_3(S)$ =</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$Z_4(S)$ =</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$Z_5(S)$ =</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
The Z-value \[ Z_i(S) \]

Given a string \( S \), \( Z_i(S) \) is the length of the longest substring in \( S \), starting at position \( i \), that matches a prefix of \( S \).

\[
\begin{align*}
0 & \quad 1 & \quad 2 & \quad 3 & \quad 4 & \quad 5 & \quad 6 & \quad 7 & \quad 8 & \quad 9 \\
S: & \quad T & \quad T & \quad C & \quad G & \quad T & \quad T & \quad A & \quad G & \quad C & \quad G \\
Z_0(S) = 10 & \quad Z_3(S) = & \quad Z_1(S) = 1 & \quad Z_4(S) = & \quad Z_2(S) = 0 & \quad Z_5(S) =
\end{align*}
\]
The Z-value $[ Z_i(S) ]$

Given a string $S$, $Z_i(S)$ is the length of the longest substring in $S$, starting at position $i > 0$, that matches a prefix of $S$.

\[
\begin{array}{cccccccccc}
0 & 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 \\
\end{array}
\]

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

\[
\begin{align*}
Z_0(S) &= 10 \\
Z_1(S) &= 1 \\
Z_2(S) &= 0 \\
Z_3(S) &= 0 \\
Z_4(S) &= 2 \\
Z_5(S) &= 1 \\
\end{align*}
\]
Calculating the Z-values

**Naive:** Compute the Z-values by *explicitly* comparing characters (left-to-right scan):

\[ Z_1 = \]

\[ \text{AAAABAACAABAA...} \quad \text{AAAABAAACAABAA...} \]

\[ Z_5 = \]

\[ \text{AAAABAACAABAA...} \quad \text{AAAABAACAABAA...} \]

What is our time complexity?
Calculating the Z-values

**Naive:** Compute the Z-values by *explicitly* comparing characters (left-to-right scan):

\[
S: 11011001 \\
1011001 \\
011001 \\
11001 \\
1001 \\
001 \\
01 \\
1
\]

*What is our time complexity?*
Pattern matching with the Z-value

Given a $Z_i$ value calculator, how do we solve pattern matching?

1. **Pattern, $P$**
2. **Text, $T$**
3. **Calculate Z-values**
4. **Z-algorithm**
5. **Find instances of $P$ in $T$**
Z-value Pattern Matching

To solve pattern matching (given $P$ and $T$), let $S = P$\$T$

$\$ = ‘terminal character’, outside alphabet
Z-value Pattern Matching

To solve pattern matching (given $P$ and $T$), let $S = P$T

$S = 'terminal character', outside alphabet
Z-value Pattern Matching

To solve pattern matching (given $P$ and $T$), let $S = P$\$T$

$\$ = ‘terminal character’, outside alphabet

$P$:  A A  $T$:  A A A A

0 1 2 3 4 5 6

$S$:  A A $\$ A A A A

0 1 2 3

$Z(S) = [-, 1, 0, 2, 2, 2, 1 ]$

What $Z_i$ values are matches?

What are the matching indices in $T$?
Z-value Pattern Matching

$P$: T T
$T$: C T T A

$S$: 

$Z(S)$:

Z-value search pseudo-code

1. Concatenate ($S=P\$T$)

2. Calculate Z-values for $S$

3. For $i<0$, match if $Z_i = \_\_\_\_\_\_$

Match is not at $i$, but instead at ___

__________________________
Assignment 2: a_zval

Learning Objective:

Construct a Z-value calculator and measure its efficiency

Demonstrate use of Z-values in pattern matching

Due: February 7th 11:59 PM

Consider: Our goal is $\theta(|P| + |T|)$. Does Z-value search match this?
End-of-class brainstorm

What information does a single Z-value tell us?

If I know $Z_{i-1}(S)$, can I use that information to help me compute $Z_i(S)$?
The Z-value (Take 2)

Given a string $S$, $Z_i(S)$ is the length of the longest substring in $S$, starting at position $i$, that matches a prefix of $S$.

What information does this give us?

$S$: TTCGTTAGCG $Z_4(S) = 2$
The Z-value (Take 2)

Given a string $S$, $Z_i(S)$ is the length of the longest substring in $S$, starting at position $i$, that matches a prefix of $S$.

What information does this give us?

$S$: TTCGTTAGCG $\quad Z_4 = 2$
The Z-value (Take 2)

Given a string $S$, $Z_i(S)$ is the length of the longest substring in $S$, starting at position $i$, that matches a prefix of $S$.

What information does this give us?

$S$: TTCGTTAGCG

$Z_4 = 2$
The Z-value (Take 2)

Given a string $S$, $Z_i(S)$ is the length of the longest substring in $S$, starting at position $i$, that matches a prefix of $S$.

$Z_i \neq 0$ means that my substring $(i, Z_i)$ matches my prefix $(0, Z_i)$

The characters after my substring and prefix must not match!
Calculating the Z-values (Take 2)

**Intuition:** We can use the previous $Z_1, \ldots, Z_i$ to compute $Z_{i+1}$!

<table>
<thead>
<tr>
<th></th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
</tr>
</thead>
<tbody>
<tr>
<td>$Z_1 = 3$</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>B</td>
<td>A</td>
<td>A</td>
<td>A</td>
</tr>
<tr>
<td></td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>B</td>
<td>A</td>
<td>A</td>
<td>A</td>
</tr>
<tr>
<td>$Z_2 = 2$</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>B</td>
<td>A</td>
<td>A</td>
</tr>
<tr>
<td></td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>B</td>
<td>A</td>
<td>A</td>
<td>A</td>
</tr>
<tr>
<td>$Z_5 = 3$</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>B</td>
<td>A</td>
<td>A</td>
<td>A</td>
</tr>
<tr>
<td></td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>B</td>
<td>A</td>
<td>A</td>
<td>A</td>
</tr>
<tr>
<td>$Z_6 = ?$</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>B</td>
<td>A</td>
<td>A</td>
<td>A</td>
</tr>
<tr>
<td></td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>B</td>
<td>A</td>
<td>A</td>
<td>A</td>
</tr>
</tbody>
</table>
Calculating the Z-values (Take 2)

**Intuition:** We can use the previous $Z_1, \ldots, Z_i$ to compute $Z_{i+1}$!

The **Z-algorithm** (next week) will formalize this process.