Compressed indexation structure for analysing collections of similar genomes

Annie CHATEAU & Alban MANCHERON (LIRMM, équipe MAB),
Gautier SARAH, Gaetan DROC & Manuel RUIZ (CIRAD, UMR AGAP, équipe ID/CIAT)

Clément AGRET

Équipes : ID : Intégration Données
MAB : Méthodes et algorithmes pour la bioinformatique
Data & Computation

Let's simplify, what's happen if we see the GENOME as a simple book?

If =

A lot of books to read!!
If all books tell more or less the same story, what about writing a SuperBook?
Data & Computation

When you read a book you should be able to answer to questions like:

- How many letters are there in the third chapter of the book volume 3 of "Around the World in 80 Days"?
- Is the word "Lustful" appear in the first chapter?
- What is the sentence beginning at 275th paragraph of Chapter 4?

Read the book and answer questions.

Read all books and answer questions! Can take more than a life.

Read the superBook and answer questions!!
Vocabulary

**Alphabet, prefix, factor, suffix**

- Alphabet → $\sum = \{A, C, G, T\}$
- Prefix
- Factor
- Suffix

**Root, node, leaves**

- Root
- Node
- Leaves

**K-mer**

A fragment of k consecutive nucleotides of a word (a sequence from a reference genome as appropriate)

→ A k-mer is a k size factor of a word
Our hypothesis

In rice genome, the number of distinct k-mers (which appear at least once) tends to stabilize from a \( x \) number of genomes.

Adding a new genome to an existing index created on 1000 genomes will be equal to adding a reduct set of positions.

To validate this hypothesis:

- 4 genomes + meta-chart (1) → Venn chart
- 8 genomes + k-mers counter (JellyFish) (2)

1 - https://www.meta-chart.com/
2 - JellyFish: Guillaume Marçais et al
Validation of hypothesis

Specific K-mers

Shared K-mers

Venn diagram of genomes: 9311, IRGSP, dj123 et Kasalath.
Our study

Distinct: Appears at least once
Unique: Appears exactly once
Our approach

Our approach

Our approach

Our approach

<table>
<thead>
<tr>
<th>Prefix Array</th>
<th>Suffix Array</th>
</tr>
</thead>
<tbody>
<tr>
<td>k_1</td>
<td>k_2</td>
</tr>
<tr>
<td>4</td>
<td>4</td>
</tr>
</tbody>
</table>

| g_1          | 0 1 ... 0 0 |
| g_i          | 0 1 ... 1 0 |
| g_n          | 1 0 ... 0 1 |
Examples

Alphabet $\rightarrow \Sigma = \{A, B\}$

$K = 6 \rightarrow K_1 = 2 \& K_2 = 4$

$g_1 = \text{ABBABAABAB}$
$g_2 = \text{ABBBAAABABB}$
$g_3 = \text{AABBBABABABA}$

Let’s Create the superBook!
**Examples**

Alphabet → $\Sigma = \{A, B\}$

$K = 6 \rightarrow K_1 = 2 \& K_2 = 4$

$g_1 = ABBABAABAB$

$g_2 = ABBBAABABB$

$g_3 = AABBBABABA$

Prefix Array

<table>
<thead>
<tr>
<th>AA</th>
<th>AB</th>
<th>BA</th>
<th>BB</th>
</tr>
</thead>
</table>

Suffix Array

| 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 |
| 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 |
| 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 |

$2^2$

$AAA, AAAB, AABA, AABB, ABAA, ABAB, ABBA, BAAA, BABA, BABB, BBAA, BBAB, BBBA, BBBB$
Examples

Alphabet $\rightarrow \Sigma = \{A, B\}$

$K = 6 \rightarrow K_1 = 2$ & $K_2 = 4$

$g_1 = ABBABAABAB$

$g_2 = ABBBBAAABABB$

$g_3 = AABBBABABA$

$g_1 = ABBABAABAB$

$g_1 = ABBBBAAABABB$

$g_3 = AABBBABABA$
Examples

Alphabet → \( \Sigma = \{A, B\} \)

\( K = 6 \rightarrow K_1 = 2 & K_2 = 4 \)

\( g_1 = \text{ABBABAABAB} \)

\( g_2 = \text{ABBBAABABB} \)

\( g_3 = \text{AABBBABABA} \)

\( g_1 = \text{ABBA} \)

\( g_2 = \text{ABBBAB} \)

\( g_3 = \text{ABBABABA} \)
Examples

Alphabet → \( \sum = \{A, B\} \)
K = 6 → \( K_1 = 2 \) & \( K_2 = 4 \)
g1 = ABBABAABAB

g2 = ABBAABABBB

g3 = AABBBABABA

\( g_1 = \) ABBABAABAB
\( g_2 = \) ABBAABABBB
\( g_3 = \) AABBBABABA

### Prefix Array

<table>
<thead>
<tr>
<th>AA</th>
<th>AB</th>
<th>BA</th>
<th>BB</th>
</tr>
</thead>
</table>

### Suffix Array

| 00000000000000000000 | 00000000000000000000 | 00000000000000000000 |
Examples

Alphabet $\rightarrow \Sigma = \{A, B\}$
$K = 6 \rightarrow K_1 = 2 & K_2 = 4$

$g_1 = \text{ABBABAABAB}$
$g_2 = \text{ABBBAABABB}$
$g_3 = \text{AABBBABABA}$

$g_1 = \text{ABBABAABAB}$
$g_1 = \text{ABBABA}$
$g_1 = \text{BABA}$
$g_1 = \text{ABAB}$
$g_1 = \text{ABBA}$
$g_1 = \text{BBB}$
$g_1 = \text{BBB}$
$g_1 = \text{BBB}$
$g_1 = \text{BBB}$
$g_1 = \text{BBB}$
$g_1 = \text{BBB}$
$g_1 = \text{BBB}$
$g_1 = \text{BBB}$

$\text{AAAA, AAAB, AABA, ABBB, ABAA, ABAB, ABBA, ABB, BAAA, BAAB, BAB, BABB, BBAA, BBAB, BBAB, BBBA, BBBB}$
Examples

Alphabet → $\Sigma = \{A, B\}$
$K = 6 \rightarrow K_1 = 2 & K_2 = 4$

$g_1 = ABBABAABAB$
$g_2 = ABBBAABABB$
$g_3 = AABBBABABA$

$g_1 = ABBABAABAB$
$g_2 = ABBBAABABB$
$g_3 = AABBBABABA$

Prefix Array

Suffix Array

AAAA, AAAB, AABA, ABBB, ABAA, ABAB, ABBB, BAAB, BAAB, BAAA, BAB, BAB, BABB, BBAB, BBAB, BBAB, BBAB, BBAB, BBAB
Examples

Alphabet → $\Sigma = \{A, B\}$
$K = 6 \rightarrow K_1 = 2 \& K_2 = 4$

$g_1 = ABBABAABAB$
$g_2 = ABBBAABABB$
$g_3 = AABBBABABA$

Prefix Array

Suffix Array

AAAA, AAAB, AABA, AABB, ABAA, ABAB, ABBA, ABBB, BAAA, BAAB, BABA, BABB, BBAA, BBAB, BBBA, BBBB
Examples

Alphabet $\rightarrow \Sigma = \{A, B\}$
$K = 6 \rightarrow K_1 = 2 \& K_2 = 4$

$g_1 = ABBABAABAB$
$g_2 = ABBBAABABB$
$g_3 = AABBBBABABA$

$g_1 = ABBABAABAB$
$ABBABA$

$g_1 = ABBABAABAB$
$ABBABAB$
Examples

Alphabet $\rightarrow \Sigma = \{A, B\}$

$K = 6 \rightarrow K_1 = 2 \& K_2 = 4$

$g_1 = ABBABAABAB$

$g_2 = ABBBAABABB$

$g_3 = AABBBABABA$

$g_1 = ABBABAABAB$

$g_2 = ABBBAABABB$

$g_3 = AABBBABABA$

$g_1 = ABBABAABAB$

$g_2 = ABBBAABABB$

$g_3 = AABBBABABA$

$g_1 = ABBABAABAB$

$g_2 = ABBBAABABB$

$g_3 = AABBBABABA$

$g_1 = ABBABAABAB$

$g_2 = ABBBAABABB$

$g_3 = AABBBABABA$

Prefix Array

Suffix Array

$AAAA, AAAB, AABA, AABB, ABAA, ABAB, ABBA, AABB, BAAA, BABA, BAB, BABB,BBAA, BBAB, BBBA, BBBB$
Examples

Alphabet → $\Sigma = \{A, B\}$

$K = 6 \rightarrow K_1 = 2 & K_2 = 4$

$g_1 = ABBABAABAB$

$g_2 = ABBBAABABB$

$g_3 = AABBBBABABA$

$g_1 = ABBABAABAB$

$g_1 = ABBABAABAB$

$g_1 = ABBABAABAB$

Prefix Array

<table>
<thead>
<tr>
<th>Prefix</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>
<th>10</th>
<th>11</th>
<th>12</th>
<th>13</th>
<th>14</th>
<th>15</th>
</tr>
</thead>
<tbody>
<tr>
<td>AA</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>7</td>
<td>8</td>
<td>9</td>
<td>10</td>
<td>11</td>
<td>12</td>
<td>13</td>
<td>14</td>
<td>15</td>
</tr>
<tr>
<td>AB</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BA</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BB</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Suffix Array

<table>
<thead>
<tr>
<th>Suffix</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>
<th>10</th>
<th>11</th>
<th>12</th>
<th>13</th>
<th>14</th>
<th>15</th>
</tr>
</thead>
<tbody>
<tr>
<td>AA</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>7</td>
<td>8</td>
<td>9</td>
<td>10</td>
<td>11</td>
<td>12</td>
<td>13</td>
<td>14</td>
<td>15</td>
</tr>
<tr>
<td>AB</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>10</td>
<td>11</td>
<td>12</td>
<td>13</td>
<td>14</td>
<td>15</td>
</tr>
<tr>
<td>BA</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BB</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

$g_1 = 1$

$g_2 = 0$

$g_3 = 0$
Examples

Alphabet → $\sum = \{A, B\}$
K = 6 → $K_1 = 2$ & $K_2 = 4$

$g_1 = ABBABAABAB$
$g_2 = ABBBAABABB$
$g_3 = AABBBABABA$

2^2

<table>
<thead>
<tr>
<th>Prefix Array</th>
<th>Suffix Array</th>
</tr>
</thead>
<tbody>
<tr>
<td>AA</td>
<td>0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0</td>
</tr>
<tr>
<td>AB</td>
<td>0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0</td>
</tr>
<tr>
<td>BA</td>
<td>0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0</td>
</tr>
<tr>
<td>BB</td>
<td>0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0</td>
</tr>
</tbody>
</table>

$g_1 = ABBABAABAB$
$g_2 = ABBBAABABB$
$g_3 = AABBBABABA$

AAAA, AAAB, AABA, ABBB, ABAA, ABAB, ABBB, BAAA, BAAB, BABA, BABB, BBAA, BBAB, BBBB
Examples

Alphabet → \( \sum = \{A, B\} \)

\( K = 6 \rightarrow K_1 = 2 & K_2 = 4 \)

\( g_1 = ABBABAABAB \)
\( g_2 = ABBBAABABB \)
\( g_3 = AABBBBABABA \)

\( g_1 = ABBABAABAB \)

ABBABA
BBABAA
BABAAB
ABAABA
BAABAB

Suffix Array

AAAA, AAAB, AABA, ABBB, ABAA, ABAB, ABBB, BAAA, BAAB, BABA, BABB, BBAA, BBAB, BBBB

Prefix Array
Examples

Alphabet → $\Sigma = \{A, B\}$

$K = 6 \rightarrow K_1 = 2 \& K_2 = 4$

g1 = ABBABAABAB

g2 = ABBBAABABB

g3 = AABBBABABA

2^2

Prefix Array

<table>
<thead>
<tr>
<th>AA</th>
<th>AB</th>
<th>BA</th>
<th>BB</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

Suffix Array

| 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 |
| 0 0 1 0 0 0 0 0 0 1 0 0 0 0 0 |
| 0 0 0 0 1 0 0 0 1 0 0 0 0 0 |
| 0 0 0 1 0 0 0 0 0 0 0 0 0 |

AAA, AAAB, AABA, AABB, ABAA, ABAB, ABBB, BAAA, BAAB, BABA, BABB, BBAA, BBAB, BBBA, BBBB
Examples

Alphabet → $\Sigma = \{A, B\}$

$K = 6 → K_1 = 2 & K_2 = 4$

g1 = ABBABAABAB

g2 = ABBBAABABB

g3 = AABBBABABA

Prefix Array

<table>
<thead>
<tr>
<th>Prefix</th>
<th>00</th>
<th>10</th>
<th>01</th>
</tr>
</thead>
<tbody>
<tr>
<td>AA</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>AB</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>BA</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>BB</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
</tbody>
</table>

$2^2$

Suffix Array

<table>
<thead>
<tr>
<th>Suffix</th>
<th>00000000000010010</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>00100000001011000</td>
</tr>
<tr>
<td></td>
<td>0000010001100000</td>
</tr>
<tr>
<td></td>
<td>0010110001100000</td>
</tr>
</tbody>
</table>

g1 = 00

g2 = 10

g3 = 01

g1 = 100

g2 = 0010

g3 = 0001

g1 = 110

g2 = 100

g3 = 001

g1 = 01000

g2 = 10010

g3 = 00101
Examples

Alphabet → $\Sigma = \{A, B\}$
K = 6 → $K_1 = 2$ & $K_2 = 4$

$g_1 = \text{ABBABAABAB}$
$g_2 = \text{ABBBAAABABB}$
$g_3 = \text{AABBBABABA}$

$\text{Rank}_x(i)$:
Rank return the number of elements $x$ in the range $[0, i]$.

$\text{Select}_x(i)$:
Select is the inverse operation to rank; it answers the question “at which position is the $i^{th}$ set bit?”

Prefix Array

<table>
<thead>
<tr>
<th>AA</th>
<th>AB</th>
<th>BA</th>
<th>BB</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
</tbody>
</table>

Suffix Array

|   | 0  | 1  | 0  | 0  | 0  | 0  | 0  | 0  | 0  | 0  | 0  | 1  | 0  | 0  | 1  | 0  | 1  | 0  |
|---|----|----|----|----|
|   | 0  | 0  | 1  | 0  | 1  | 0  | 0  | 0  | 1  | 0  | 0  | 1  | 1  | 0  | 0  | 0  | 0  | 0  |
|   | 0  | 0  | 0  | 0  | 1  | 0  | 0  | 1  | 1  | 0  | 0  | 0  | 0  | 0  | 0  | 1  | 0  | 1  |
|   | 0  | 1  | 0  | 1  | 1  | 0  | 0  | 0  | 1  | 1  | 0  | 0  | 0  | 0  | 0  | 0  | 1  | 0  | 1  | 0  | 0  | 0  | 0  | 0  |

$g_1 = \text{ABBABAABAB}$
$g_2 = \text{ABBBAAABABB}$
$g_3 = \text{AABBBABABA}$

AAA, AAAB, AABA, AABB, ABAA, ABAB, ABBB, BAAA, BAAB, BABA, BABB, BBAA, BBAB, BBBA, BBBBB
Examples

Is the word BBABAA exist?

BBABAA

Prefix Array

| AA | AB | BA | BB |

Suffix Array

| 0 0 0 0 0 0 0 0 1 0 0 1 0 |
| 0 0 1 0 0 0 0 0 0 1 0 1 1 0 0 |
| 0 0 0 0 0 1 0 0 0 1 1 0 0 0 0 |
| 0 0 1 0 1 1 0 0 0 1 1 0 0 0 0 |

g1 = 00
g2 = 10
g3 = 01
g1 = 1100
g2 = 0010
g3 = 0001
g1 = 110
g2 = 100
g3 = 001
g1 = 01000
g2 = 10010
g3 = 00101
Examples

Is the word BBABAA exist?

BBABAA

Prefix Array

<table>
<thead>
<tr>
<th>AA</th>
<th>AB</th>
<th>BA</th>
<th>BB</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
</tbody>
</table>

Suffix Array

<table>
<thead>
<tr>
<th></th>
<th>0</th>
<th>0</th>
<th>0</th>
<th>0</th>
<th>0</th>
<th>0</th>
<th>0</th>
<th>0</th>
<th>0</th>
<th>0</th>
<th>0</th>
<th>1</th>
<th>0</th>
<th>0</th>
<th>1</th>
<th>0</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

\[ g_1 = 00 \]
\[ g_2 = 10 \]
\[ g_3 = 01 \]
\[ g_1 = 1100 \]
\[ g_2 = 0010 \]
\[ g_3 = 0001 \]
\[ g_1 = 110 \]
\[ g_2 = 100 \]
\[ g_3 = 001 \]
\[ g_1 = 01000 \]
\[ g_2 = 10010 \]
\[ g_3 = 00101 \]
Examples

Is the word BBABAA exist?

BBABAA

AAAA, AAAB, AABA, AABB, ABAA, ABAB, ABBA, ABBB, BAAA, BABA, BABB, BBAA, BBAB, BBBA, BBBB

Prefix Array

<table>
<thead>
<tr>
<th>AA</th>
<th>AB</th>
<th>BA</th>
<th>BB</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Suffix Array

| 0 0 0 0 0 0 0 0 0 1 0 0 1 0 |
| 0 0 1 0 0 0 0 0 0 1 0 1 1 0 0 |
| 0 0 0 0 0 1 0 0 1 1 0 0 0 0 0 |
| 0 0 1 0 1 1 0 0 1 1 0 0 0 0 0 |

\[ g_1 = 00 \]
\[ g_2 = 10 \]
\[ g_3 = 01 \]
\[ g_1 = 1100 \]
\[ g_2 = 0010 \]
\[ g_3 = 0001 \]
\[ g_1 = 110 \]
\[ g_2 = 100 \]
\[ g_3 = 001 \]
\[ g_1 = 01000 \]
\[ g_2 = 10010 \]
\[ g_3 = 00101 \]
Examples

Is the word BBABAA exist?

BBABAA

In which genomes?

Rank_1(ABAA) = 2

AA, AB, ABA, ABB, ABAA, ABAB, ABBB, BAAA, BAAB, BABA, BABB, BBAA, BBAB, BBBA, BBBB

<table>
<thead>
<tr>
<th>Prefix Array</th>
<th>Suffix Array</th>
</tr>
</thead>
<tbody>
<tr>
<td>AA</td>
<td>0 0 0 0 0 0 0 0 0 0 1 0 0 1 0</td>
</tr>
<tr>
<td>AB</td>
<td>0 0 1 0 0 0 0 0 0 1 0 1 1 0 0</td>
</tr>
<tr>
<td>BA</td>
<td>0 0 0 0 1 0 0 0 1 1 0 0 0 0 0</td>
</tr>
<tr>
<td>BB</td>
<td>0 0 1 0 1 1 0 0 0 1 1 0 0 0 0</td>
</tr>
</tbody>
</table>

g1 = 00

g2 = 10

g3 = 01

g1 = 1100

g2 = 0010

g3 = 0001

g1 = 110

g2 = 100

g3 = 001

g1 = 01000

g2 = 10010

g3 = 00101

AA

AB

BA

BB
Examples

Is the word BBABAA exist?

BBABAA

In which genomes?
Rank\_1(ABAA) = 2

Prefix Array

<table>
<thead>
<tr>
<th>Prefix</th>
<th>Rank</th>
</tr>
</thead>
<tbody>
<tr>
<td>AA</td>
<td>1</td>
</tr>
<tr>
<td>AB</td>
<td>2</td>
</tr>
<tr>
<td>BA</td>
<td>3</td>
</tr>
<tr>
<td>BB</td>
<td>4</td>
</tr>
</tbody>
</table>

Suffix Array

<table>
<thead>
<tr>
<th>Suffix</th>
<th>Rank</th>
</tr>
</thead>
<tbody>
<tr>
<td>AA</td>
<td>1</td>
</tr>
<tr>
<td>AB</td>
<td>2</td>
</tr>
<tr>
<td>BA</td>
<td>3</td>
</tr>
<tr>
<td>BB</td>
<td>4</td>
</tr>
</tbody>
</table>

2^2
Examples

Is the word BBABAA exist?

BBABAA

In which genomes?
Rank₁(ABAA) = 2
→ g₁ only!

Reminder:
g₁ = ABBABABAABAB
g₂ = ABBBBABABB
g₃ = ABBBBABABA
RRR was first proposed by Raman et al [1]

→ \(O(1)\) time binary rank queries

→ \(N H_0(S) + o(N)\)  \(H_0(S)\) is the zeroth-order empirical entropy of \(S\)

B : Size of the blocks = Numbers of 1 and 0 in the block
F : Superblock factor
C : Class number = Numbers of 1 in the block b
O : Offset = Index into the table

B : Size of the blocks = Numbers of 1 and 0 in the block
F : Superblock factor
C : Class number = Numbers of 1 in the block b
O : Offset = Index into the table

Sum of ranks for all previous blocks

Initial offset addresses
RUBIKS : RRR Update for Bit Indexing in K-mer Structure

0000 0100 0000 1000 0000 0100 0000

- Prefixes: $4^8$
- Kmers: $10^9$
  → 15258 "1"
- Suffixes: $4^{20}$
RUBIKS : RRR Update for Bit Indexing in K-mer Structure

```
0000 0100 0000 1000 0000 0100 0000
NULL  NULL  NULL  NULL  NULL  NULL
```

- Prefixes : $4^8$
- Kmers : $10^9$
  → 15258 “1”
- Suffixes: $4^{20}$

Pointeur vers une RRR
RUBIKS : RRR Update for Bit Indexing in K-mer Structure

- Prefixes: $4^8$
- Kmers: $10^9$
  $\rightarrow$ 15258 “1”
- Suffixes: $4^{20}$

Pointeur vers une RRR
RUBIKS : RRR Update for Bit Indexing in K-mer Structure

- Prefixes: $4^8$
- Kmers: $10^9$
  - $\rightarrow$ 15258 “1”
- Suffixes: $4^{20}$

We still can do a Rank and Select!
Our outlook

How to answer the following questions:

Is the word "ATATAAGATTACA" present in the first chromosome of genomes 644?

Which sequences are common to the kasalath and 9311 genomes?

Index of 3000 genomes

→ Tools based on this index
→ Integrate these tools into the GenomeHarvest project
→ Create tools to make these index structure easy to use
Thank you for your attention.
Do you have questions?
That's all Folks!
Annexes

If searching for a word takes 1 sec per line:
→ Search for a word in a book of 500 000 000 lines:
500,000,000 sec: 5,787 Days

→ Search for a word in a dictionary of 500 000 000 lines:
Log (500,000,000) sec: 28,897 sec
## Results

(1) **SDSL Lite**

<table>
<thead>
<tr>
<th>$K_2$</th>
<th>compression</th>
<th>$b$</th>
<th>$\text{rrrb}$ $\log \frac{n}{m}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>14</td>
<td>32 Mo</td>
<td>2 Mo</td>
<td></td>
</tr>
<tr>
<td>15</td>
<td>128 Mo</td>
<td>4 Mo</td>
<td></td>
</tr>
<tr>
<td>16</td>
<td>512 Mo</td>
<td>32 Mo</td>
<td></td>
</tr>
<tr>
<td>17</td>
<td>2 Go</td>
<td>128 Mo</td>
<td></td>
</tr>
</tbody>
</table>
Validation of hypothesis

Venn diagram of genomes: 9311, musa_v1, musa_v2

Specific K-mers

Shared K-mers

54 030 473
51 619 980

259 616 512

312 007 403

207 353
22 854
21 109
#ifdef __aquoiicassert
Prefix Array

Suffix 1: #G  Suffix 2: #G  ...  Suffix i: #G  Suffix s: #G

Suffix Array

\( k_1 \)

\( 4 \)

\( g_1 \) 0 1 ... 0 0

\( g_i \) 0 1 ... 1 0

\( g_n \) 1 0 ... 0 1