Tailoring *r*-index for document listing towards metagenomics applications

Dustin Cobas[1] Veli Mäkinen[2] Massimiliano Rossi[3]

SPIRE 2020

[1]CeBiB — Center for Biotechnology and Bioengineering, Department of Computer Science, University of Chile, Santiago, Chile

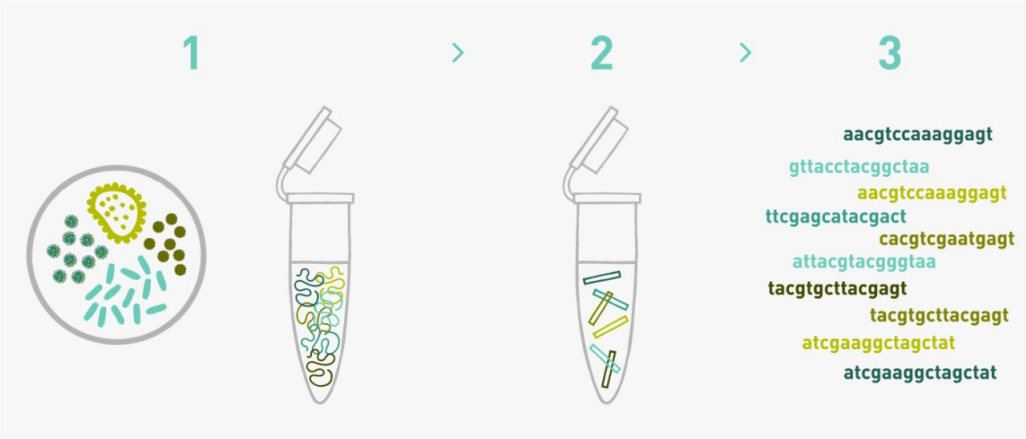
[2]Department of Computer Science, University of Helsinki, Helsinki, Finland

[3]Department of Computer and Information Science and Engineering, University of Florida, Gainesville, USA

Outline

Introduction

Metagenomics



Source: https://www.allgenetics.eu

Pseudoaligners: Kallisto, MetaKallisto, Themisto, ...

Pseudoalignment Criterion

Given a set of references $T_1, \ldots, T_t$ (representing t distinct species) and read P: the read P is pseudoaligned with T_i if there exists a k-mer of P that occurs in T_i and for all other k-mers u of P, either u occurs in T_i or u does not occur in $T_1, \ldots, T_t$.

Document Listing with Frequencies

Problem

Given $D = \{T_1, T_2, \ldots, T_t\}$, and a pattern P, return the set of documents where P occurs and their frequencies.

(Välimäki and Mäkinen, 2007)

Basics

Suffix	SA	DA
…	…	…
AAAACGACAAACCGT…	47737902	0
AAAACGACAAACCGT…	36966161	0
AAAACGACAAACGTA…	77296147	1
AAAACGACAAACGTA…	59705026	1
AAAACGACAAACGTA…	53683742	1
AAAACGACAAAGAAA…	3332196	0
AAAACGACAAAGATA…	67002068	1
AAAACGACAAAGCGA…	106042043	2
AAAACGACAAAGCGA…	97519792	2
AAAACGACAAAGCGA…	92078024	2
AAAACGACAAAGCGA…	94675771	2
AAAACGACAAAGGAT…	12625901	0
AAAACGACAAAGGAT…	7548512	0
AAAACGACAAAGGAT…	34127797	0
AAAACGACAAAGGAT…	25240293	0
AAAACGACAAAGGAT…	17783154	0
AAAACGACAAAGGAT…	44514277	0
AAAACGACAAAGGAT…	50426443	0
AAAACGACAAAGTAG…	93233237	2
AAAACGACAAAGTAG…	90395352	2
AAAACGACAAAGTTT…	94295208	2
AAAACGACAAAGTTT…	99741549	2
AAAACGACAAATAAA…	94013490	2
AAAACGACAAATAAA…	88500889	2
AAAACGACAAATAAA…	91252230	2
AAAACGACAAATAAA…	108106766	2
AAAACGACAAATAAA…	102593450	2
…	…	…

- $\mathcal{D}[1..n] = T_1 T_2 \cdots T_t$
- Suffixes in lexicographical order
- Suffix Array (SA)
- Document Array (DA)
- $\mathrm{tf}(\mathcal{D}, P) \rightarrow$ counting document occurrences in $\mathrm{DA}[s_p..e_p]$

Basics

Suffix	SA	DA
…	…	…
AAAACGACAAACCGT…	47737902	0
AAAACGACAAACCGT…	36966161	0
AAAACGACAAACGTA…	77296147	1
AAAACGACAAACGTA…	59705026	1
AAAACGACAAACGTA…	53683742	1
AAAACGACAAAGAAA…	3332196	0
AAAACGACAAAGATA…	67002068	1
AAAACGACAAAGCGA…	106042043	2
AAAACGACAAAGCGA…	97519792	2
AAAACGACAAAGCGA…	92078024	2
AAAACGACAAAGCGA…	94675771	2
AAAACGACAAAGGAT…	12625901	0
AAAACGACAAAGGAT…	7548512	0
AAAACGACAAAGGAT…	34127797	0
AAAACGACAAAGGAT…	25240293	0
AAAACGACAAAGGAT…	17783154	0
AAAACGACAAAGGAT…	44514277	0
AAAACGACAAAGGAT…	50426443	0
AAAACGACAAAGTAG…	93233237	2
AAAACGACAAAGTAG…	90395352	2
AAAACGACAAAGTTT…	94295208	2
AAAACGACAAAGTTT…	99741549	2
AAAACGACAAATAAA…	94013490	2
AAAACGACAAATAAA…	88500889	2
AAAACGACAAATAAA…	91252230	2
AAAACGACAAATAAA…	108106766	2
AAAACGACAAATAAA…	102593450	2
…	…	…

- $\mathcal{D}[1..n] = T_1 T_2 \cdots T_t$
- Suffixes in lexicographical order
- Suffix Array (SA)
- Document Array (DA)
- $\mathrm{tf}(\mathcal{D}, P) \rightarrow$ counting document occurrences in $\mathrm{DA}[s_p..e_p]$

Basics

Suffix	SA	DA
…	…	…
AAAACGACAAACCGT…	47737902	0
AAAACGACAAACCGT…	36966161	0
AAAACGACAAACGTA…	77296147	1
AAAACGACAAACGTA…	59705026	1
AAAACGACAAACGTA…	53683742	1
AAAACGACAAAGAAA…	3332196	0
AAAACGACAAAGATA…	67002068	1
AAAACGACAAAGCGA…	106042043	2
AAAACGACAAAGCGA…	97519792	2
AAAACGACAAAGCGA…	92078024	2
AAAACGACAAAGCGA…	94675771	2
AAAACGACAAAGGAT…	12625901	0
AAAACGACAAAGGAT…	7548512	0
AAAACGACAAAGGAT…	34127797	0
AAAACGACAAAGGAT…	25240293	0
AAAACGACAAAGGAT…	17783154	0
AAAACGACAAAGGAT…	44514277	0
AAAACGACAAAGGAT…	50426443	0
AAAACGACAAAGTAG…	93233237	2
AAAACGACAAAGTAG…	90395352	2
AAAACGACAAAGTTT…	94295208	2
AAAACGACAAAGTTT…	99741549	2
AAAACGACAAATAAA…	94013490	2
AAAACGACAAATAAA…	88500889	2
AAAACGACAAATAAA…	91252230	2
AAAACGACAAATAAA…	108106766	2
AAAACGACAAATAAA…	102593450	2
…	…	…

- $\mathcal{D}[1..n] = T_1 T_2 \cdots T_t$
- Suffixes in lexicographical order
- Suffix Array (SA)
- Document Array (DA)
- $\mathrm{tf}(\mathcal{D}, P) \rightarrow$ counting document occurrences in $\mathrm{DA}[s_p..e_p]$

Basics

Suffix	SA	DA
…	…	…
AAAACGACAAACCGT…	47737902	0
AAAACGACAAACCGT…	36966161	0
AAAACGACAAACGTA…	77296147	1
AAAACGACAAACGTA…	59705026	1
AAAACGACAAACGTA…	53683742	1
AAAACGACAAAGAAA…	3332196	0
AAAACGACAAAGATA…	67002068	1
AAAACGACAAAGCGA…	106042043	2
AAAACGACAAAGCGA…	97519792	2
AAAACGACAAAGCGA…	92078024	2
AAAACGACAAAGCGA…	94675771	2
AAAACGACAAAGGAT…	12625901	0
AAAACGACAAAGGAT…	7548512	0
AAAACGACAAAGGAT…	34127797	0
AAAACGACAAAGGAT…	25240293	0
AAAACGACAAAGGAT…	17783154	0
AAAACGACAAAGGAT…	44514277	0
AAAACGACAAAGGAT…	50426443	0
AAAACGACAAAGTAG…	93233237	2
AAAACGACAAAGTAG…	90395352	2
AAAACGACAAAGTTT…	94295208	2
AAAACGACAAAGTTT…	99741549	2
AAAACGACAAATAAA…	94013490	2
AAAACGACAAATAAA…	88500889	2
AAAACGACAAATAAA…	91252230	2
AAAACGACAAATAAA…	108106766	2
AAAACGACAAATAAA…	102593450	2
…	…	…

- $\mathcal{D}[1..n] = T_1 T_2 \cdots T_t$
- Suffixes in lexicographical order
- Suffix Array (SA)
- Document Array (DA)
- $\mathrm{tf}(\mathcal{D}, P) \rightarrow$ counting document occurrences in $\mathrm{DA}[s_p..e_p]$

Basics

Suffix	SA	DA
…	…	…
AAAACGACAAACCGT…	47737902	0
AAAACGACAAACCGT…	36966161	0
AAAACGACAAACGTA…	77296147	1
AAAACGACAAACGTA…	59705026	1
AAAACGACAAACGTA…	53683742	1
AAAACGACAAAGAAA…	3332196	0
AAAACGACAAAGATA…	67002068	1
AAAACGACAAAGCGA…	106042043	2
AAAACGACAAAGCGA…	97519792	2
AAAACGACAAAGCGA…	92078024	2
AAAACGACAAAGCGA…	94675771	2
AAAACGACAAAGGAT…	12625901	0
AAAACGACAAAGGAT…	7548512	0
AAAACGACAAAGGAT…	34127797	0
AAAACGACAAAGGAT…	25240293	0
AAAACGACAAAGGAT…	17783154	0
AAAACGACAAAGGAT…	44514277	0
AAAACGACAAAGGAT…	50426443	0
AAAACGACAAAGTAG…	93233237	2
AAAACGACAAAGTAG…	90395352	2
AAAACGACAAAGTTT…	94295208	2
AAAACGACAAAGTTT…	99741549	2
AAAACGACAAATAAA…	94013490	2
AAAACGACAAATAAA…	88500889	2
AAAACGACAAATAAA…	91252230	2
AAAACGACAAATAAA…	108106766	2
AAAACGACAAATAAA…	102593450	2
…	…	…

- $\mathcal{D}[1..n] = T_1 T_2 \cdots T_t$
- Suffixes in lexicographical order
- Suffix Array (SA)
- Document Array (DA)
- $\mathsf{tf}(\mathcal{D}, P) \rightarrow$ counting document occurrences in DA$[s_p..e_p]$

Term Frequency with r—index

Term Frequency with $r-$index

r-index	DA
…	…
47737902	0
36966161	0
77296147	1
59705026	1
53683742	1
3332196	0
67002068	1
106042043	2
97519792	2
92078024	2
94675771	2
12625901	0
7548512	0
34127797	0
25240293	0
17783154	0
44514277	0
50426443	0
93233237	2
90395352	2
94295208	2
99741549	2
94013490	2
88500889	2
91252230	2
108106766	2
102593450	2
…	…

- $[s_p..e_p]$ with $r-$index
- $T \in \mathrm{DA}[s_p..e_p]$
- $r-\mathrm{index}_d, d \in [1..t]$
- *Leftmost* and *rightmost* occurrences
- Only corresponding positions on SA_d are required!

Access SA^{-1} with $r-$index

The access operation $\mathrm{SA}_d^{-1}[i]$ takes $\mathcal{O}(\log(n_d/r_d))$ time using $\mathcal{O}(r_d \log(n_d/r_d)$ space.

(Sadakane, 2007)

Term Frequency with $r-$index

r-index	DA
…	…
47737902	0
36966161	0
77296147	1
59705026	1
53683742	1
3332196	0
67002068	1
106042043	2
97519792	2
92078024	2
94675771	2
12625901	0
7548512	0
34127797	0
25240293	0
17783154	0
44514277	0
50426443	0
93233237	2
90395352	2
94295208	2
99741549	2
94013490	2
88500889	2
91252230	2
108106766	2
102593450	2
…	…

- $[s_p..e_p]$ with $r-$index
- $T \in \mathrm{DA}[s_p..e_p]$
- $r-\mathrm{index}_d, d \in [1..t]$
- *Leftmost* and *rightmost* occurrences
- Only corresponding positions on SA_d are required!

Access SA^{-1} with $r-$index

The access operation $\mathrm{SA}_d^{-1}[i]$ takes $\mathcal{O}(\log(n_d/r_d))$ time using $\mathcal{O}(r_d \log(n_d/r_d)$ space.

(Sadakane, 2007)

Term Frequency with $r-$index

r-index	DA
…	…
47737902	0
36966161	0
77296147	1
59705026	1
53683742	1
3332196	0
67002068	1
106042043	2
97519792	2
92078024	2
94675771	2
12625901	0
7548512	0
34127797	0
25240293	0
17783154	0
44514277	0
50426443	0
93233237	2
90395352	2
94295208	2
99741549	2
94013490	2
88500889	2
91252230	2
108106766	2
102593450	2
…	…

#	r-index$_3$
…	…
185271	36966161
185272	47737902
185273	31828339
185274	3332196
185275	12625901
185276	7548512
185277	34127797
185278	25240293
185279	17783154
185280	44514277
185281	50426443
185282	50068803
185283	22613683
185284	1250302
…	…

- $[s_p..e_p]$ with $r-$index
- $T \in \mathrm{DA}[s_p..e_p]$
- $r-\mathrm{index}_d, d \in [1..t]$
- *Leftmost* and *rightmost* occurrences
- Only corresponding positions on SA_d are required!

Access SA^{-1} with $r-$index

The access operation $\mathrm{SA}_d^{-1}[i]$ takes $\mathcal{O}(\log(n_d/r_d))$ time using $\mathcal{O}(r_d \log(n_d/r_d)$ space.

(Sadakane, 2007)

Term Frequency with $r-$index

r-index	DA
…	…
47737902	0
36966161	0
77296147	1
59705026	1
53683742	1
3332196	0
67002068	1
106042043	2
97519792	2
92078024	2
94675771	2
12625901	0
7548512	0
34127797	0
25240293	0
17783154	0
44514277	0
50426443	0
93233237	2
90395352	2
94295208	2
99741549	2
94013490	2
88500889	2
91252230	2
108106766	2
102593450	2
…	…

#	r-index$_3$
…	…
185271	36966161
185272	47737902
185273	31828339
185274	3332196
185275	12625901
185276	7548512
185277	34127797
185278	25240293
185279	17783154
185280	44514277
185281	50426443
185282	50068803
185283	22613683
185284	1250302
…	…

- $[s_p..e_p]$ with $r-$index
- $T \in \mathrm{DA}[s_p..e_p]$
- $r-\mathrm{index}_d, d \in [1..t]$
- *Leftmost* and *rightmost* occurrences
- Only corresponding positions on SA_d are required!

Access SA^{-1} with $r-$index

The access operation $\mathrm{SA}_d^{-1}[i]$ takes $\mathcal{O}(\log(n_d/r_d))$ time using $\mathcal{O}(r_d \log(n_d/r_d)$ space.

(Sadakane, 2007)

Term Frequency with $r-$index

r-index	DA
…	…
47737902	0
36966161	0
77296147	1
59705026	1
53683742	1
3332196	0
67002068	1
106042043	2
97519792	2
92078024	2
94675771	2
12625901	0
7548512	0
34127797	0
25240293	0
17783154	0
44514277	0
50426443	0
93233237	2
90395352	2
94295208	2
99741549	2
94013490	2
88500889	2
91252230	2
108106766	2
102593450	2
…	…

#	r-index$_3$
…	…
185271	36966161
185272	47737902
185273	31828339
185274	3332196
185275	12625901
185276	7548512
185277	34127797
185278	25240293
185279	17783154
185280	44514277
185281	50426443
185282	50068803
185283	22613683
185284	1250302
…	…

- $[s_p..e_p]$ with $r-$index
- $T \in \mathrm{DA}[s_p..e_p]$
- $r-\mathrm{index}_d, d \in [1..t]$
- *Leftmost* and *rightmost* occurrences
- Only corresponding positions on SA_d are required!

Access SA^{-1} with $r-$index

The access operation $\mathrm{SA}_d^{-1}[i]$ takes $\mathcal{O}(\log(n_d/r_d))$ time using $\mathcal{O}(r_d \log(n_d/r_d)$ space.

(Sadakane, 2007)

Term Frequency with $r-$index

r-index	DA
…	…
47737902	0
36966161	0
77296147	1
59705026	1
53683742	1
3332196	0
67002068	1
106042043	2
97519792	2
92078024	2
94675771	2
12625901	0
7548512	0
34127797	0
25240293	0
17783154	0
44514277	0
50426443	0
93233237	2
90395352	2
94295208	2
99741549	2
94013490	2
88500889	2
91252230	2
108106766	2
102593450	2
…	…

#	r-index$_3$
…	…
185271	36966161
185272	47737902
185273	31828339
185274	3332196
185275	12625901
185276	7548512
185277	34127797
185278	25240293
185279	17783154
185280	44514277
185281	50426443
185282	50068803
185283	22613683
185284	1250302
…	…

- $[s_p..e_p]$ with $r-$index
- $T \in \mathrm{DA}[s_p..e_p]$
- $r-\mathrm{index}_d, d \in [1..t]$
- *Leftmost* and *rightmost* occurrences
- Only corresponding positions on SA_d are required!

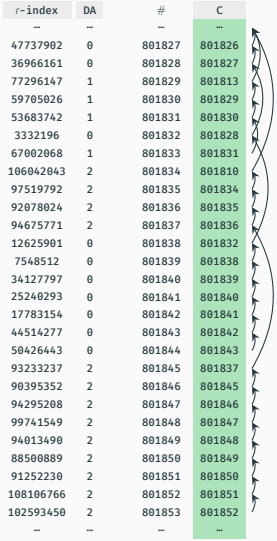
Access SA^{-1} with $r-$index

The access operation $\mathrm{SA}_d^{-1}[i]$ takes $\mathcal{O}(\log(n_d/r_d))$ time using $\mathcal{O}(r_d \log(n_d/r_d)$ space.

(Sadakane, 2007)

Solutions

Sadakane



r-index	DA	#	C
…	…	…	…
47737902	0	801827	801826
36966161	0	801828	801827
77296147	1	801829	801813
59705026	1	801830	801829
53683742	1	801831	801830
3332196	0	801832	801828
67002068	1	801833	801831
106042043	2	801834	801810
97519792	2	801835	801834
92078024	2	801836	801835
94675771	2	801837	801836
12625901	0	801838	801832
7548512	0	801839	801838
34127797	0	801840	801839
25240293	0	801841	801840
17783154	0	801842	801841
44514277	0	801843	801842
50426443	0	801844	801843
93233237	2	801845	801837
90395352	2	801846	801845
94295208	2	801847	801846
99741549	2	801848	801847
94013490	2	801849	801848
88500889	2	801850	801849
91252230	2	801851	801850
108106766	2	801852	801851
102593450	2	801853	801852
…	…	…	…

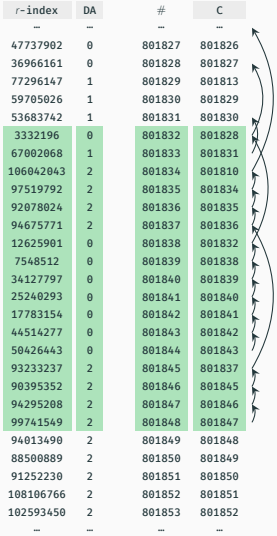
- Links to predecessor occurrence of the document (C)
- Leftmost occurrences in $[s_p..e_p]$
- Range Minimum Query on C (RMQ_C)

Document Listing with Frequencies

The total time is $\mathcal{O}(m + ndoc \cdot \log(n/r))$.

(Sadakane, 2007)

Sadakane



r-index	DA	#	C
…	…	…	…
47737902	0	801827	801826
36966161	0	801828	801827
77296147	1	801829	801813
59705026	1	801830	801829
53683742	1	801831	801830
3332196	0	801832	801828
67002068	1	801833	801831
106042043	2	801834	801810
97519792	2	801835	801834
92078024	2	801836	801835
94675771	2	801837	801836
12625901	0	801838	801832
7548512	0	801839	801838
34127797	0	801840	801839
25240293	0	801841	801840
17783154	0	801842	801841
44514277	0	801843	801842
50426443	0	801844	801843
93233237	2	801845	801837
90395352	2	801846	801845
94295208	2	801847	801846
99741549	2	801848	801847
94013490	2	801849	801848
88500889	2	801850	801849
91252230	2	801851	801850
108106766	2	801852	801851
102593450	2	801853	801852
…	…	…	…

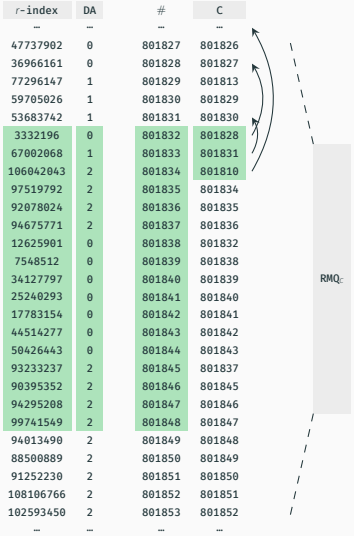
- Links to predecessor occurrence of the document (C)
- Leftmost occurrences in $[s_p..e_p]$
- Range Minimum Query on C (RMQ_C)

Document Listing with Frequencies

The total time is $\mathcal{O}(m + ndoc \cdot \log(n/r))$.

(Sadakane, 2007)

Sadakane



r-index	DA	#	C
…	…	…	…
47737902	0	801827	801826
36966161	0	801828	801827
77296147	1	801829	801813
59705026	1	801830	801829
53683742	1	801831	801830
3332196	0	801832	801828
67002068	1	801833	801831
106042043	2	801834	801810
97519792	2	801835	801834
92078024	2	801836	801835
94675771	2	801837	801836
12625901	0	801838	801832
7548512	0	801839	801838
34127797	0	801840	801839
25240293	0	801841	801840
17783154	0	801842	801841
44514277	0	801843	801842
50426443	0	801844	801843
93233237	2	801845	801837
90395352	2	801846	801845
94295208	2	801847	801846
99741549	2	801848	801847
94013490	2	801849	801848
88500889	2	801850	801849
91252230	2	801851	801850
108106766	2	801852	801851
102593450	2	801853	801852
…	…	…	…

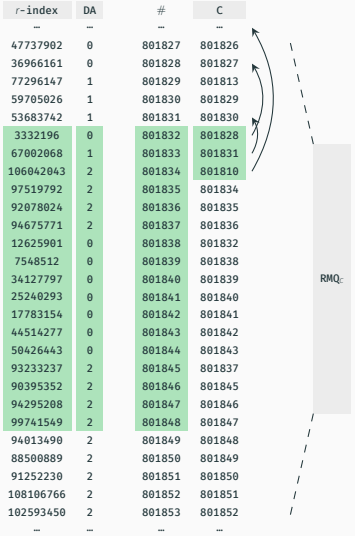
- Links to predecessor occurrence of the document (C)
- Leftmost occurrences in $[s_p..e_p]$
- Range Minimum Query on C (RMQ_C)

Document Listing with Frequencies

The total time is $\mathcal{O}(m + ndoc \cdot \log(n/r))$.

(Sadakane, 2007)

Sadakane



r-index	DA	#	C
…	…	…	…
47737902	0	801827	801826
36966161	0	801828	801827
77296147	1	801829	801813
59705026	1	801830	801829
53683742	1	801831	801830
3332196	0	801832	801828
67002068	1	801833	801831
106042043	2	801834	801810
97519792	2	801835	801834
92078024	2	801836	801835
94675771	2	801837	801836
12625901	0	801838	801832
7548512	0	801839	801838
34127797	0	801840	801839
25240293	0	801841	801840
17783154	0	801842	801841
44514277	0	801843	801842
50426443	0	801844	801843
93233237	2	801845	801837
90395352	2	801846	801845
94295208	2	801847	801846
99741549	2	801848	801847
94013490	2	801849	801848
88500889	2	801850	801849
91252230	2	801851	801850
108106766	2	801852	801851
102593450	2	801853	801852
…	…	…	…

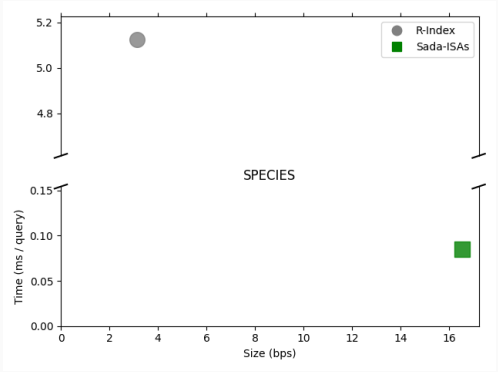
- Links to predecessor occurrence of the document (C)
- Leftmost occurrences in $[s_p..e_p]$
- Range Minimum Query on C (RMQ_C)

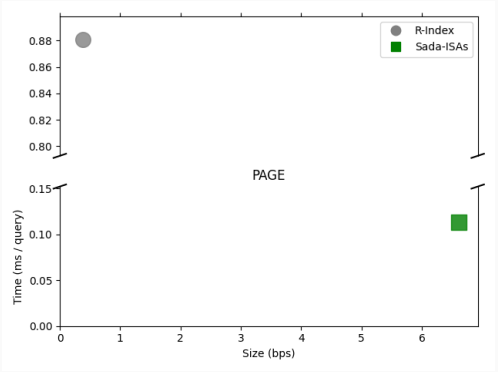
Document Listing with Frequencies

The total time is $\mathcal{O}(m + ndoc \cdot \log(n/r))$.

(Sadakane, 2007)

Sadakane: Results





Interleaved Longest Common Prefix

r-index	DA	#	ILCP
…	…	…	…
47737902	0	801827	15117
36966161	0	801828	14861
77296147	1	801829	12
59705026	1	801830	30
53683742	1	801831	202
3332196	0	801832	11
67002068	1	801833	11
106042043	2	801834	11
97519792	2	801835	153070
92078024	2	801836	47893
94675771	2	801837	152969
12625901	0	801838	12
7548512	0	801839	8697
34127797	0	801840	8735
25240293	0	801841	8734
17783154	0	801842	8734
44514277	0	801843	8735
50426443	0	801844	8736
93233237	2	801845	12
90395352	2	801846	250822
94295208	2	801847	13
99741549	2	801848	105863
94013490	2	801849	11
88500889	2	801850	225776
91252230	2	801851	56148
108106766	2	801852	3969
102593450	2	801853	56148
…	…	…	…

- Interleaved Longest Common Prefix (ILCP)
- Leftmost occurrences in $[s_p..e_p]$
- Equal-value runs
- Range Minimum Query on run heads of ILCP (RMQ_{ILCP})

Document Listing with Frequencies

The total time is $\mathcal{O}(m + ndoc \cdot \log(n/r))$.

(Gagie, Hartikainen, Karhu, Kärkkäinen, Navarro, Puglisi and Sirén, 2017)

Interleaved Longest Common Prefix

r-index	DA	#	ILCP
…	…	…	…
47737902	0	801827	15117
36966161	0	801828	14861
77296147	1	801829	12
59705026	1	801830	30
53683742	1	801831	202
3332196	0	801832	11
67002068	1	801833	11
106042043	2	801834	11
97519792	2	801835	153070
92078024	2	801836	47893
94675771	2	801837	152969
12625901	0	801838	12
7548512	0	801839	8697
34127797	0	801840	8735
25240293	0	801841	8734
17783154	0	801842	8734
44514277	0	801843	8735
50426443	0	801844	8736
93233237	2	801845	12
90395352	2	801846	250822
94295208	2	801847	13
99741549	2	801848	105863
94013490	2	801849	11
88500889	2	801850	225776
91252230	2	801851	56148
108106766	2	801852	3969
102593450	2	801853	56148
…	…	…	…

- Interleaved Longest Common Prefix (ILCP)
- Leftmost occurrences in $[s_p..e_p]$
- Equal-value runs
- Range Minimum Query on run heads of ILCP (RMQ_{ILCP})

Document Listing with Frequencies

The total time is $\mathcal{O}(m + ndoc \cdot \log(n/r))$.

(Gagie, Hartikainen, Karhu, Kärkkäinen, Navarro, Puglisi and Sirén, 2017)

Interleaved Longest Common Prefix

r-index	DA	#	ILCP
…	…	…	…
47737902	0	801827	15117
36966161	0	801828	14861
77296147	1	801829	12
59705026	1	801830	30
53683742	1	801831	202
3332196	0	801832	11
67002068	1	801833	11
106042043	2	801834	11
97519792	2	801835	153070
92078024	2	801836	47893
94675771	2	801837	152969
12625901	0	801838	12
7548512	0	801839	8697
34127797	0	801840	8735
25240293	0	801841	8734
17783154	0	801842	8734
44514277	0	801843	8735
50426443	0	801844	8736
93233237	2	801845	12
90395352	2	801846	250822
94295208	2	801847	13
99741549	2	801848	105863
94013490	2	801849	11
88500889	2	801850	225776
91252230	2	801851	56148
108106766	2	801852	3969
102593450	2	801853	56148
…	…	…	…

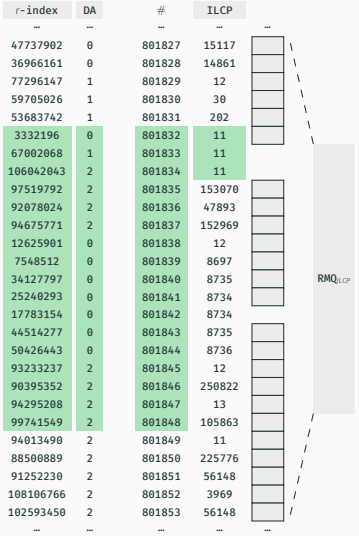
- Interleaved Longest Common Prefix (ILCP)
- Leftmost occurrences in $[s_p..e_p]$
- Equal-value runs
- Range Minimum Query on run heads of ILCP (RMQ_{ILCP})

Document Listing with Frequencies

The total time is $\mathcal{O}(m + ndoc \cdot \log(n/r))$.

(Gagie, Hartikainen, Karhu, Kärkkäinen, Navarro, Puglisi and Sirén, 2017)

Interleaved Longest Common Prefix



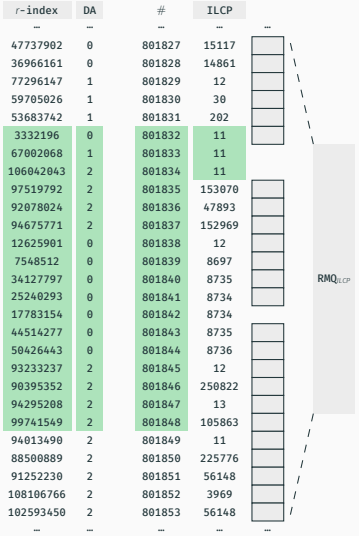
- Interleaved Longest Common Prefix (ILCP)
- Leftmost occurrences in $[s_p..e_p]$
- Equal-value runs
- Range Minimum Query on run heads of ILCP ($\mathrm{RMQ}_{\mathrm{ILCP}}$)

Document Listing with Frequencies

The total time is $\mathcal{O}(m + ndoc \cdot \log(n/r))$.

(Gagie, Hartikainen, Karhu, Kärkkäinen, Navarro, Puglisi and Sirén, 2017)

Interleaved Longest Common Prefix



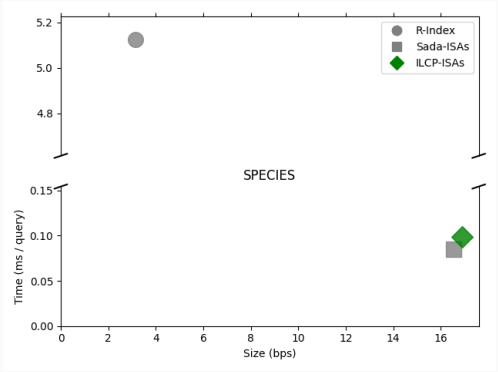
- Interleaved Longest Common Prefix (ILCP)
- Leftmost occurrences in $[s_p..e_p]$
- Equal-value runs
- Range Minimum Query on run heads of ILCP ($\mathrm{RMQ}_{\mathrm{ILCP}}$)

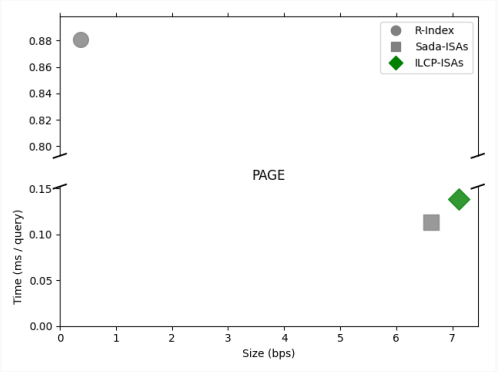
Document Listing with Frequencies

The total time is $\mathcal{O}(m + ndoc \cdot \log(n/r))$.

(Gagie, Hartikainen, Karhu, Kärkkäinen, Navarro, Puglisi and Sirén, 2017)

Interleaved Longest Common Prefix: Results





Interleaved Longest Common Prefix with Double Encoding

r-index	DA	#	ILCP	
…	…	…	…	…
47737902	0	801827	15117	
36966161	0	801828	14861	
77296147	1	801829	12	
59705026	1	801830	30	
53683742	1	801831	202	
3332196	0	801832	11	
67002068	1	801833	11	
106042043	2	801834	11	
97519792	2	801835	153070	
92078024	2	801836	47893	
94675771	2	801837	152969	
12625901	0	801838	12	
7548512	0	801839	8697	
34127797	0	801840	8735	
25240293	0	801841	8734	
17783154	0	801842	8734	
44514277	0	801843	8735	
50426443	0	801844	8736	
93233237	2	801845	12	
90395352	2	801846	250822	
94295208	2	801847	13	
99741549	2	801848	105863	
94013490	2	801849	11	
88500889	2	801850	225776	
91252230	2	801851	56148	
108106766	2	801852	3969	
102593450	2	801853	56148	
…	…	…	…	…

- Equal-value runs
- Equal-document runs
- Range Minimum Query on run heads of doubled-encoded ILCP ($\mathrm{RMQ}_{\mathrm{ILCP}^\star}$)

Document Listing with Frequencies

The total time is $\mathcal{O}(m + ndoc \cdot \log(n/r))$.

Interleaved Longest Common Prefix with Double Encoding

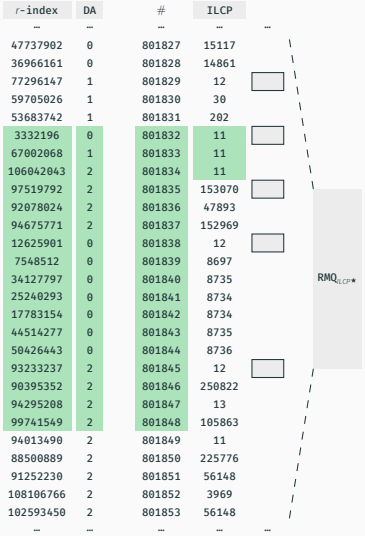
r-index	DA	#	ILCP	
…	…	…	…	…
47737902	0	801827	15117	
36966161	0	801828	14861	
77296147	1	801829	12	
59705026	1	801830	30	
53683742	1	801831	202	
3332196	0	801832	11	
67002068	1	801833	11	
106042043	2	801834	11	
97519792	2	801835	153070	
92078024	2	801836	47893	
94675771	2	801837	152969	
12625901	0	801838	12	
7548512	0	801839	8697	
34127797	0	801840	8735	
25240293	0	801841	8734	
17783154	0	801842	8734	
44514277	0	801843	8735	
50426443	0	801844	8736	
93233237	2	801845	12	
90395352	2	801846	250822	
94295208	2	801847	13	
99741549	2	801848	105863	
94013490	2	801849	11	
88500889	2	801850	225776	
91252230	2	801851	56148	
108106766	2	801852	3969	
102593450	2	801853	56148	
…	…	…	…	…

- Equal-value runs
- Equal-document runs
- Range Minimum Query on run heads of doubled-encoded ILCP ($\mathrm{RMQ}_{\mathrm{ILCP}^\star}$)

Document Listing with Frequencies

The total time is $\mathcal{O}(m + ndoc \cdot \log(n/r))$.

Interleaved Longest Common Prefix with Double Encoding

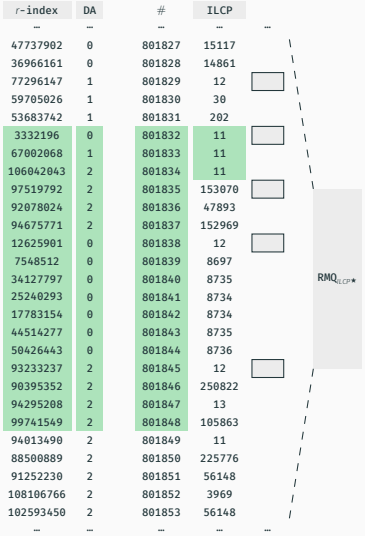


- Equal-value runs
- Equal-document runs
- Range Minimum Query on run heads of doubled-encoded ILCP ($\mathrm{RMQ}_{\mathrm{ILCP}\star}$)

Document Listing with Frequencies

The total time is $\mathcal{O}(m + ndoc \cdot \log(n/r))$.

Interleaved Longest Common Prefix with Double Encoding

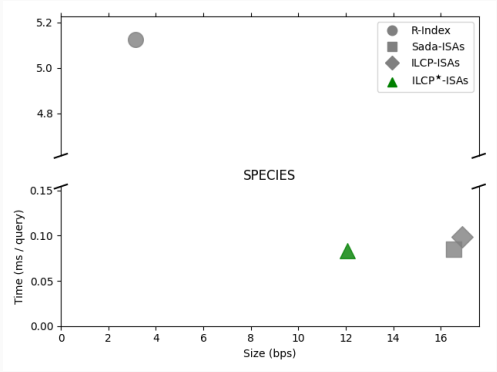


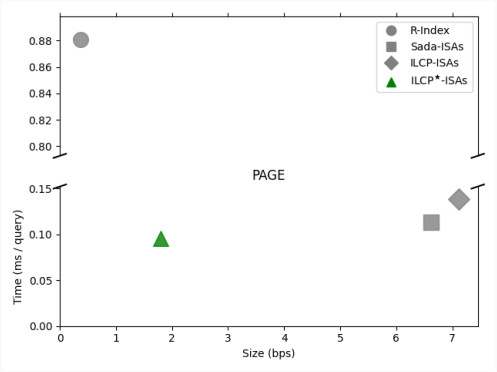
- Equal-value runs
- Equal-document runs
- Range Minimum Query on run heads of doubled-encoded ILCP ($\text{RMQ}_{\text{ILCP}\star}$)

Document Listing with Frequencies

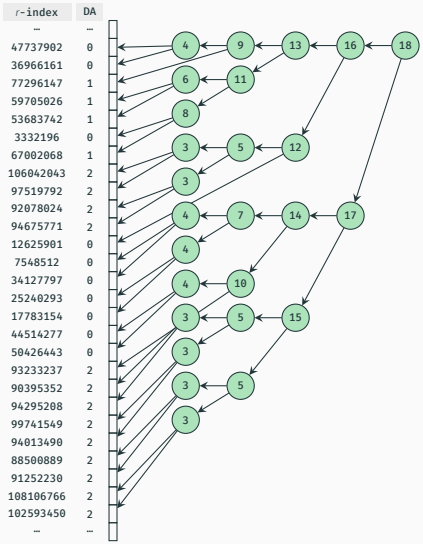
The total time is $\mathcal{O}(m + ndoc \cdot \log(n/r))$.

Interleaved Longest Common Prefix with Double Encoding: Results





Grammar-compressed Document Array with Bitvectors



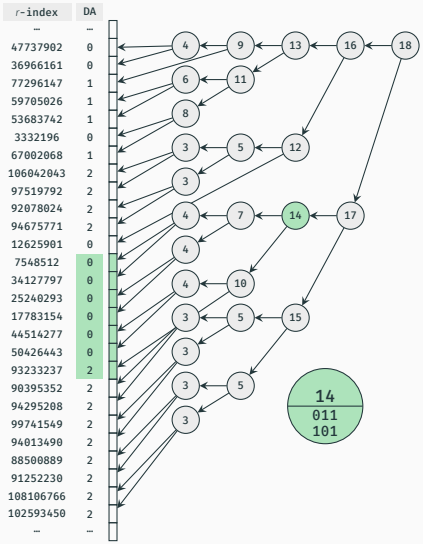
- Grammar-compressed Document Array (GCDA)
- Bitvectors L and R
- Coverage of $[s_p..e_p]$ on the parse tree T

Document Listing with Frequencies

The total time is
$\mathcal{O}(m + ndoc((t/w)\log n + \log(n/r)))$

(Cobas and Navarro, 2019)

Grammar-compressed Document Array with Bitvectors



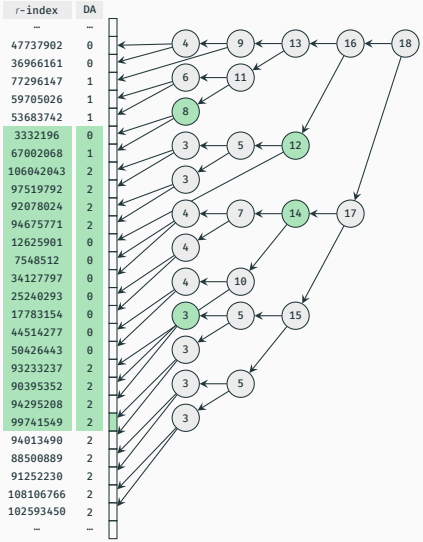
- Grammar-compressed Document Array (GCDA)
- Bitvectors L and R
- Coverage of $[s_p..e_p]$ on the parse tree T

Document Listing with Frequencies

The total time is
$\mathcal{O}(m + ndoc((t/w)\log n + \log(n/r)))$

(Cobas and Navarro, 2019)

Grammar-compressed Document Array with Bitvectors



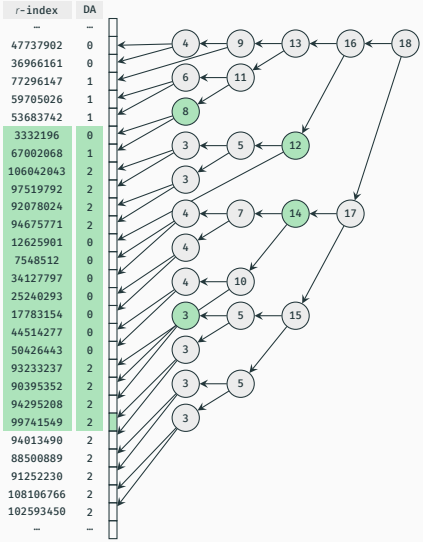
- Grammar-compressed Document Array (GCDA)
- Bitvectors L and R
- Coverage of $[s_p..e_p]$ on the parse tree T

Document Listing with Frequencies

The total time is
$\mathcal{O}(m + ndoc((t/w)\log n + \log(n/r)))$

(Cobas and Navarro, 2019)

Grammar-compressed Document Array with Bitvectors



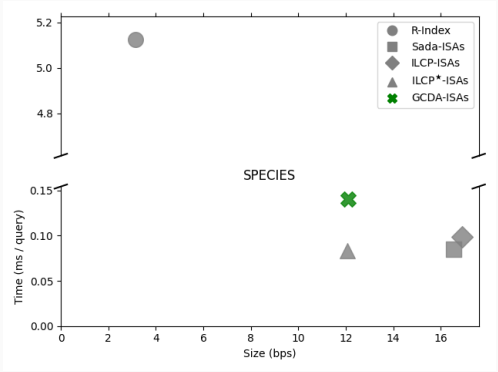
- Grammar-compressed Document Array (GCDA)
- Bitvectors L and R
- Coverage of $[s_p..e_p]$ on the parse tree T

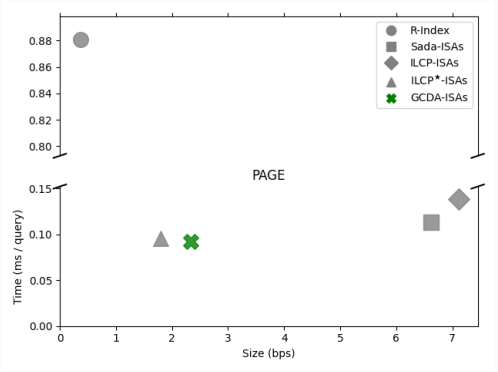
Document Listing with Frequencies

The total time is
$\mathcal{O}(m + ndoc((t/w)\log n + \log(n/r)))$

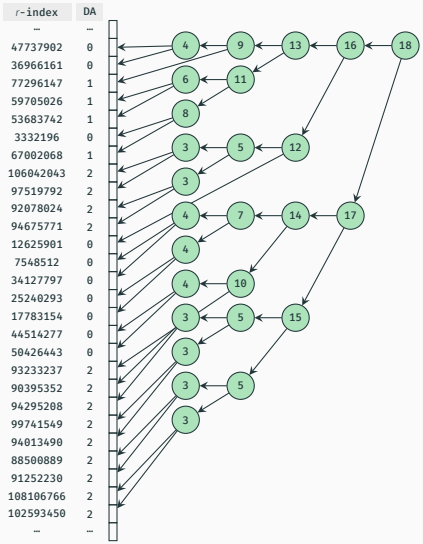
(Cobas and Navarro, 2019)

Grammar-compressed Document Array with Bitvectors: Results





Grammar-compressed Document Array with Precomputed Document Lists



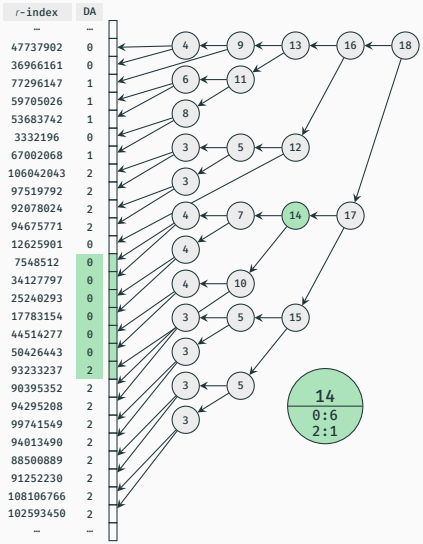
- Grammar-compressed Document Array (GCDA)
- Precomputed Document Lists (PDL) with Frequency
- Coverage of $[s_p..e_p]$ on the parse tree T

Document Listing with Frequencies

The total time is $\mathcal{O}(m + ndoc \cdot \log n)$

(Cobas and Navarro, 2019)

Grammar-compressed Document Array with Precomputed Document Lists



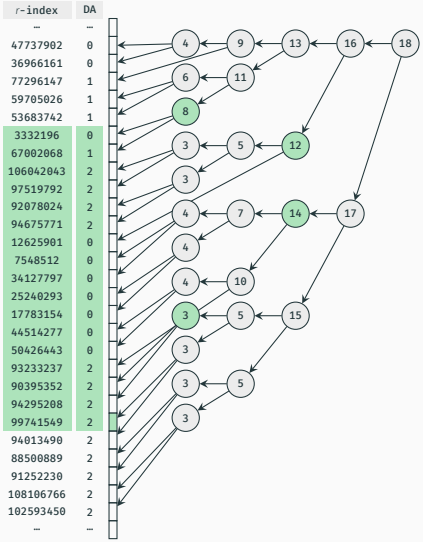
- Grammar-compressed Document Array (GCDA)
- Precomputed Document Lists (PDL) with Frequency
- Coverage of $[s_p..e_p]$ on the parse tree T

Document Listing with Frequencies

The total time is $\mathcal{O}(m + ndoc \cdot \log n)$

(Cobas and Navarro, 2019)

Grammar-compressed Document Array with Precomputed Document Lists



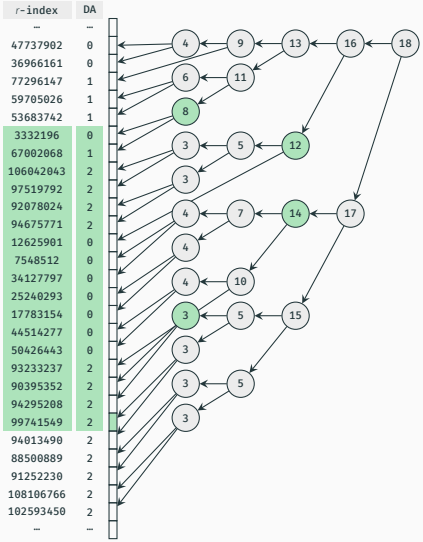
- Grammar-compressed Document Array (GCDA)
- Precomputed Document Lists (PDL) with Frequency
- Coverage of $[s_p..e_p]$ on the parse tree T

Document Listing with Frequencies

The total time is $\mathcal{O}(m + ndoc \cdot \log n)$

(Cobas and Navarro, 2019)

Grammar-compressed Document Array with Precomputed Document Lists



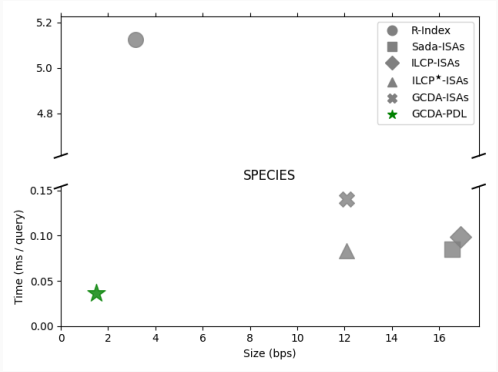
- Grammar-compressed Document Array (GCDA)
- Precomputed Document Lists (PDL) with Frequency
- Coverage of $[s_p..e_p]$ on the parse tree T

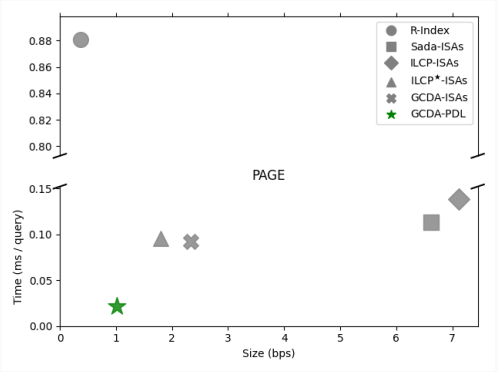
Document Listing with Frequencies

The total time is $\mathcal{O}(m + ndoc \cdot \log n)$

(Cobas and Navarro, 2019)

GCDA with Precomputed Document Lists: Results





Conclusion

Conclusion

Document Listing with Frequencies extending previous Document Listing solutions.

- Term Frequency Scheme.
- Precomputed Document Lists with Frequency.

Future Work

- Study of new solutions.
- Integration of the results with real pseudoaligners.

Thank You!

Wavelet Tree Version: Results