b-move: Faster Lossless Approximate Pattern Matching in a Run-Length Compressed Index

Additional File

Lore Depuydt
Corresponding author
Ghent University - imec, Technologiepark 126, 9052 Ghent, Belgium

Luca Renders
Ghent University - imec, Technologiepark 126, 9052 Ghent, Belgium

Simon Van de Vyver
Ghent University, Technologiepark 126, 9052 Ghent, Belgium

Lennart Veys
Ghent University, Technologiepark 126, 9052 Ghent, Belgium

Travis Gagie
Dalhousie University, 6050 University Avenue, PO BOX 15000, Halifax, NS B3H 4R2, Canada

Jan Fostier
Corresponding author
Ghent University - imec, Technologiepark 126, 9052 Ghent, Belgium

Table S1 Example index for reverse search text T^{rev} = "\$CTGGTTGTATACTGTATC" with BWT run details.

i	T^{rev}	SA^{rev}	BWT^{rev}	LF^{rev}	$T^{\text{rev}}_{\text{SA}^{\text{rev}}[i]}$
0	\$	0	C	4	\$CTGGTTGTATACTGTATC
1	C	11	T	11	ACTGTATC
2	T	9	T	12	ATACTGTATC
3	G	16	T	13	ATC
4	G	18	T	14	C
5	T	1	\$	0	CTGGTTGTATACTGTATC
6	T	12	A	1	CTGTATC
7	G	3	T	15	GGTTGTATACTGTATC
8	T	7	T	16	GTATACTGTATC
9	A	14	T	17	GTATC
10	T	4	G	7	GTTGTATACTGTATC
11	A	10	A	2	TACTGTATC
12	C	8	G	8	TATACTGTATC
13	T	15	G	9	TATC
14	G	17	A	3	TC
15	T	2	C	5	TGGTTGTATACTGTATC
16	A	6	T	18	TGTATACTGTATC
17	T	13	C	6	TGTATC
18	C	5	G	10	TTGTATACTGTATC

Reverse search text T^{rev} = "\$CTGGTTGTATACTGTATC" with its suffix array SA^{rev}, Burrows-Wheeler transform BWT^{rev}, LF mapping LF^{rev}, and suffixes. BWT runs are delineated by horizontal lines.

Table S2 Overview of the components in the bidirectional r-index and the bidirectional move structure.

Component	Description	Memory usage for 3 264 *E. coli* genomes [MB]
Bidirectional r-index as described in [1]: character extension support		
Reg. bv. runs	Sparse bitvector with r 1-bits indicating run endings in the BWT of T	78.56
Reg. bv. runs-per-letter	Division of the runs bitvector into $\|\Sigma\|$ shorter sparse bitvectors per character	142.37
Reg. run heads	Characters corresponding to runs in the BWT of T	49.13
Rev. bv. runs	Analogous for T^{rev}	78.56
Rev. bv. runs-per-letter	Analogous for T^{rev}	142.37
Rev. run heads	Analogous for T^{rev}	49.13
Total	-	540.13
Bidirectional move structure (this paper): character extension support		
Reg. move table M_{LF}	Bit-packed LF move table for T	1 535.97
Rev. move table $M_{\text{LF}}^{\text{rev}}$	Bit-packed LF move table for T^{rev}	1 535.95
Total	-	3 071.92
Bidirectional r-index as described in [1]: ϕ and ϕ^{-1} support		
Text to run mappings	Mappings to link a ϕ/ϕ^{-1} run index back to its BWT run	797.53
ϕ/ϕ^{-1} run bv.'s	Bitvectors to indicate the ϕ/ϕ^{-1} runs	284.60
Total	-	1 082.13
Bidirectional move structure (this paper): ϕ and ϕ^{-1} support		
ϕ move table M_{ϕ}	Bit-packed ϕ move table	2 146.95
ϕ^{-1} move table $M_{\phi^{-1}}$	Bit-packed ϕ^{-1} move table	2 146.64
ϕ/ϕ^{-1} run bv.'s	Bitvectors to indicate the ϕ/ϕ^{-1} runs, necessary to initiate the locate process	406.32
Total	-	4 699.91
Both indexes: toehold update and locate support [1]		
SA samples	SA and SA^{rev} samples for toehold updates	2 008.56
PLCP	PLCP for locate support	196.99
Total	-	2 205.55

Algorithm S1 Update intervals $([s, e], R_s, R_e)$ and $([s^{\text{rev}}, e^{\text{rev}}], R_s^{\text{rev}}, R_e^{\text{rev}})$ corresponding to P, to intervals $([s', e'], R_s, R_e)$ and $([s^{\text{rev}\prime}, e^{\text{rev}\prime}], R_s^{\text{rev}\prime}, R_e^{\text{rev}\prime})$ that correspond to Pc.

def `addCharToRight`$(([s, e], R_s, R_e), ([s^{\text{rev}}, e^{\text{rev}}], R_s^{\text{rev}}, R_e^{\text{rev}}), c)$**:**
  $([s^{\text{rev}\prime}, e^{\text{rev}\prime}], R_s^{\text{rev}\prime}, R_e^{\text{rev}\prime}) \leftarrow M_{\text{LF}}^{\text{rev}}$.`addChar`$([s^{\text{rev}}, e^{\text{rev}}], R_s^{\text{rev}}, R_e^{\text{rev}}, c)$
  if $s^{\text{rev}\prime} = -1$ **then**
    return empty intervals
  if $e^{\text{rev}\prime} - s^{\text{rev}\prime} = e^{\text{rev}} - s^{\text{rev}}$ **then**
    return $([s, e], R_s, R_e), ([s^{\text{rev}\prime}, e^{\text{rev}\prime}], R_s^{\text{rev}\prime}, R_e^{\text{rev}\prime})$
  $x \leftarrow 0$
  for a **where** $a \prec c$ **do**
    $([s^{\text{rev}\prime\prime}, e^{\text{rev}\prime\prime}], R_s^{\text{rev}\prime\prime}, R_e^{\text{rev}\prime\prime}) \leftarrow M_{\text{LF}}^{\text{rev}}$.`addChar`$([s^{\text{rev}}, e^{\text{rev}}], R_s^{\text{rev}}, R_e^{\text{rev}}, a)$
    if $s^{\text{rev}\prime\prime} \neq -1$ **then**
      $x \leftarrow x + e^{\text{rev}\prime\prime} - s^{\text{rev}\prime\prime} + 1$
  $s' \leftarrow s + x$
  $e' \leftarrow s + x + (e^{\text{rev}\prime} - s^{\text{rev}\prime})$
  return $([s', e'], R_s, R_e), ([s^{\text{rev}\prime}, e^{\text{rev}\prime}], R_s^{\text{rev}\prime}, R_e^{\text{rev}\prime})$

References

1 Yuma Arakawa, Gonzalo Navarro, and Kunihiko Sadakane. Bi-Directional r-Indexes. In *33rd Annual Symposium on Combinatorial Pattern Matching (CPM 2022)*, volume 223 of *Leibniz International Proceedings in Informatics (LIPIcs)*, pages 11:1–11:14, Dagstuhl, Germany, 2022. Schloss Dagstuhl – Leibniz-Zentrum für Informatik. doi:10.4230/LIPIcs.CPM.2022.11.