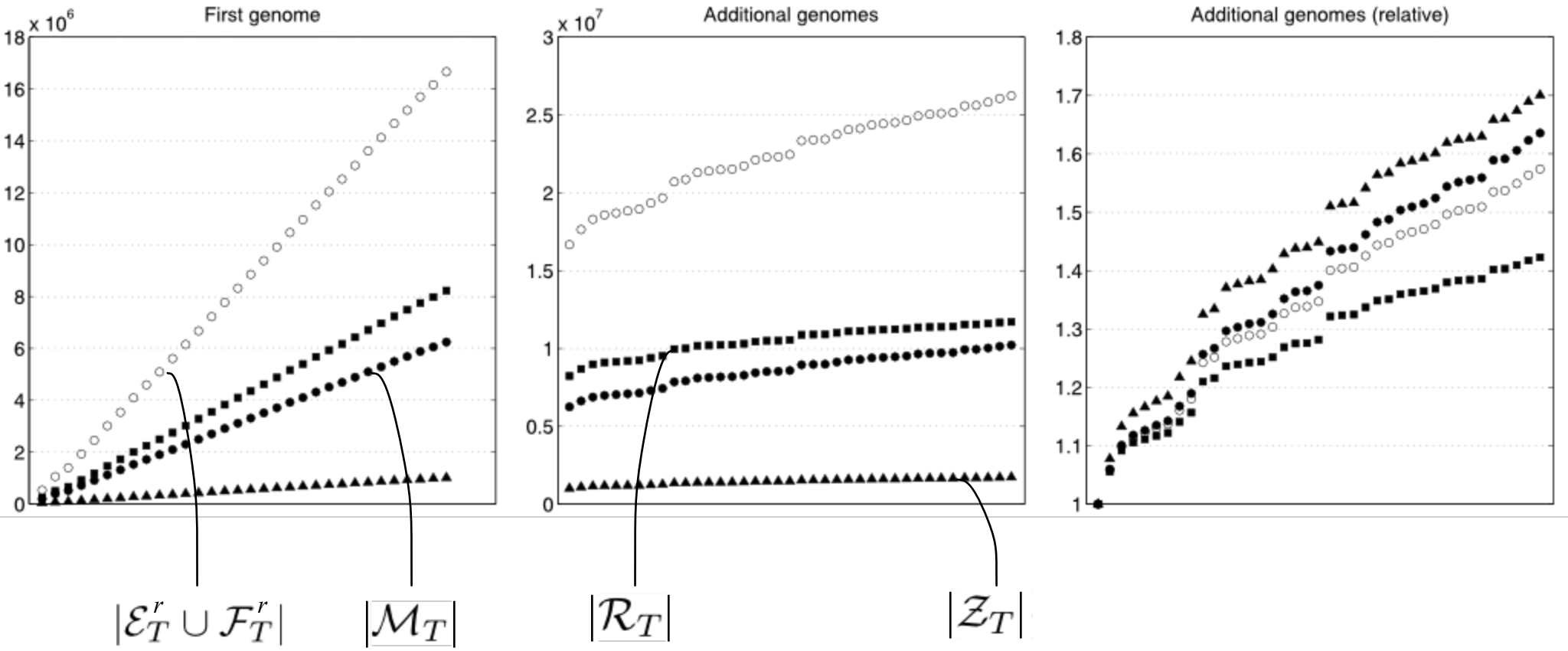
Composite repetition-aware data structures

Djamal Belazzougui[1], **Fabio Cunial[2]**, Travis Gagie[1], Nicola Prezza[3], Mathieu Raffinot[4]

(1) Department of Computer Science, University of Helsinki, Finland.
(2) Max Planck Institute for Molecular Cell Biology and Genetics, Dresden, Germany.
(3) Department of Mathematics and Computer Science, University of Udine, Italy.
(4) LIAFA, Paris Diderot University - Paris 7, France.

Highly-repetitive strings

39 *Saccharomyces cerevisiae* genomes



Distinct measures of repetition all grow sublinearly

[1] Paolo Ferragina and Gonzalo Navarro. Pizza&Chili repetitive corpus. Accessed: 2015-01-25. `http://pizzachili.dcc.uchile.cl/repcorpus.html`

Combining repetition-aware data structures

RLBWT$_T$

$\mathcal{R}_T$

LZ77 index

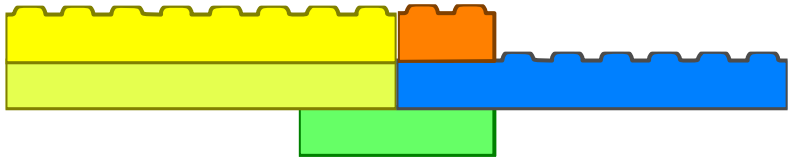
$\mathcal{Z}_T$

CDAWG$_T$

$\mathcal{M}_T$, $\mathcal{E}_T \cup \mathcal{F}_T$

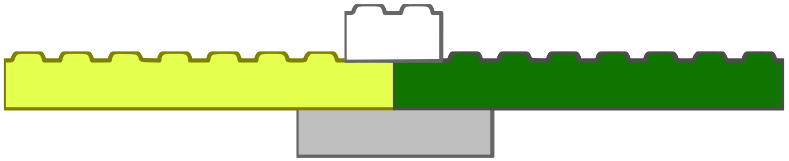
Locating

RLBWT$_{\overline{T}}$



Locating

Suffix tree representations



Locating

Words:

RLBWT+CDAWG	$O(\lvert\mathcal{E}_T^r \cup \mathcal{F}_T^r\rvert)$
RLBWT+LZ77	$O(\lvert\mathcal{Z}_T\rvert + \lvert\mathcal{R}_T\rvert + \lvert\mathcal{R}_{\overline{T}}\rvert)$
[1]	$O(n/k + \lvert\mathcal{R}_T\rvert)$

Time:

RLBWT+CDAWG	$O(m \log\log n + \mathtt{occ})$
RLBWT+LZ77	$O(m(\log\log n + \log\lvert\mathcal{Z}_T\rvert) + \mathtt{pOcc} \log^{\epsilon}\lvert\mathcal{Z}_T\rvert + \mathtt{sOcc} \log\log n)$
[2]	$O(m^2 h + (m + \mathtt{occ}) \log\lvert\mathcal{Z}_T\rvert)$
[1]	$O(m \log\log n + k \cdot \mathtt{occ} \log\log n)$

[1] Veli Mäkinen, Gonzalo Navarro, Jouni Sirén, and Niko Välimäki. *Storage and retrieval of highly repetitive sequence collections*. Journal of Computational Biology, 17(3):281–308, 2010.
[2] Sebastian Kreft and Gonzalo Navarro. *On compressing and indexing repetitive sequences*. Theoretical Computer Science, 483:115–133, 2013.

Suffix tree representation

	stringDepth locateLeaf	isAncestor	parent nextSibling	child firstChild	suffixLink	weinerLink	edgeChar	nLeaves
1	$O(1)$	$O(1)$	$O(\log\log n)$	$O(1)$	$O(\log\log n)$	$O(\log\log n)$	$O(\log\log n)$	$O(1)$
2	$O(1)$	$O(1)$	$O(\log\log n)$	$O(1)$	$O(\log\log n)$	$O(\log\log n)$		$O(1)$
3	$O(1)$		$O(\log\log n)$	$O(1)$	$O(1)$			

Words: $O(|\mathcal{E}_T^r| + |\mathcal{F}_T^r| + |\mathcal{E}_T^\ell| + |\mathcal{F}_T^\ell|)$

$O(|\mathcal{E}_T^r| + |\mathcal{F}_T^r|)$

Preliminaries

Maximal repeats

and the BWT

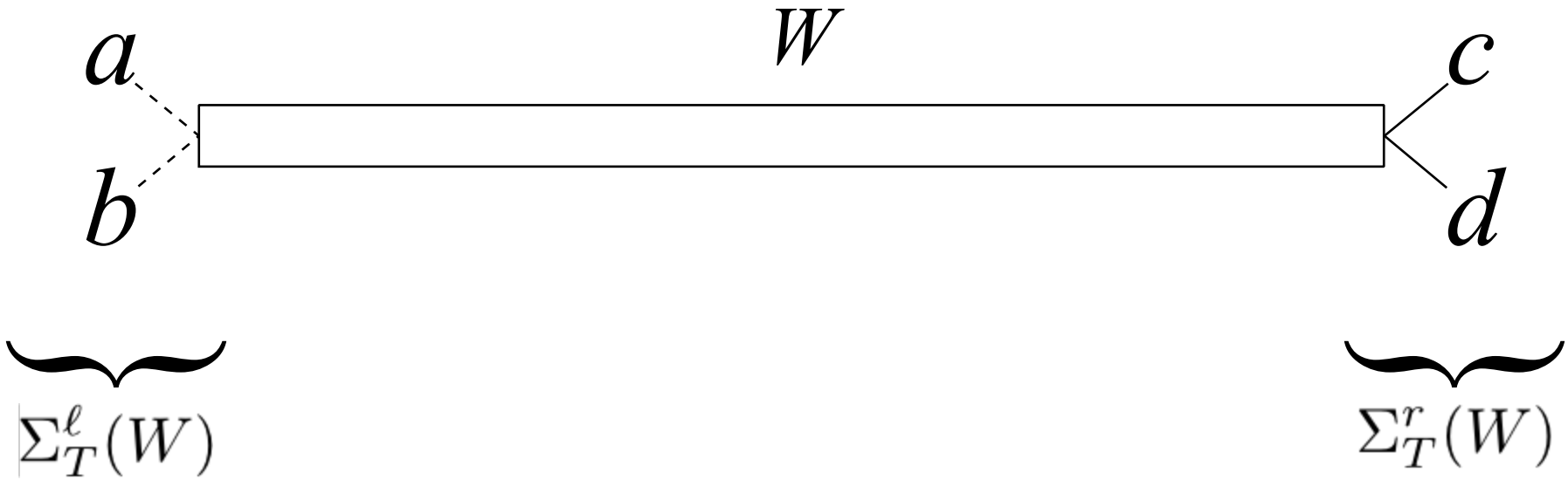
and the CDAWG

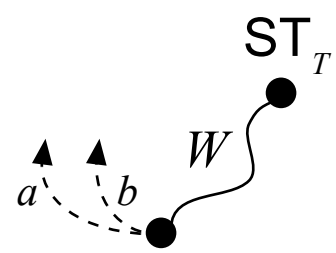
"Rightmost" maximal repeats

and the BWT

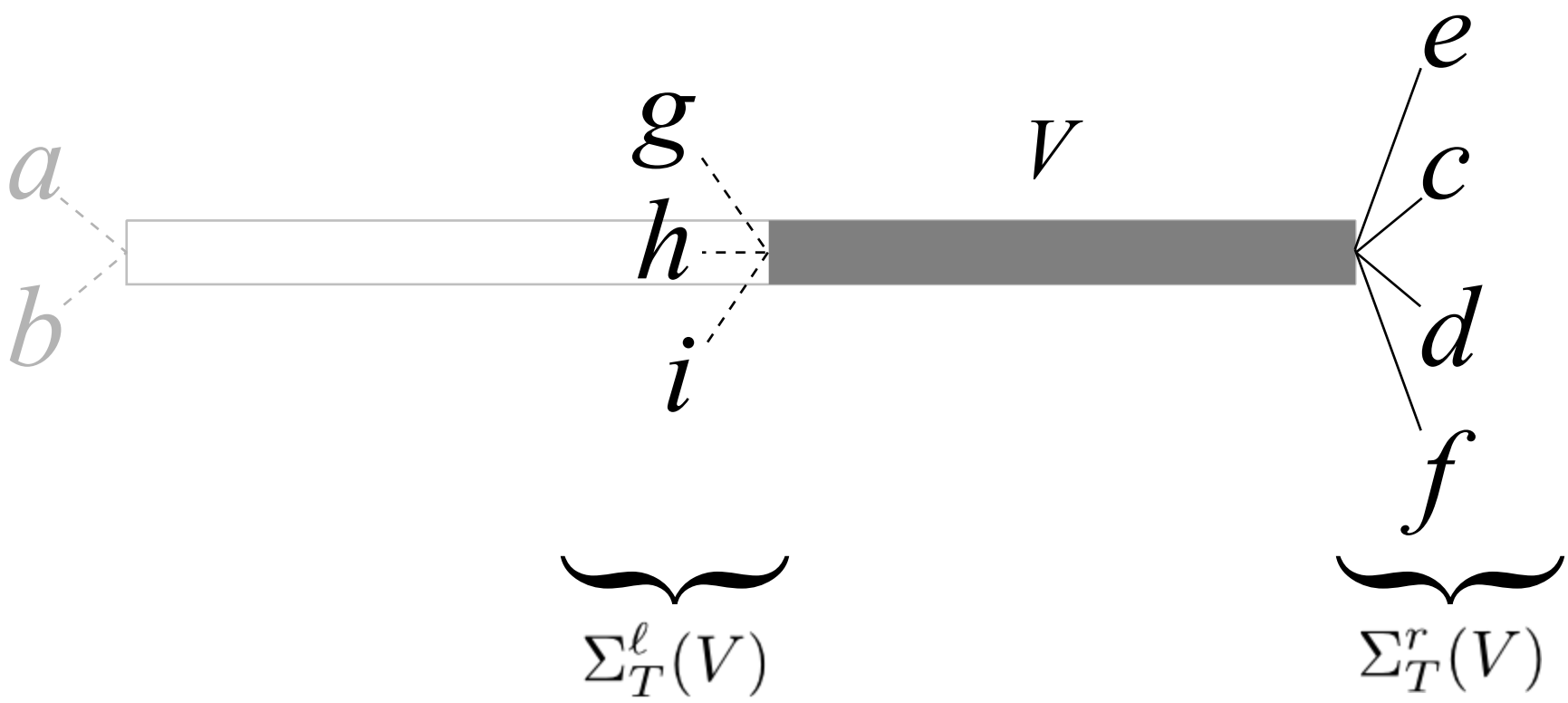
and LZ77 factors

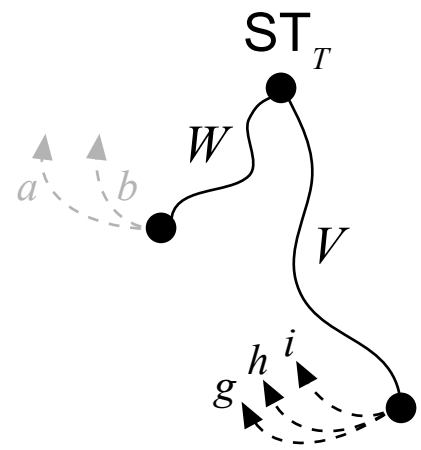
Maximal repeats



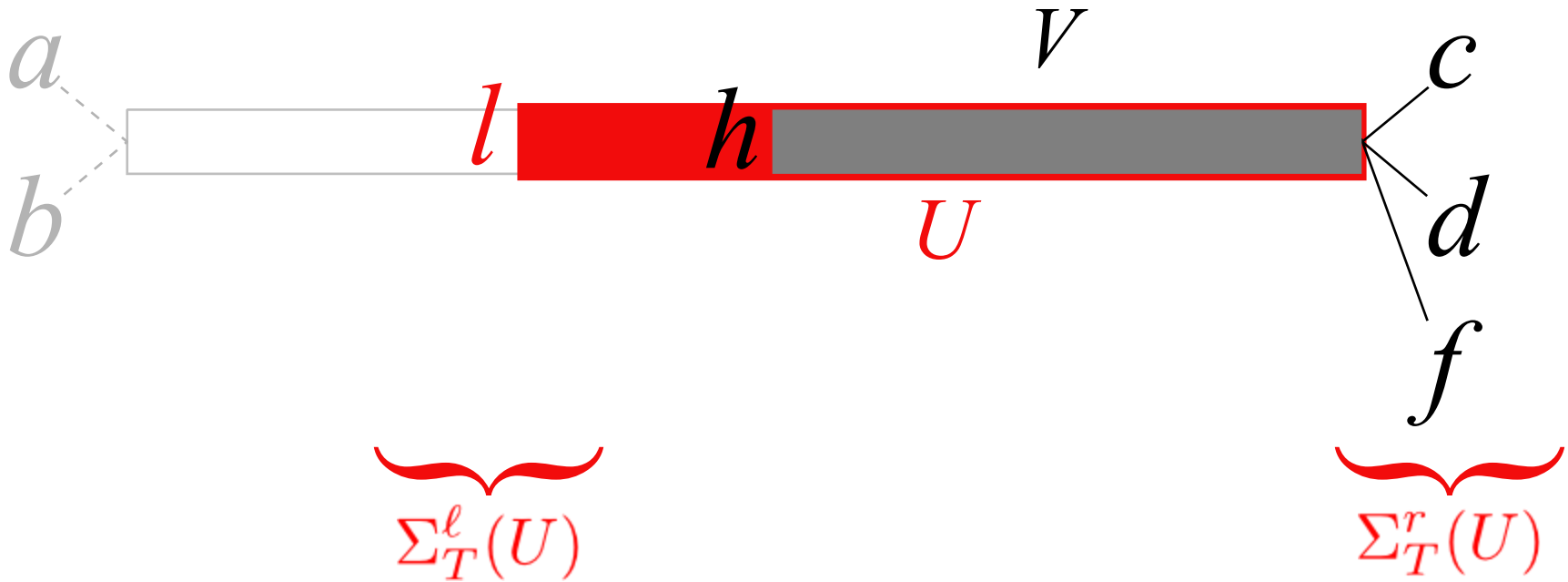


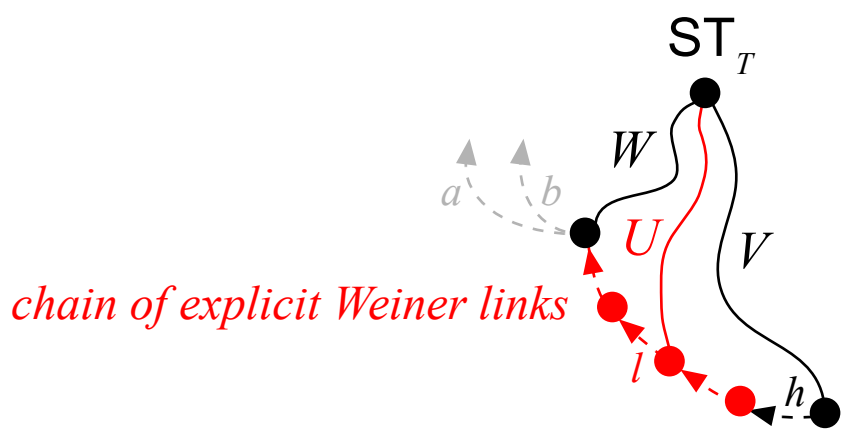
Maximal repeats



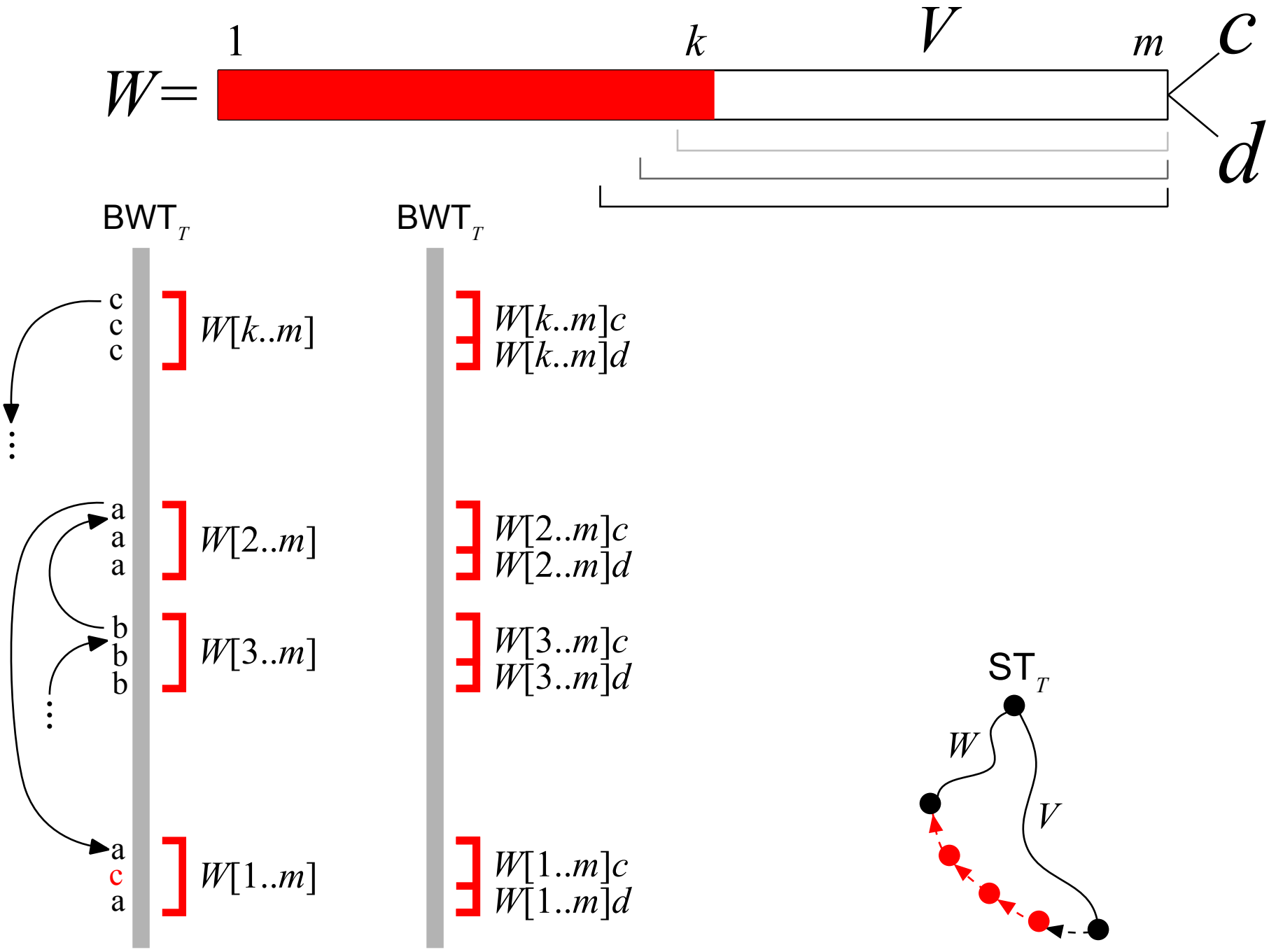


Maximal repeats

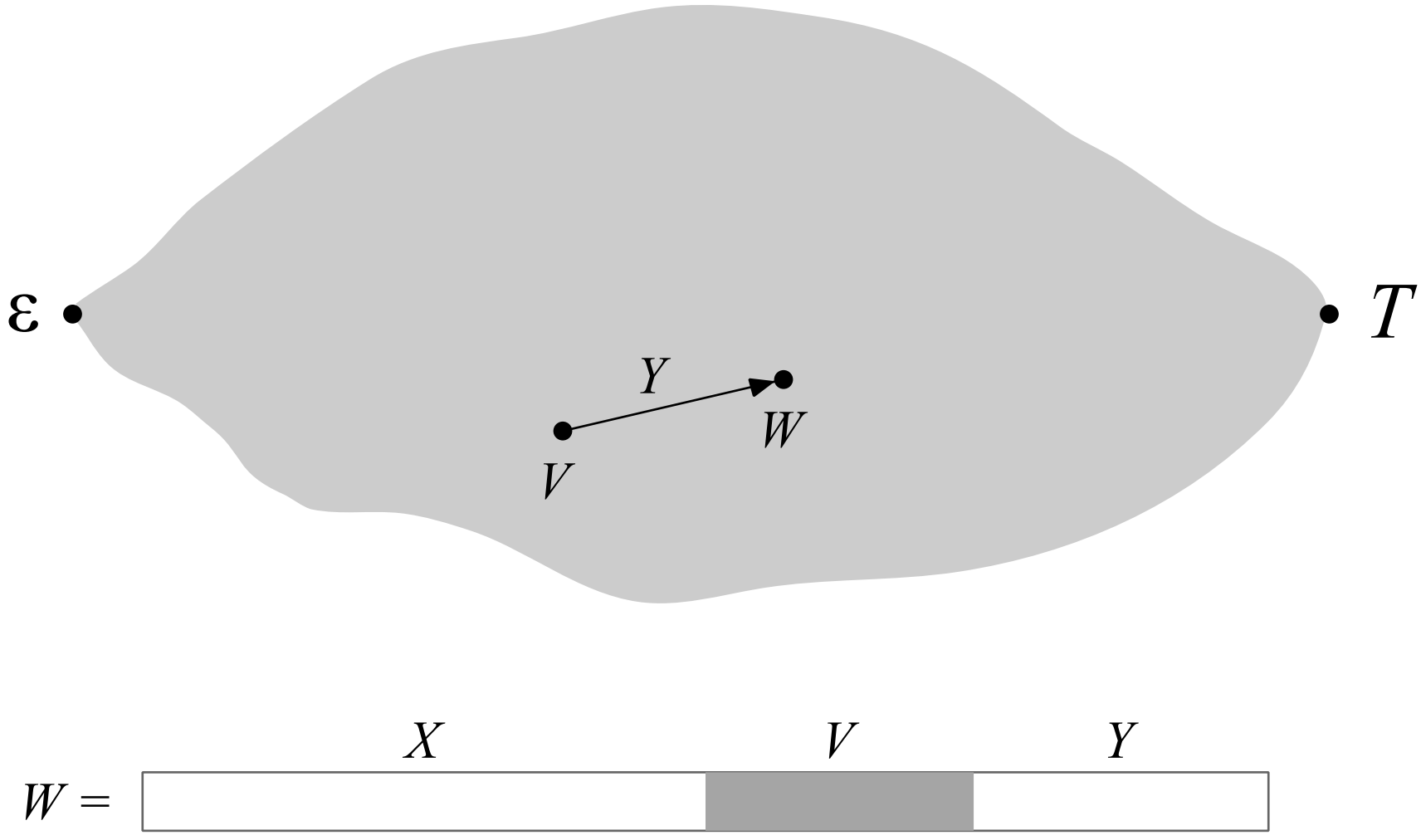




Maximal repeats and BWT

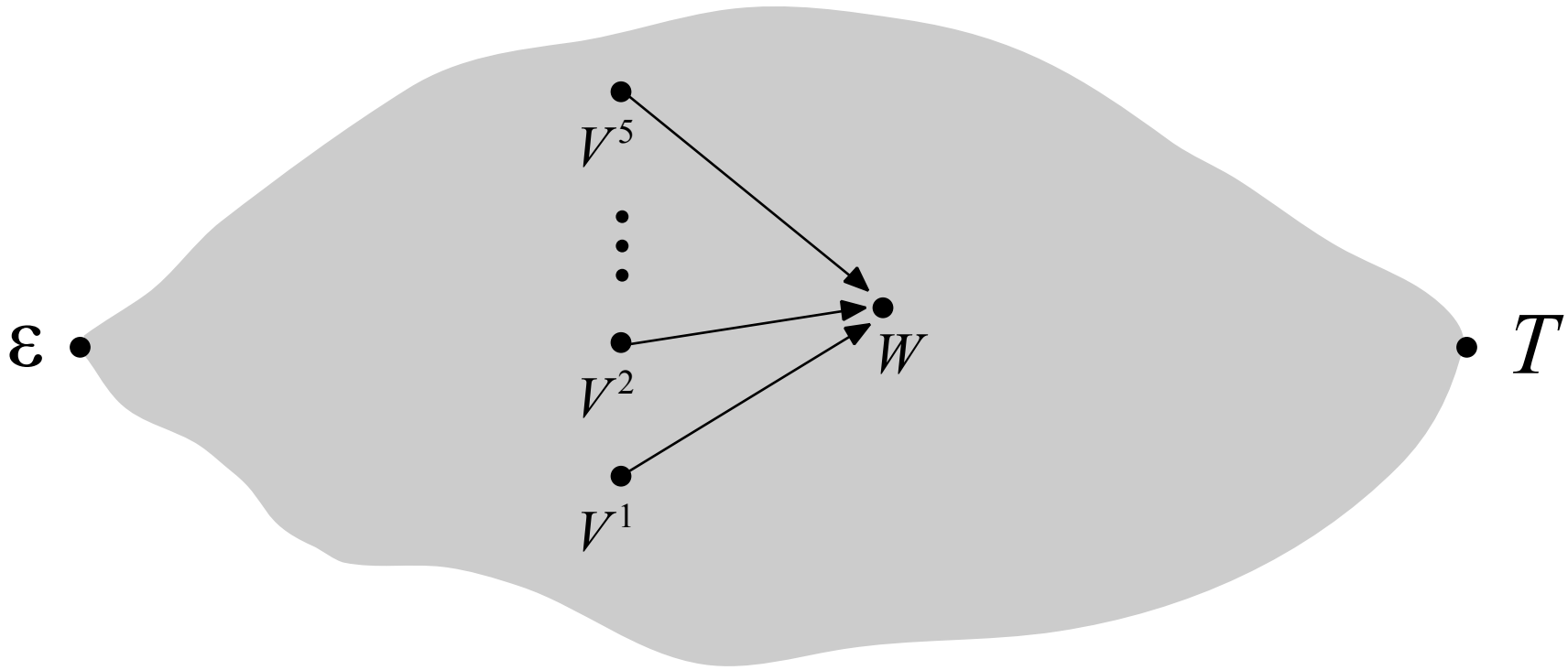


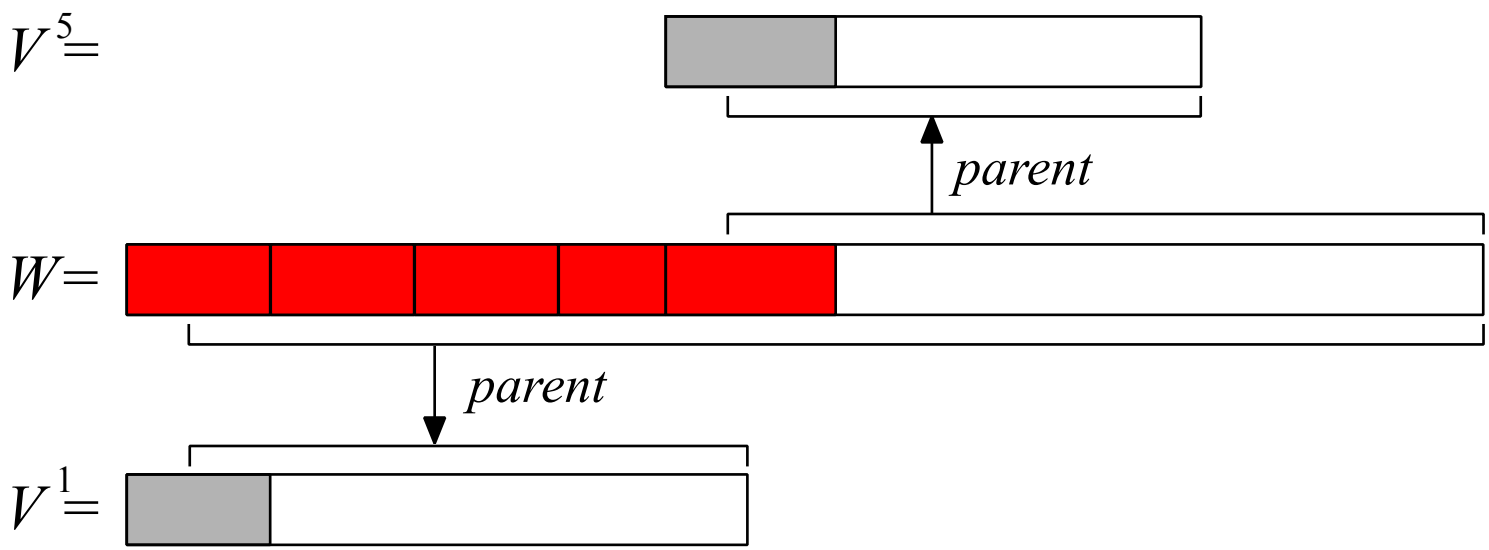
Compact Directed Acyclic Word Graph



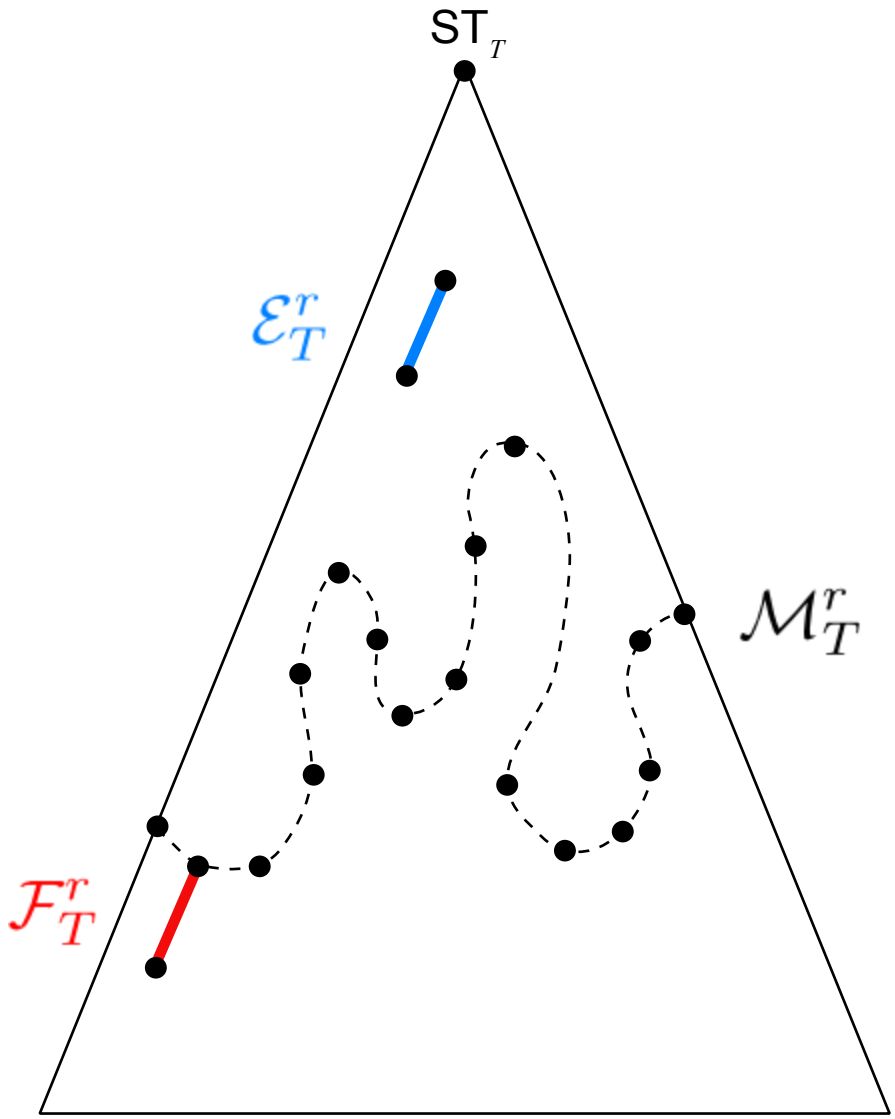
[1] Anselm Blumer, Janet Blumer, David Haussler, Ross McConnell, and Andrzej Ehrenfeucht. *Complete inverted files for efficient text retrieval and analysis*. Journal of the ACM, 34(3):578–595, 1987.
[2] Maxime Crochemore and Renaud Vérin. *Direct construction of compact directed acyclic word graphs*. In Alberto Apostolico and Jotun Hein, editors, CPM, volume 1264 of Lecture Notes in Computer Science, pages 116–129. Springer, 1997.

Maximal repeats and CDAWG

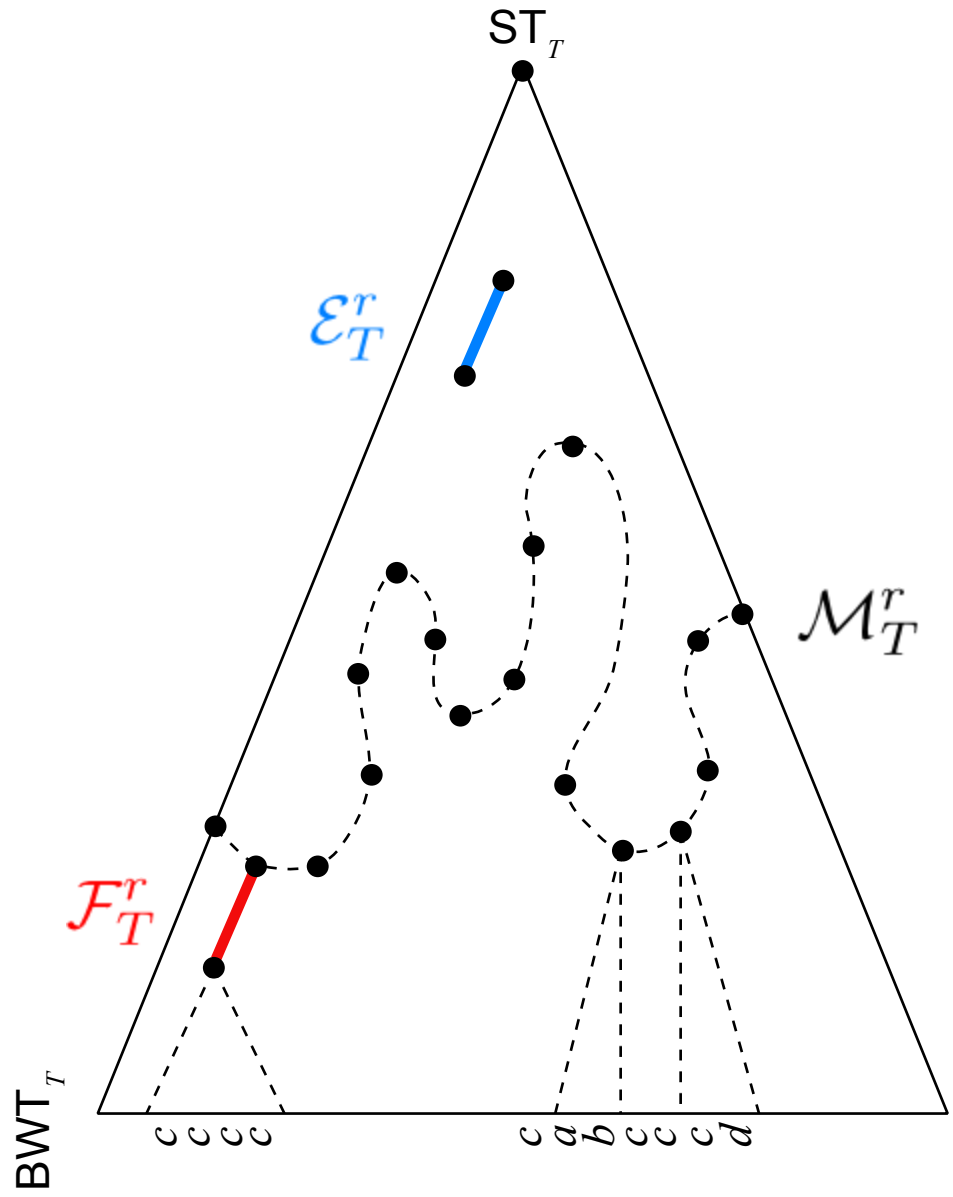




Rightmost maximal repeats



Rightmost maximal repeats and BWT runs



$$\sum_{W \in \mathcal{M}_T^r} |\Sigma_T^\ell(W)| - |\mathcal{M}_T^r| + 1 \leq |\mathcal{R}_T| \leq |\mathcal{F}_T^r|$$

Locating with RLBWT+LZ77

RLBWT$_T$

$\mathcal{R}_T$

Rank/select in $O(\log \log n)$ time, $O(|\mathcal{R}_T|)$ words (predecessor data structure).

LZ77 index

$\mathcal{Z}_T$

Primary occurrences:
$O((1 + \mathtt{occ}) \log^{\epsilon}(|\mathcal{Z}_T|))$ time,
$O(|\mathcal{Z}_T|)$ words
(4-sided range reporting).

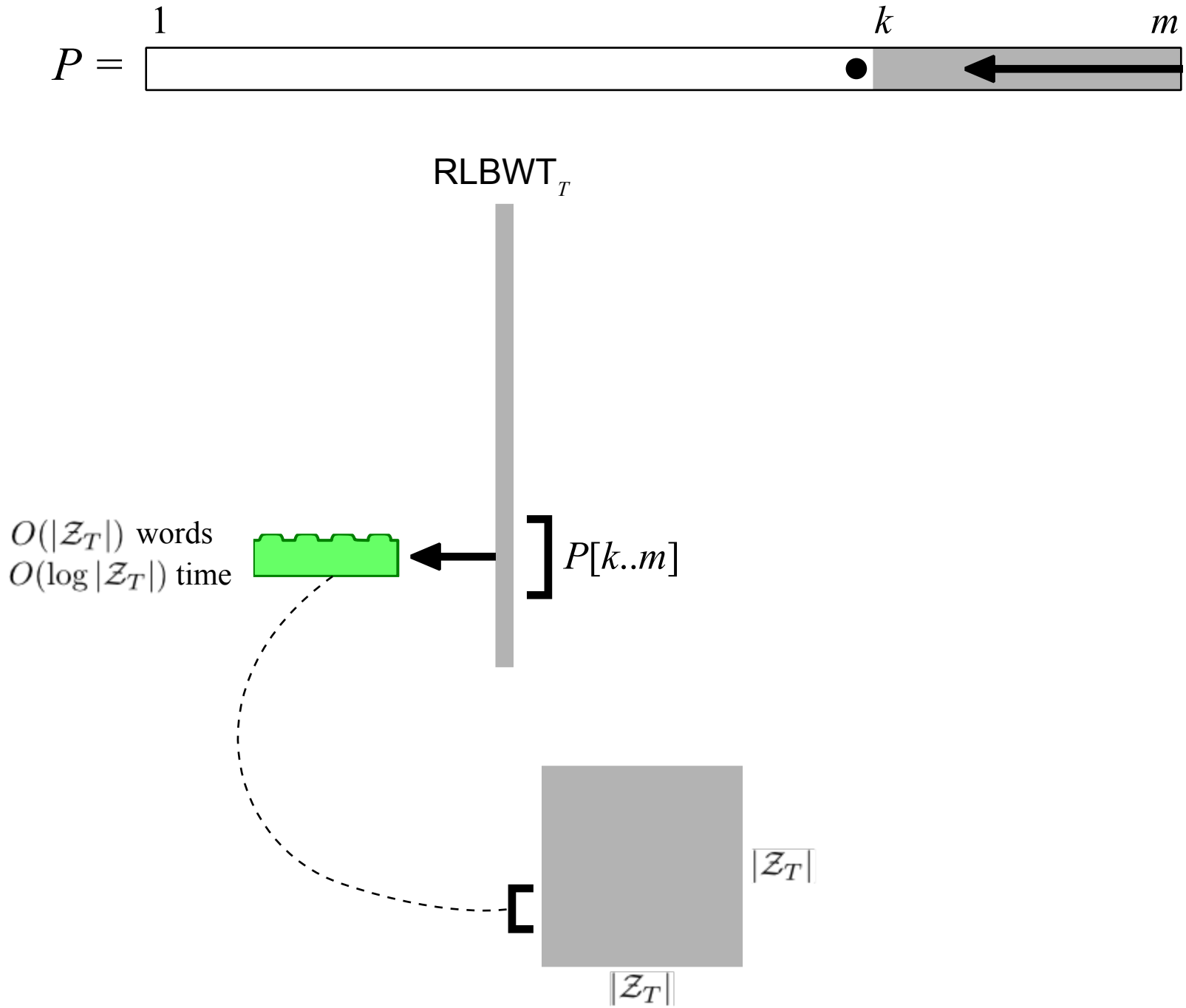
Secondary occurrences:
$O(\mathtt{occ} \log \log n)$ time,
$O(|\mathcal{Z}_T|)$ words
(2-sided range reporting).

CDAWG$_T$

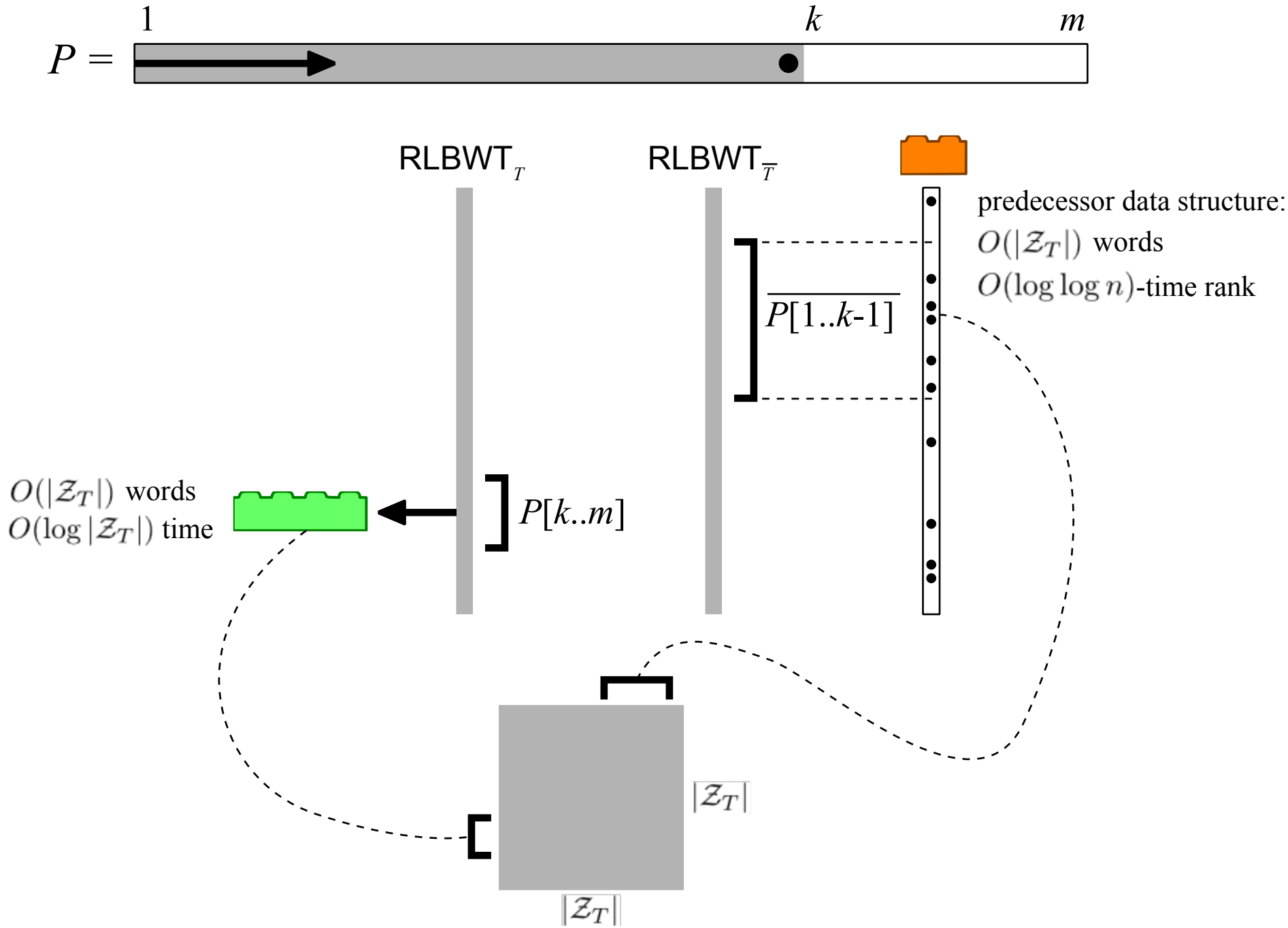
$\mathcal{M}_T$, $\mathcal{E}_T \cup \mathcal{F}_T$

[1] Dan E Willard. *Log-logarithmic worst-case range queries are possible in space Θ(N)*. Information Processing Letters, 17(2):81–84, 1983.
[2] Timothy M. Chan, Kasper Green Larsen, and Mihai Pătraşcu. *Orthogonal range searching on the RAM, revisited*. In Proceedings of the Twenty-seventh Annual Symposium on Computational Geometry, pages 1–10. ACM, 2011.
[3] Juha Kärkkäinen and Esko Ukkonen. *Lempel-Ziv parsing and sublinear-size index structures for string matching*. In Proc. 3rd South American Workshop on String Processing (WSP'96), pages 141–155, 1996.

Locating with RLBWT+LZ77



Locating with RLBWT+LZ77



Locating with RLBWT+LZ77

Words:

RLBWT+LZ77 $O(|\mathcal{Z}_T| + |\mathcal{R}_T| + |\mathcal{R}_{\overline{T}}|)$

[1] $O(n/k + |\mathcal{R}_T|)$

Time:

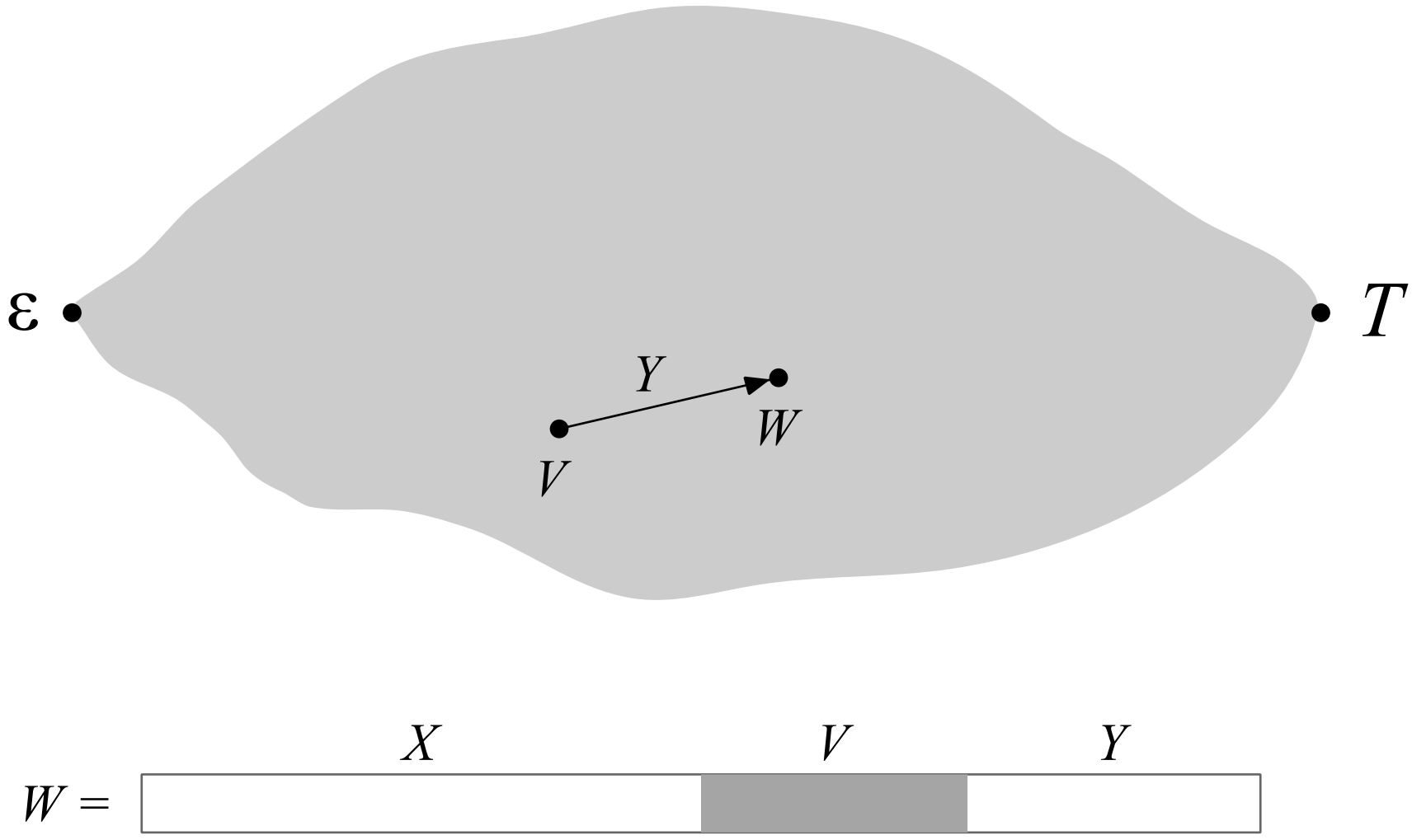
RLBWT+LZ77 $O(m(\log\log n + \log|\mathcal{Z}_T|) + \mathtt{p0cc}\log^{\epsilon}|\mathcal{Z}_T| + \mathtt{s0cc}\log\log n)$

[2] $O(m^2 h + (m + \mathtt{occ})\log|\mathcal{Z}_T|)$

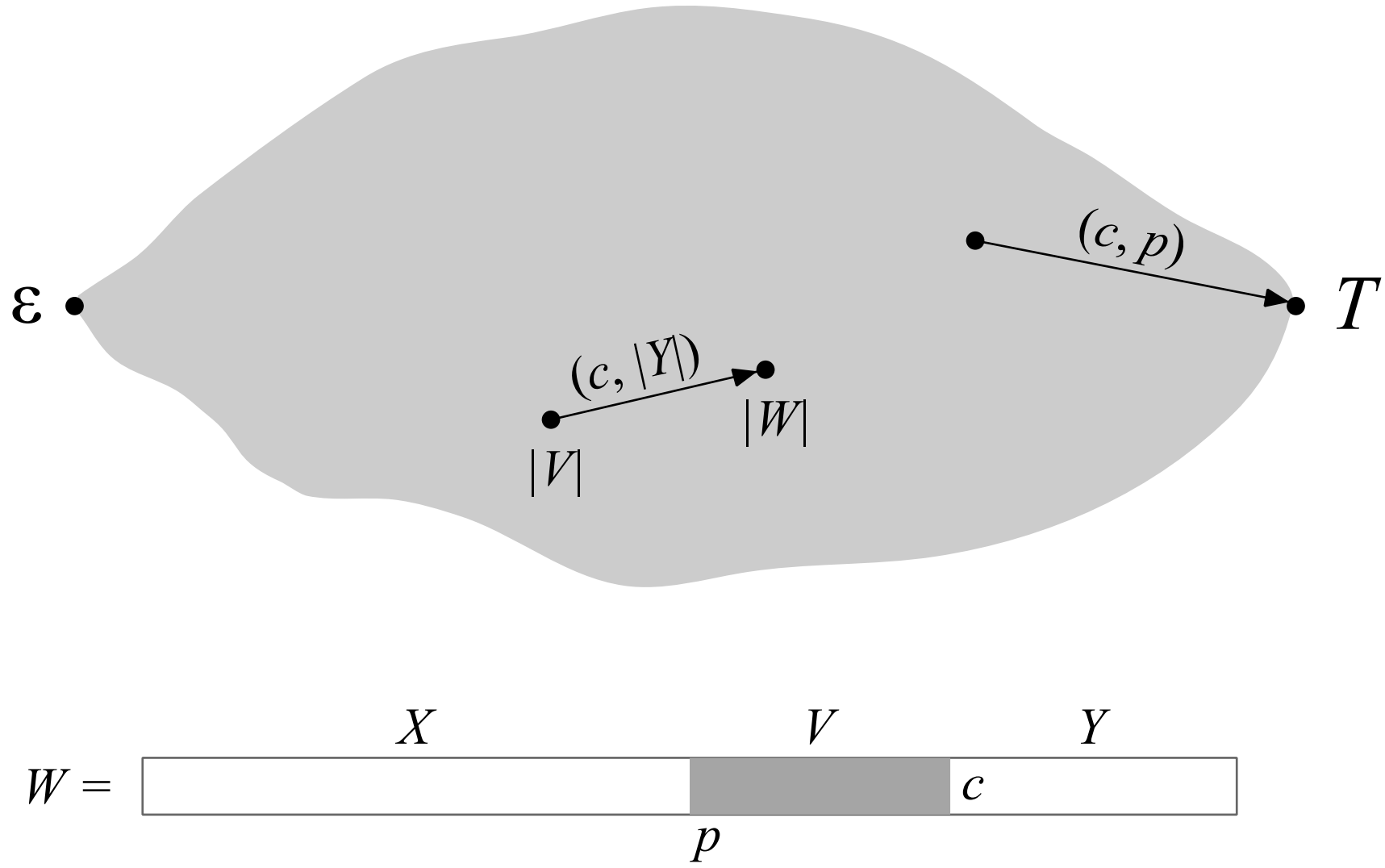
[1] $O(m\log\log n + k \cdot \mathtt{occ}\log\log n)$

[1] Veli Mäkinen, Gonzalo Navarro, Jouni Sirén, and Niko Välimäki. *Storage and retrieval of highly repetitive sequence collections*. Journal of Computational Biology, 17(3):281–308, 2010.
[2] Sebastian Kreft and Gonzalo Navarro. *On compressing and indexing repetitive sequences*. Theoretical Computer Science, 483:115–133, 2013.

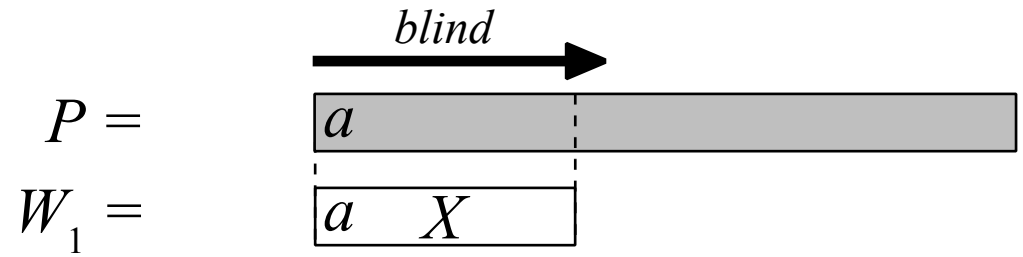
Locating with RLBWT+CDAWG

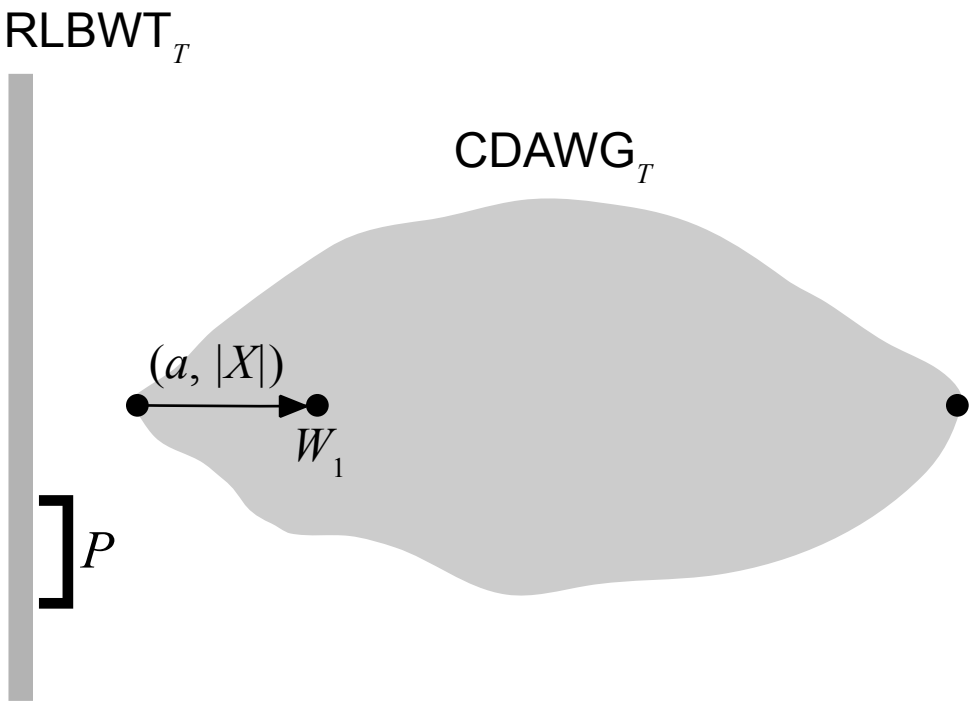


Locating with RLBWT+CDAWG



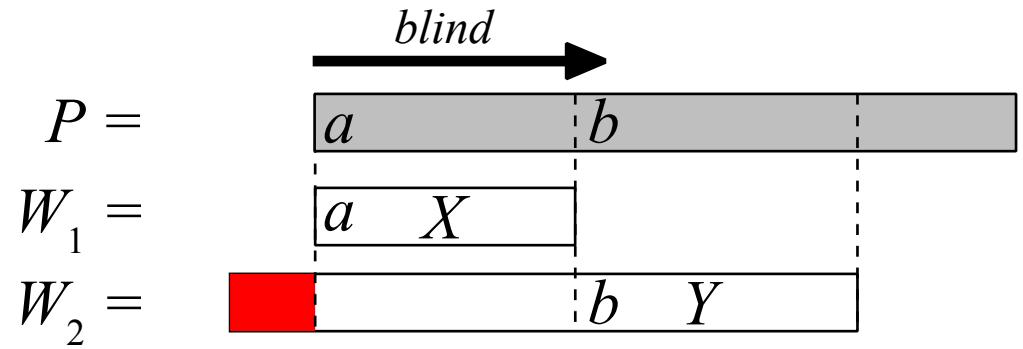
Locating with RLBWT+CDAWG

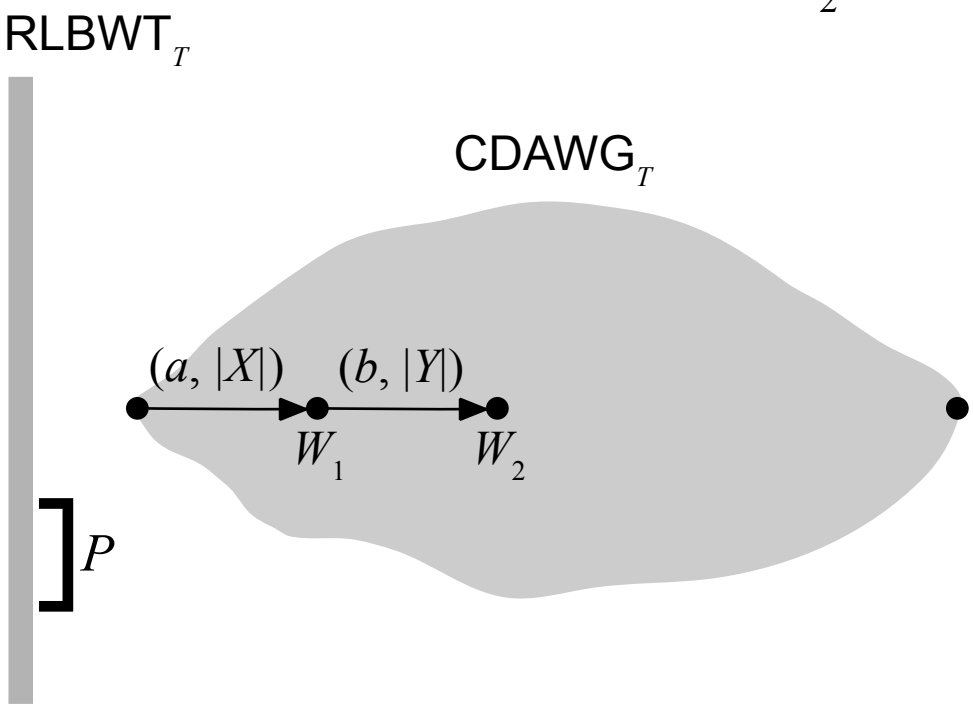




[1] Maxime Crochemore and Christophe Hancart. *Automata for matching patterns*. In Handbook of formal languages, pages 399–462. Springer, 1997.

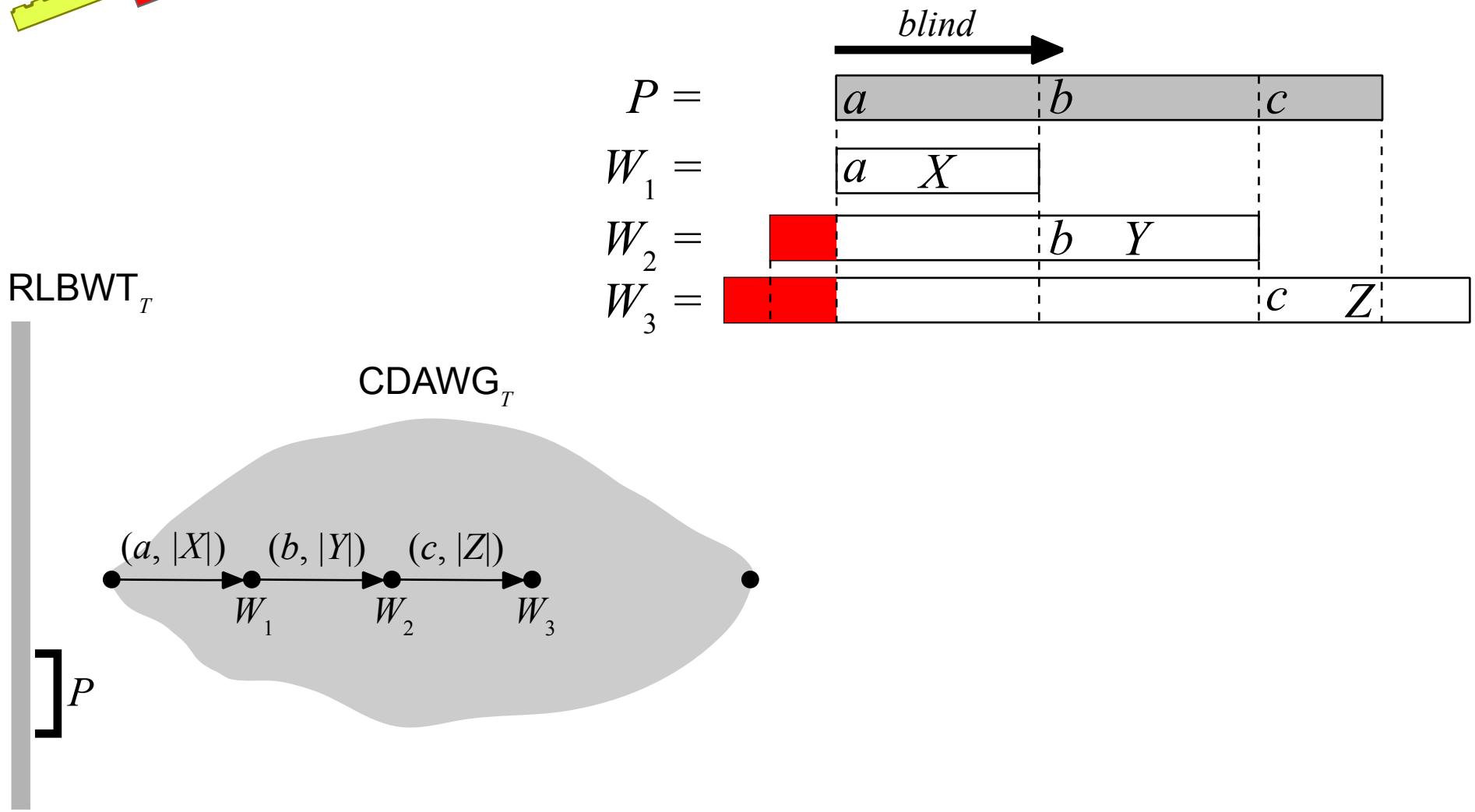
Locating with RLBWT+CDAWG





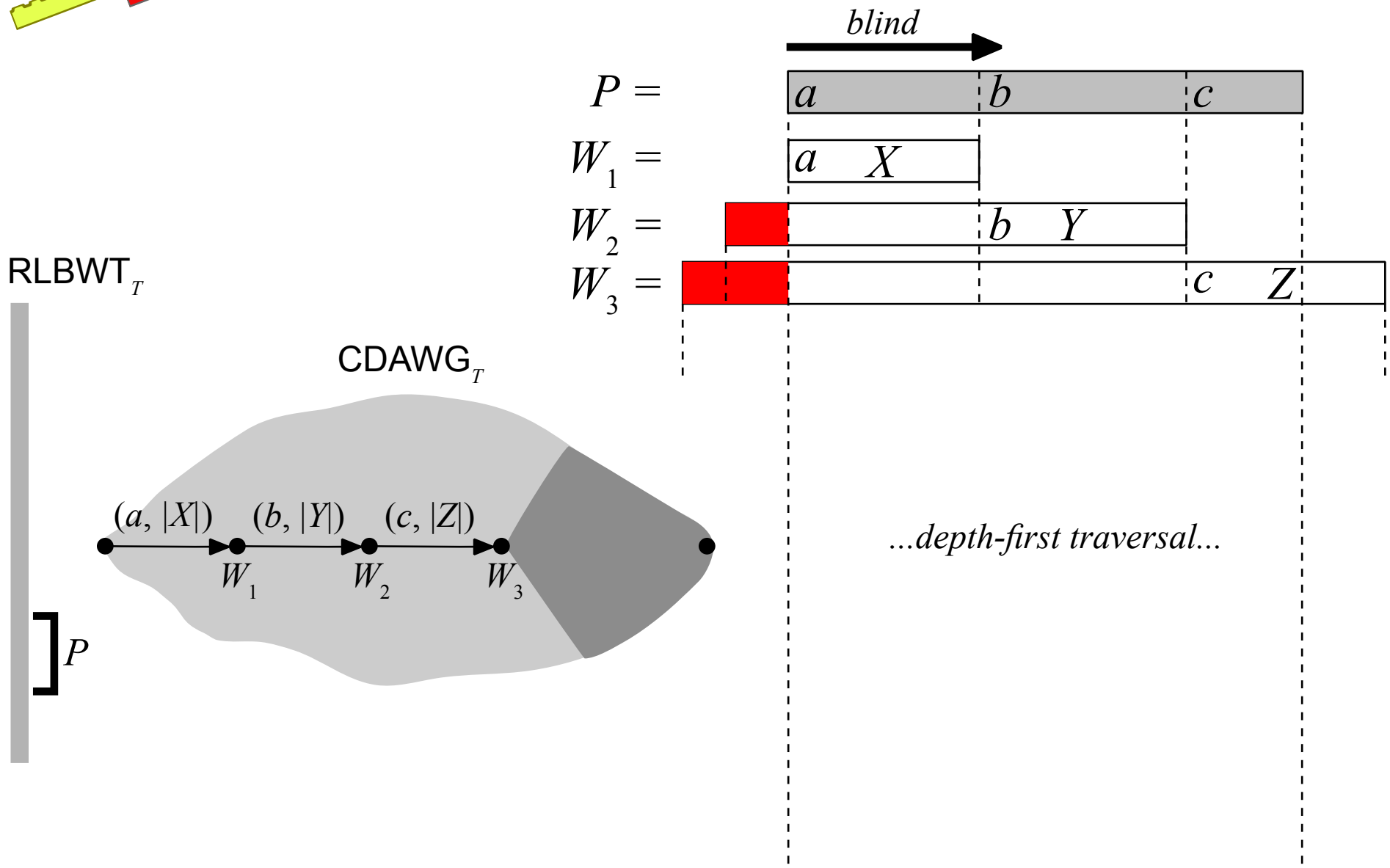
[1] Maxime Crochemore and Christophe Hancart. *Automata for matching patterns*. In Handbook of formal languages, pages 399–462. Springer, 1997.

Locating with RLBWT+CDAWG



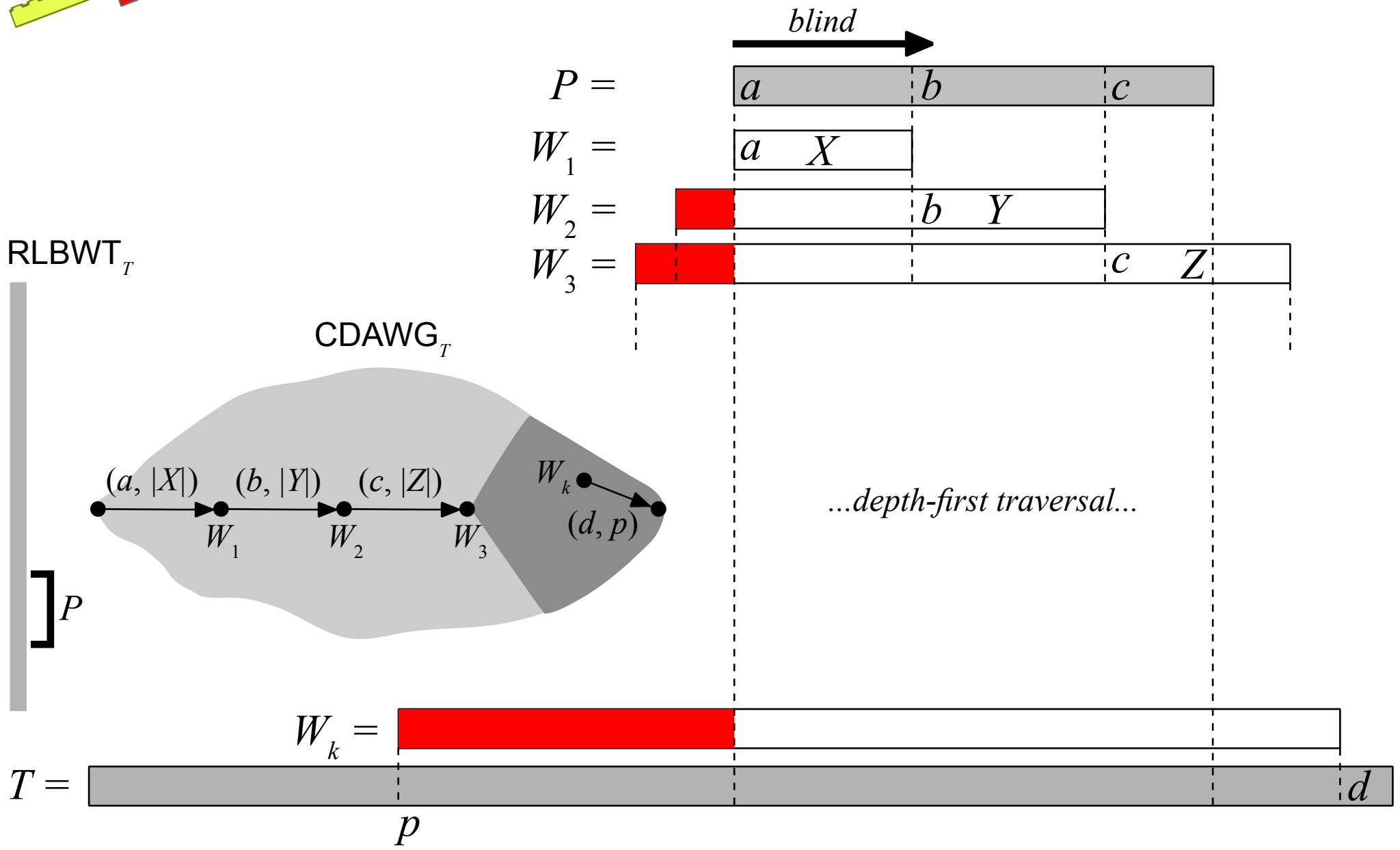
[1] Maxime Crochemore and Christophe Hancart. *Automata for matching patterns*. In Handbook of formal languages, pages 399–462. Springer, 1997.

Locating with RLBWT+CDAWG



[1] Maxime Crochemore and Christophe Hancart. *Automata for matching patterns*. In Handbook of formal languages, pages 399–462. Springer, 1997.

Locating with RLBWT+CDAWG



[1] Maxime Crochemore and Christophe Hancart. *Automata for matching patterns*. In Handbook of formal languages, pages 399–462. Springer, 1997.

Locating with RLBWT+CDAWG

Words:

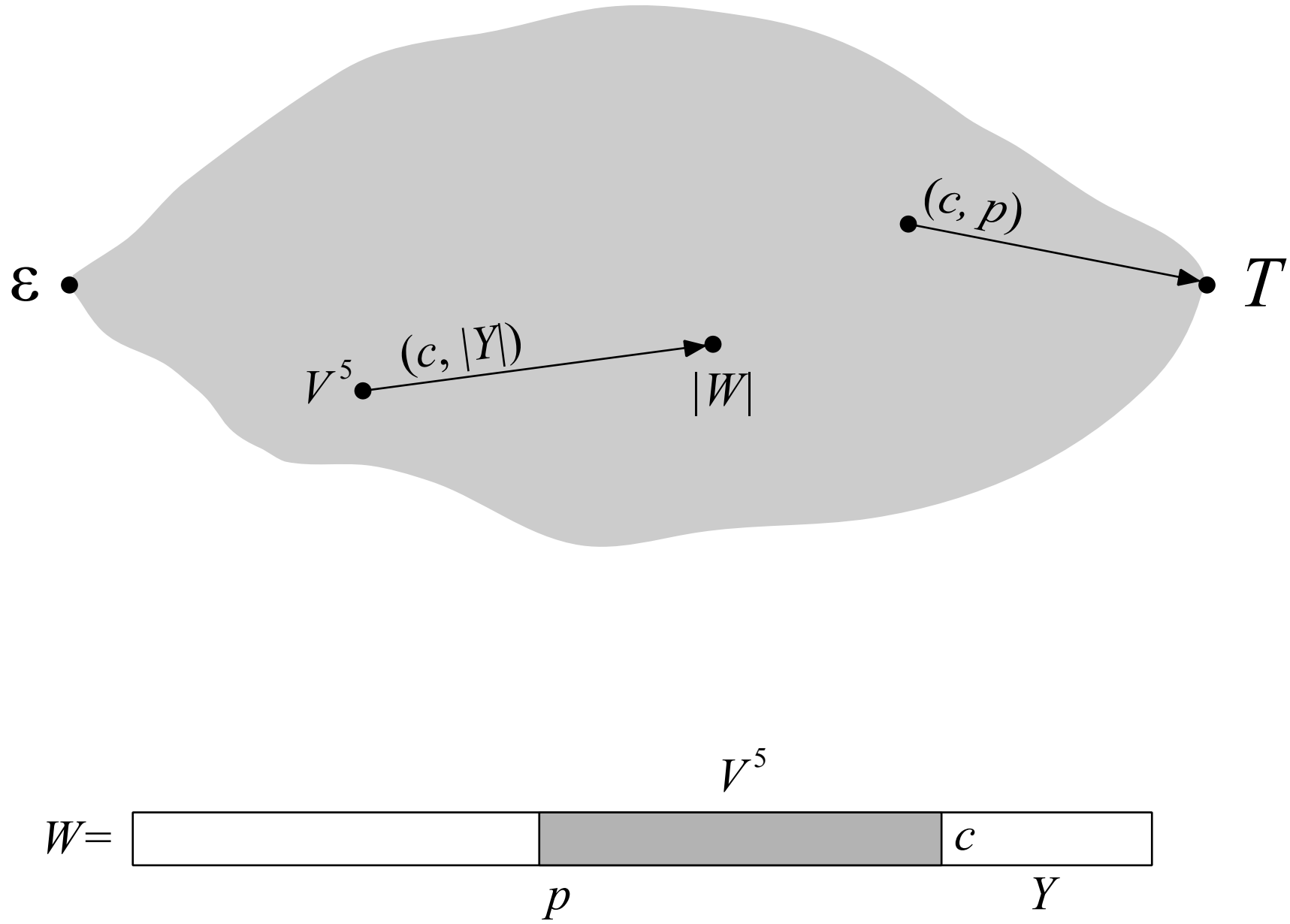
RLBWT+CDAWG	$O(\lvert\mathcal{E}_T^r \cup \mathcal{F}_T^r\rvert)$
RLBWT+LZ77	$O(\lvert\mathcal{Z}_T\rvert + \lvert\mathcal{R}_T\rvert + \lvert\mathcal{R}_{\overline{T}}\rvert)$
[1]	$O(n/k + \lvert\mathcal{R}_T\rvert)$

Time:

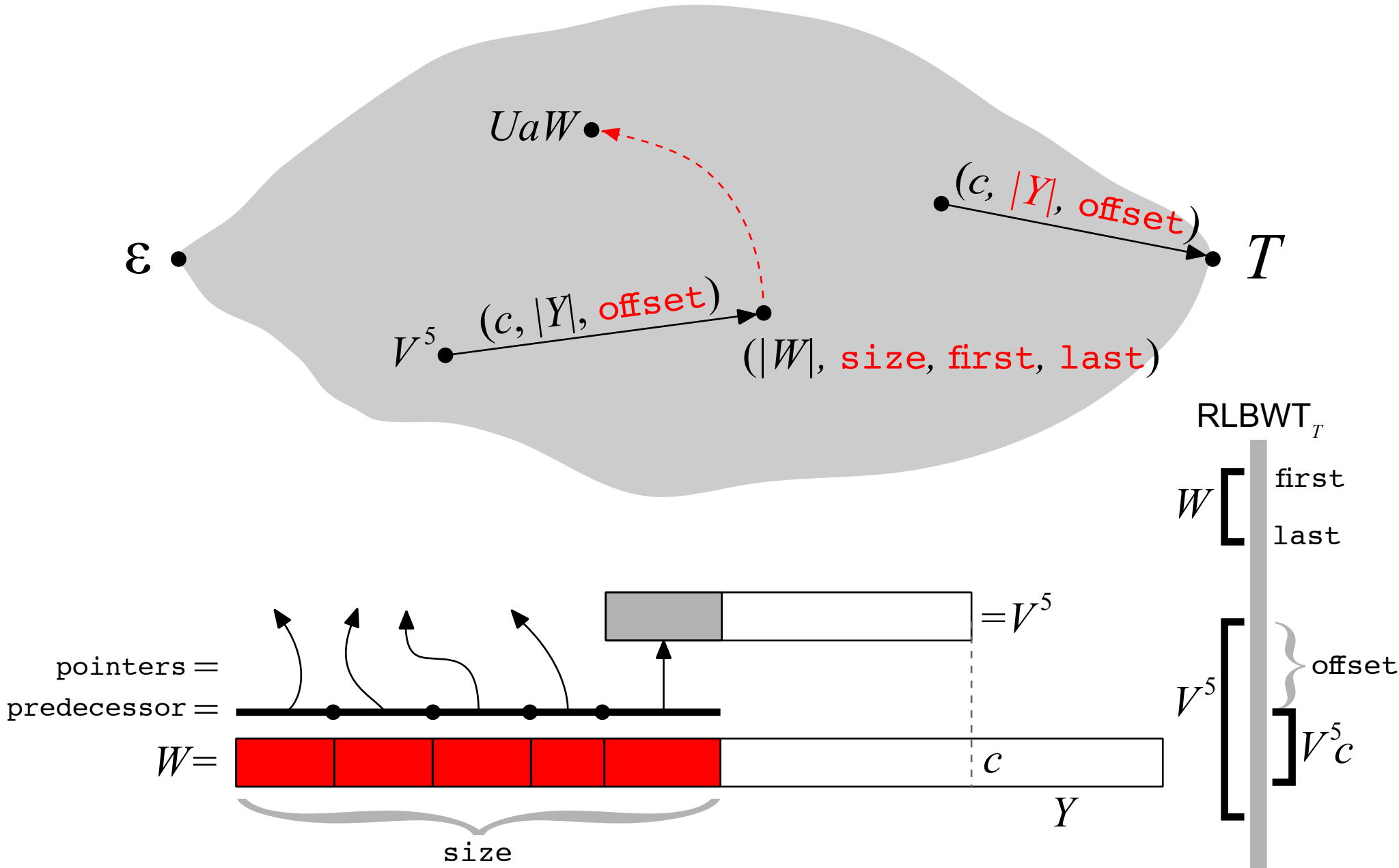
RLBWT+CDAWG	$O(m \log\log n + \mathtt{occ})$
RLBWT+LZ77	$O(m(\log\log n + \log\lvert\mathcal{Z}_T\rvert) + \mathtt{p0cc} \log^{\epsilon}\lvert\mathcal{Z}_T\rvert + \mathtt{s0cc} \log\log n)$
[2]	$O(m^2 h + (m + \mathtt{occ}) \log\lvert\mathcal{Z}_T\rvert)$
[1]	$O(m \log\log n + k \cdot \mathtt{occ} \log\log n)$

[1] Veli Mäkinen, Gonzalo Navarro, Jouni Sirén, and Niko Välimäki. *Storage and retrieval of highly repetitive sequence collections*. Journal of Computational Biology, 17(3):281–308, 2010.
[2] Sebastian Kreft and Gonzalo Navarro. *On compressing and indexing repetitive sequences*. Theoretical Computer Science, 483:115–133, 2013.

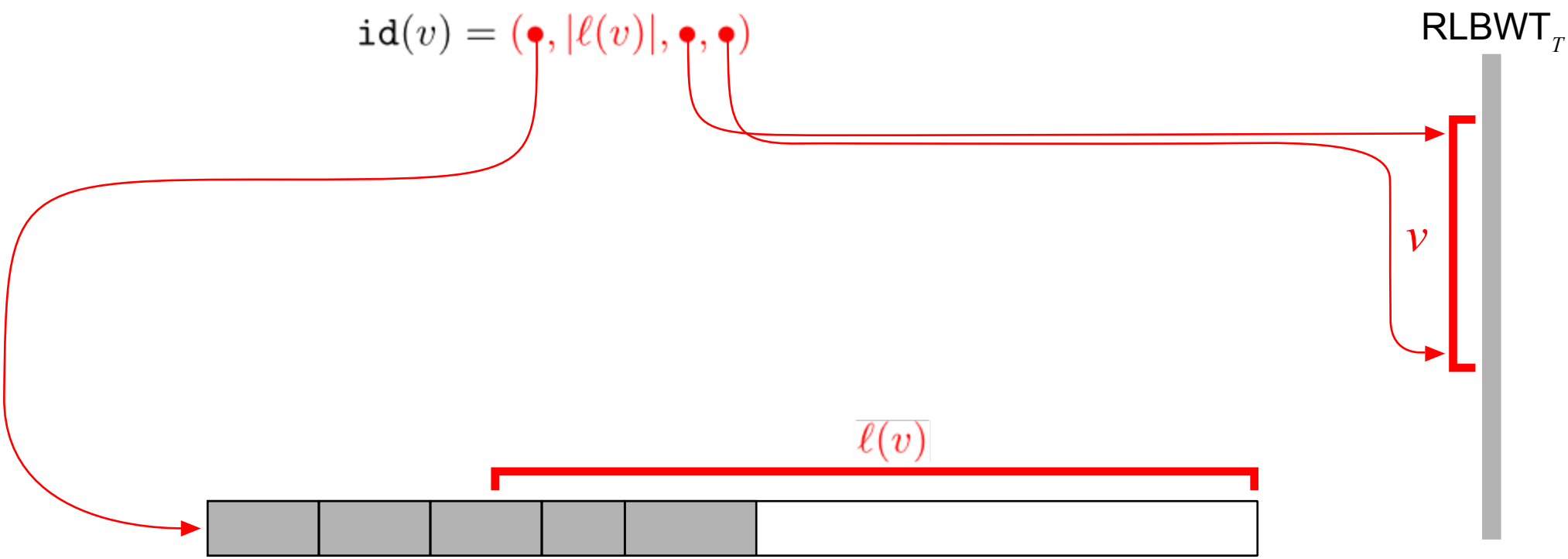
CDAWG for locating



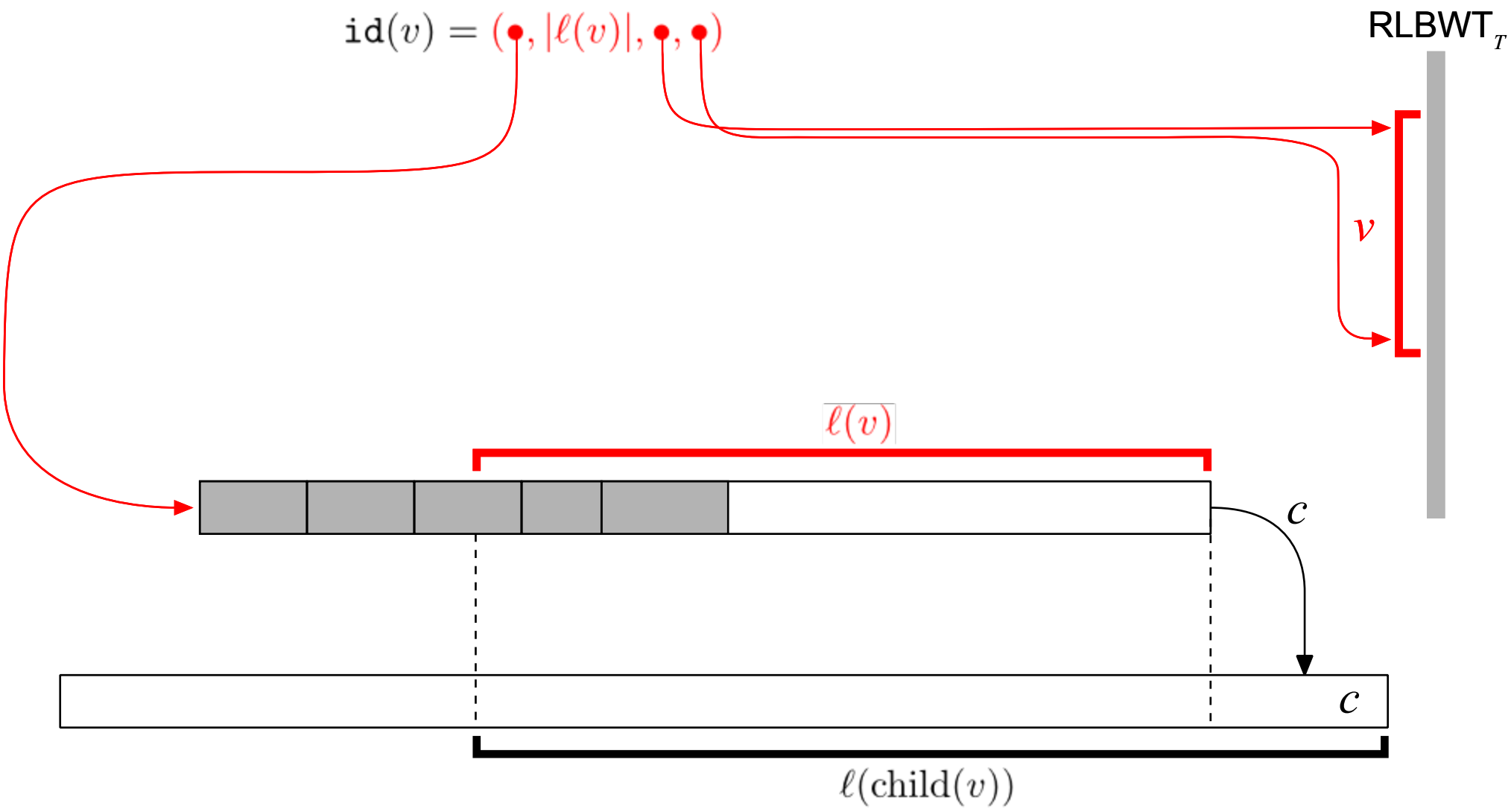
CDAWG for suffix tree



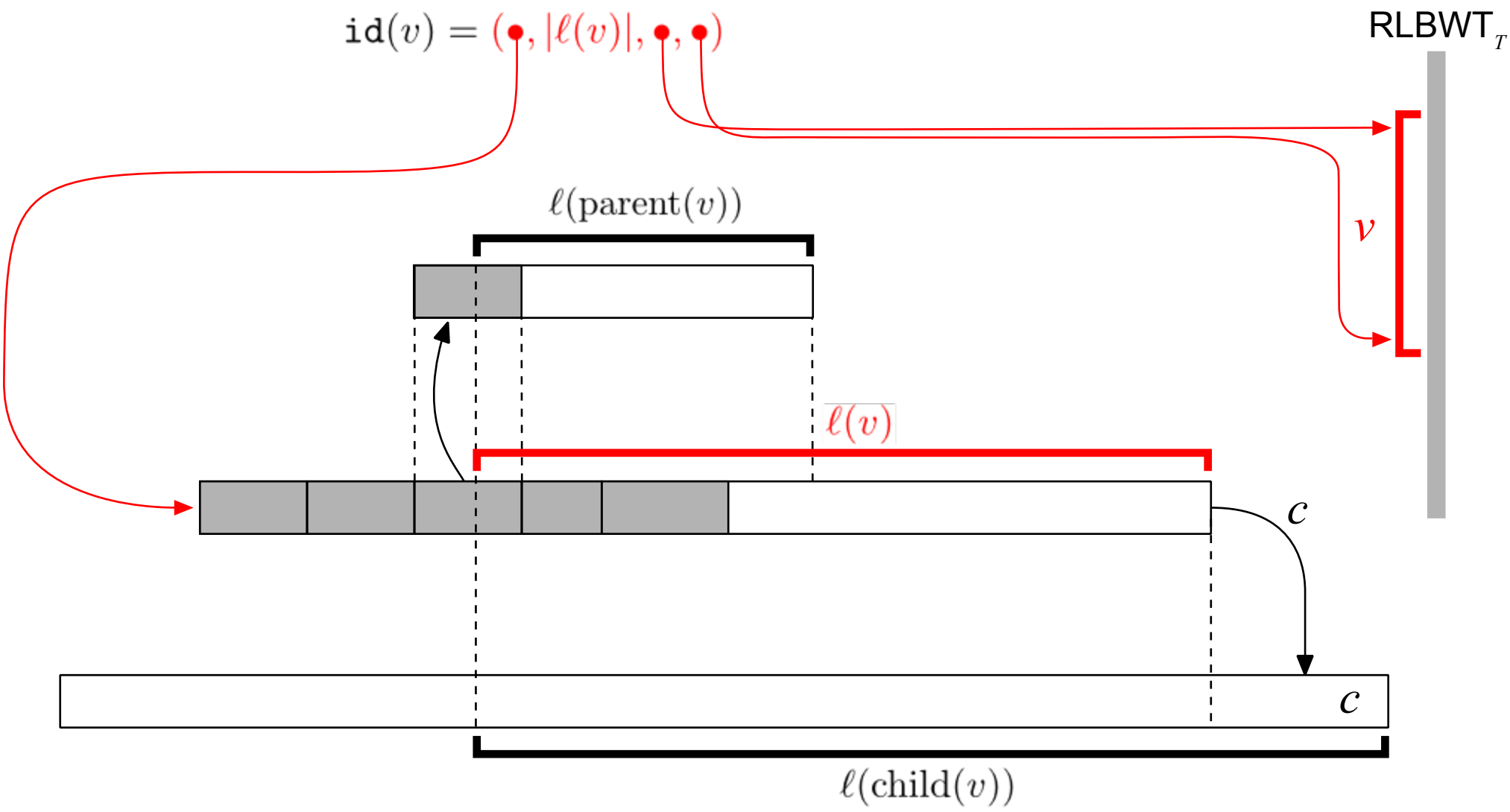
A node v of the suffix tree



A node v of the suffix tree



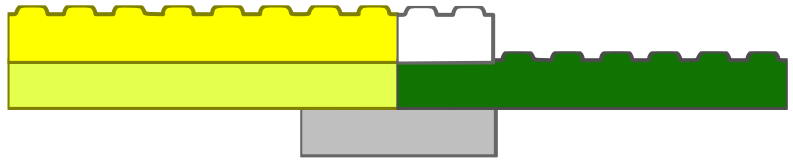
A node v of the suffix tree

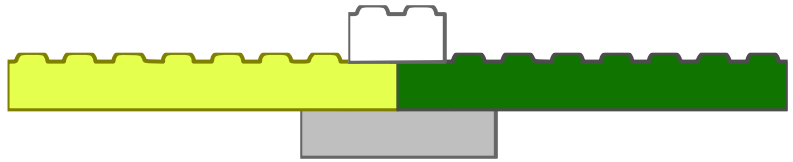


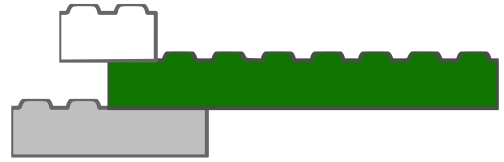
Suffix tree operations

Time:

	stringDepth locateLeaf	isAncestor	parent nextSibling	child firstChild	suffixLink	weinerLink	edgeChar	nLeaves
1	$O(1)$	$O(1)$	$O(\log\log n)$	$O(1)$	$O(\log\log n)$	$O(\log\log n)$	$O(\log\log n)$	$O(1)$
2	$O(1)$	$O(1)$	$O(\log\log n)$	$O(1)$	$O(\log\log n)$	$O(\log\log n)$		$O(1)$
3	$O(1)$		$O(\log\log n)$	$O(1)$	$O(1)$			

1)  $O(|\mathcal{E}_T^r| + |\mathcal{F}_T^r| + |\mathcal{E}_T^\ell| + |\mathcal{F}_T^\ell|)$

2)  $O(|\mathcal{E}_T^r| + |\mathcal{F}_T^r|)$

3)  $O(|\mathcal{E}_T^r| + |\mathcal{F}_T^r|)$

Suffix tree operations

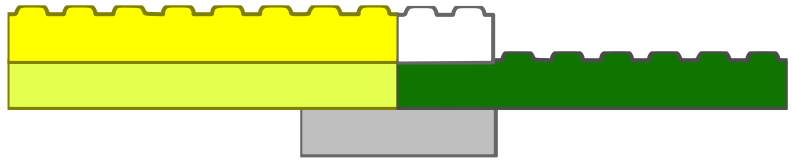
Time:

matching statistics

	stringDepth locateLeaf	isAncestor	parent nextSibling	child firstChild	suffixLink	weinerLink	edgeChar	nLeaves
1	$O(1)$	$O(1)$	$O(\log\log n)$	$O(1)$	$O(\log\log n)$	$O(\log\log n)$	$O(\log\log n)$	$O(1)$
2	$O(1)$	$O(1)$	$O(\log\log n)$	$O(1)$	$O(\log\log n)$	$O(\log\log n)$		$O(1)$
3	$O(1)$		$O(\log\log n)$	$O(1)$	$O(1)$			

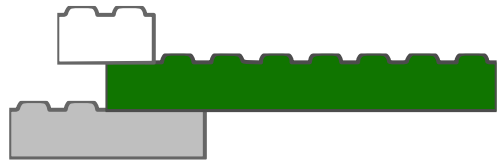
constant-space traversal

1)



$$O(|\mathcal{E}_T^r| + |\mathcal{F}_T^r| + |\mathcal{E}_T^\ell| + |\mathcal{F}_T^\ell|)$$

3)



$$O(|\mathcal{E}_T^r| + |\mathcal{F}_T^r|)$$

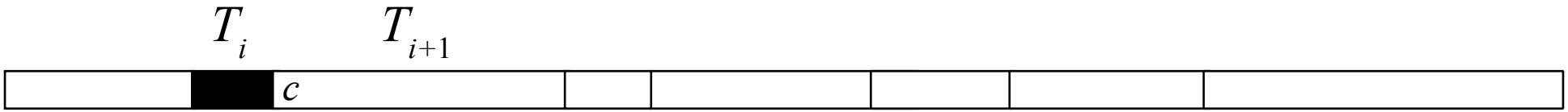
Composite repetition-aware data structures

Djamal Belazzougui[1], Fabio Cunial[2], Travis Gagie[1], Nicola Prezza[3], Mathieu Raffinot[4]

(1) Department of Computer Science, University of Helsinki, Finland.
(2) Max Planck Institute for Molecular Cell Biology and Genetics, Dresden, Germany.
(3) Department of Mathematics and Computer Science, University of Udine, Italy.
(4) LIAFA, Paris Diderot University - Paris 7, France.

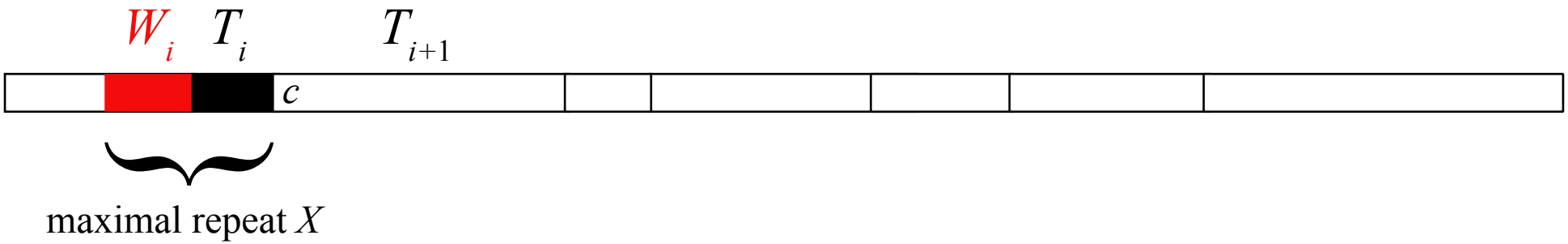
Rightmost maximal repeats and LZ factors

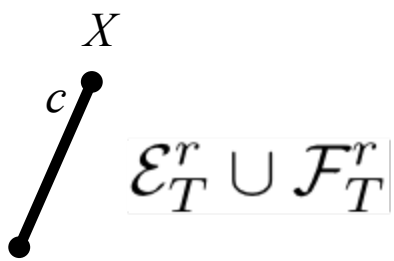
$$|\mathcal{Z}_T| \leq |\mathcal{E}_T^r \cup \mathcal{F}_T^r|$$



Rightmost maximal repeats and LZ factors

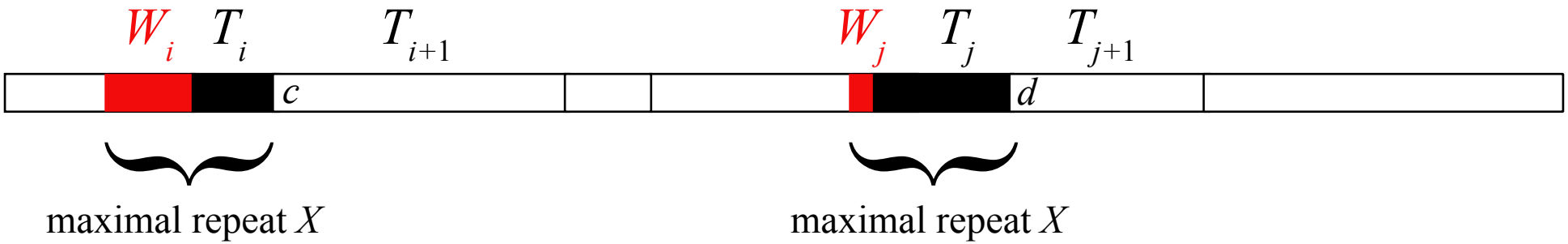
$$|\mathcal{Z}_T| \leq |\mathcal{E}_T^r \cup \mathcal{F}_T^r|$$





Rightmost maximal repeats and LZ factors

$$|\mathcal{Z}_T| \leq |\mathcal{E}_T^r \cup \mathcal{F}_T^r|$$



$\mathcal{E}_T^r \cup \mathcal{F}_T^r$

Measures of repetition