String Matching with Variable Length Gaps

By Philip Bille, Inge Li Gørtz, Hjalte Wedel Vildhøj and David Kofoed Wind

Presented by Hjalte Wedel Vildhøj

October 13, 2010

SPIRE 2010, Los Cabos, Mexico

DTU Informatics

Department of Informatics and Mathematical Modelling

The Variable Length Gap Problem

Given some string $T \in \Sigma^+$ and a *variable length gap pattern*

$$P = P_1 \cdot \ g\{a_1, b_1\} \cdot P_2 \cdot g\{a_2, b_2\} \cdots g\{a_{k-1}, b_{k-1}\} \cdot P_k \ .$$

Find the *end positions* for all occurrences of P in T.

The Variable Length Gap Problem

Given some string $T \in \Sigma^+$ and a *variable length gap pattern*

$$P = P_1 \cdot \underbrace{g\{a_1, b_1\}}_{\boxed{\text{Some } x \in \Sigma^* \text{ s.t. } a_1 \leq |x| \leq b_1}} \cdot P_2 \cdot g\{a_2, b_2\} \cdots g\{a_{k-1}, b_{k-1}\} \cdot P_k \ .$$

Find the *end positions* for all occurrences of P in T.

The Variable Length Gap Problem

Given some string $T \in \Sigma^+$ and a *variable length gap pattern*

$$P = P_1 \cdot \underbrace{g\{a_1, b_1\}}_{\boxed{\text{Some } x \in \Sigma^* \text{ s.t. } a_1 \leq |x| \leq b_1}} \cdot P_2 \cdot g\{a_2, b_2\} \cdots g\{a_{k-1}, b_{k-1}\} \cdot P_k \ .$$

Find the *end positions* for all occurrences of P in T.

Example: $P = \mathrm{A} \cdot g\{6, 7\} \cdot \mathrm{CC} \cdot g\{2, 6\} \cdot \mathrm{GT}$

$T =$ ATCGGCTCCAGACCAGTACCCGTTCCGTGGT

Solution: $\{\}$

The Variable Length Gap Problem

Given some string $T \in \Sigma^+$ and a *variable length gap pattern*

$$P = P_1 \cdot \underbrace{g\{a_1, b_1\}}_{\boxed{\text{Some } x \in \Sigma^* \text{ s.t. } a_1 \le |x| \le b_1}} \cdot P_2 \cdot g\{a_2, b_2\} \cdots g\{a_{k-1}, b_{k-1}\} \cdot P_k \ .$$

Find the *end positions* for all occurrences of P in T.

Example: $P = \mathrm{A} \cdot g\{6, 7\} \cdot \mathrm{CC} \cdot g\{2, 6\} \cdot \mathrm{GT}$

$$T = \mathrm{A}\overbrace{\mathrm{TCGGCT}}^{6}\mathrm{CC}\overbrace{\mathrm{AGACCA}}^{6}\mathrm{GT}\mathrm{ACCCGTTCCGTGGT}$$

end pos in T

Solution: $\{17\}$

The Variable Length Gap Problem

Given some string $T \in \Sigma^+$ and a *variable length gap pattern*

$$P = P_1 \cdot \underbrace{g\{a_1, b_1\}}_{\boxed{\text{Some } x \in \Sigma^* \text{ s.t. } a_1 \leq |x| \leq b_1}} \cdot P_2 \cdot g\{a_2, b_2\} \cdots g\{a_{k-1}, b_{k-1}\} \cdot P_k \ .$$

Find the *end positions* for all occurrences of P in T.

Example: $P = \mathrm{A} \cdot g\{6, 7\} \cdot \mathrm{CC} \cdot g\{2, 6\} \cdot \mathrm{GT}$

$$T = \mathrm{ATCGGCTCCA}\overbrace{\mathrm{GACCAGTA}}^{8}\mathrm{CC}\overbrace{\mathrm{CGTTCC}}^{6}\mathrm{GTGGT}$$

Solution: $\{17\}$

Not a valid match!

The Variable Length Gap Problem

Given some string $T \in \Sigma^+$ and a *variable length gap pattern*

$$P = P_1 \cdot \underbrace{g\{a_1, b_1\}}_{\boxed{\text{Some } x \in \Sigma^* \text{ s.t. } a_1 \leq |x| \leq b_1}} \cdot P_2 \cdot g\{a_2, b_2\} \cdots g\{a_{k-1}, b_{k-1}\} \cdot P_k \ .$$

Find the *end positions* for all occurrences of P in T.

Example: $P = \mathrm{A} \cdot g\{6, 7\} \cdot \mathrm{CC} \cdot g\{2, 6\} \cdot \mathrm{GT}$

$$T = \ \mathrm{ATCGGCTCCAGA}\,\mathrm{A}\overbrace{\mathrm{CCAGTA}}^{6}\mathrm{CC}\overbrace{\mathrm{CGTTCC}}^{6}\mathrm{GT}\,\mathrm{GGT}$$

end pos in T

Solution: $\{17, 28\}$

The Variable Length Gap Problem

Given some string $T \in \Sigma^+$ and a *variable length gap pattern*

$$P = P_1 \cdot \underbrace{g\{a_1, b_1\}}_{\boxed{\text{Some } x \in \Sigma^* \text{ s.t. } a_1 \leq |x| \leq b_1}} \cdot P_2 \cdot g\{a_2, b_2\} \cdots g\{a_{k-1}, b_{k-1}\} \cdot P_k \ .$$

Find the *end positions* for all occurrences of P in T.

Example: $P = \mathrm{A} \cdot g\{6,7\} \cdot \mathrm{CC} \cdot g\{2,6\} \cdot \mathrm{GT}$

$$T = \mathrm{ATCGGCTCCAGA}\,\mathrm{A}\overbrace{\mathrm{CCAGTAC}}^{7}\mathrm{CC}\overbrace{\mathrm{GTTCC}}^{5}\mathrm{GT}\,\mathrm{GGT}$$

end pos in T

Solution: $\{17, 28\}$

The Variable Length Gap Problem

Given some string $T \in \Sigma^+$ and a *variable length gap pattern*

$$P = P_1 \cdot \underbrace{g\{a_1, b_1\}}_{\boxed{\text{Some } x \in \Sigma^* \text{ s.t. } a_1 \le |x| \le b_1}} \cdot P_2 \cdot g\{a_2, b_2\} \cdots g\{a_{k-1}, b_{k-1}\} \cdot P_k \; .$$

Find the *end positions* for all occurrences of P in T.

Example: $P = \mathrm{A} \cdot g\{6, 7\} \cdot \mathrm{CC} \cdot g\{2, 6\} \cdot \mathrm{GT}$

$$T = \text{ATCGGCTCCAGACCAGTA}\overbrace{\text{CCCGTT}}^{7}\text{CC}\overbrace{\text{GTG}}^{3}\text{GT}$$

end pos in T

Solution: $\left\{17, 28, 31\right\}$

The Variable Length Gap Problem

Given some string $T \in \Sigma^+$ and a *variable length gap pattern*

$$P = P_1 \cdot \underbrace{g\{a_1, b_1\}}_{\boxed{\text{Some } x \in \Sigma^* \text{ s.t. } a_1 \le |x| \le b_1}} \cdot P_2 \cdot g\{a_2, b_2\} \cdots g\{a_{k-1}, b_{k-1}\} \cdot P_k \ .$$

Find the *end positions* for all occurrences of P in T.

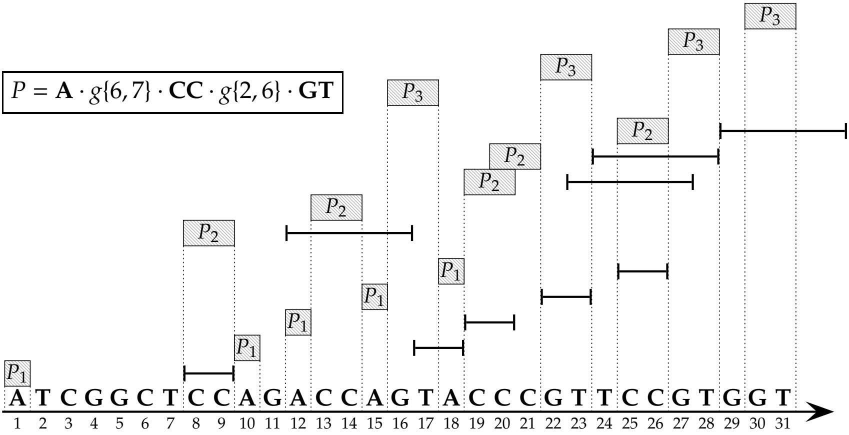
Example: $P = \mathrm{A} \cdot g\{6,7\} \cdot \mathrm{CC} \cdot g\{2,6\} \cdot \mathrm{GT}$

$T =$ ATCGGCTCCAGACCAGTACCCGTTCCGTGGT

Solution: $\{17, 28, 31\}$

A Closer Look At The Problem

$$P = \mathbf{A} \cdot g\{6,7\} \cdot \mathbf{CC} \cdot g\{2,6\} \cdot \mathbf{GT}$$



A Closer Look At The Problem

Parameters

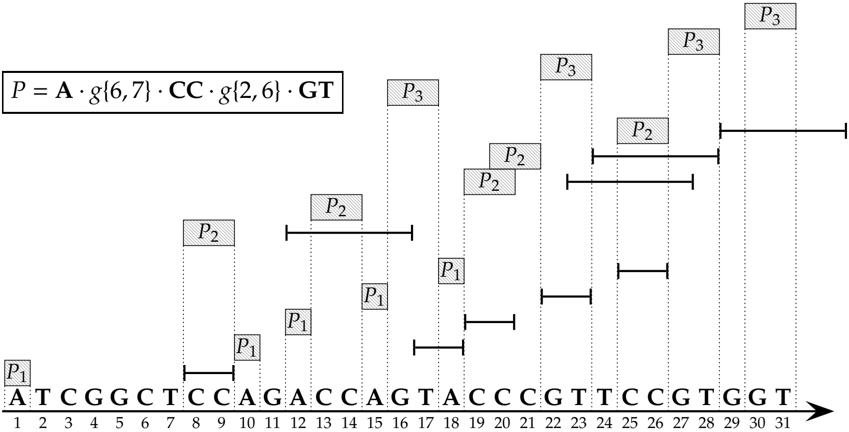
$n = |T|$

$\alpha = \#\text{ occ. of } P_1, P_2, \ldots, P_k \text{ in } T$

$$m = \sum_{i=1}^{k} |P_i|$$

$$A = \sum_{i=1}^{k} a_i$$

$\boxed{P = \mathbf{A} \cdot g\{6,7\} \cdot \mathbf{CC} \cdot g\{2,6\} \cdot \mathbf{GT}}$



A Closer Look At The Problem

Parameters

$n = |T|$

$\alpha = \#$ occ. of $P_1, P_2, \ldots, P_k$ in T

$$m = \sum_{i=1}^{k} |P_i|$$

$$A = \sum_{i=1}^{k} a_i$$

Known Upper Bounds

By	*Time*	*Space*
Bille & Thorup[1]	$O\left(n(k\frac{\log w}{w} + \log k) + m \log m + A\right)$	$O(m + A)$
Morgante et al.[2]	$O((n + m) \log k + \alpha)$	$O(m + \alpha)$

[1] P. Bille and M. Thorup. Regular expression matching with multi-strings and intervals. In *Proc. 21st SODA, 2010*

[2] M. Morgante, A. Policriti, N. Vitacolonna, and A. Zuccolo. Structured motifs search. *J. Comput. Bio.*, 12(8):1065-1082, 2005

A Closer Look At The Problem

Parameters

$n = |T|$

$\alpha = \#$ occ. of $P_1, P_2, \ldots, P_k$ in T

$$m = \sum_{i=1}^{k} |P_i|$$

$$A = \sum_{i=1}^{k} a_i$$

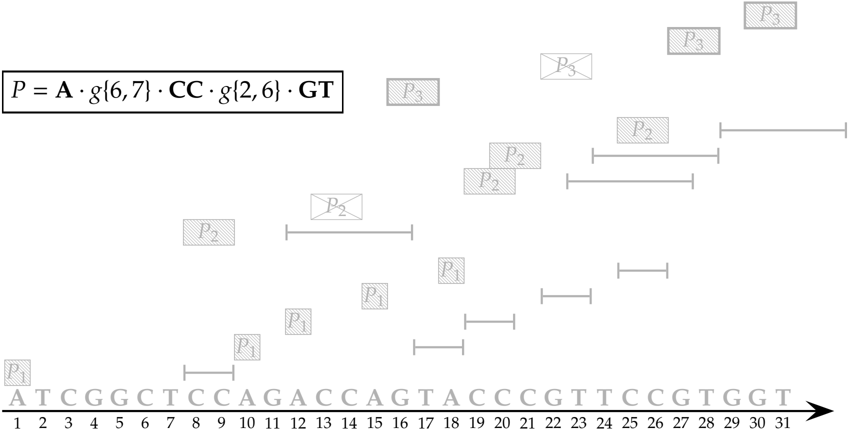
Known Upper Bounds

By	*Time*	*Space*
Bille & Thorup	$O\left(n(k\frac{\log w}{w} + \log k) + m \log m + A\right)$	$O(m + A)$
Morgante et al.	$O((n + m) \log k + \alpha)$	$O(m + \alpha)$

Can you get the best of both?

Illustrating the Algorithm

$$\boxed{P = \mathbf{A} \cdot g\{6,7\} \cdot \mathbf{CC} \cdot g\{2,6\} \cdot \mathbf{GT}}$$

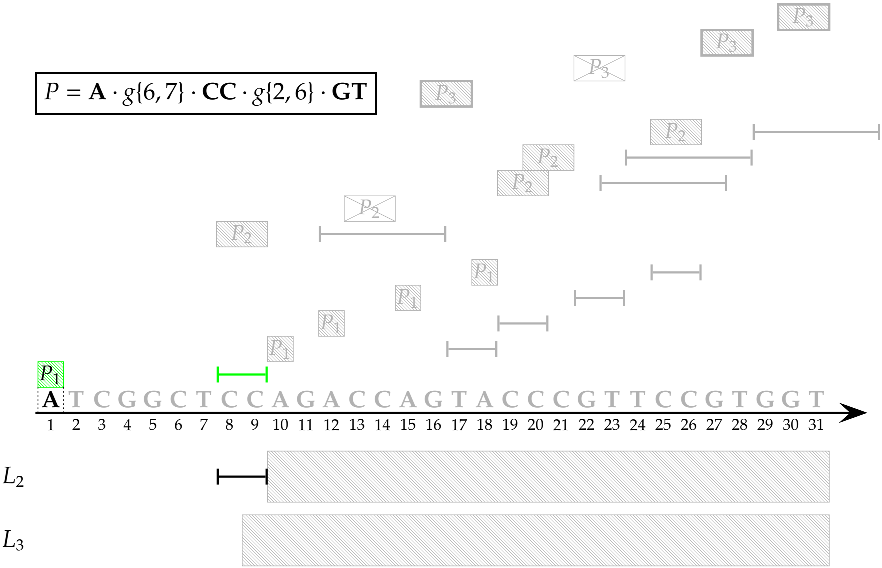


L_2

L_3

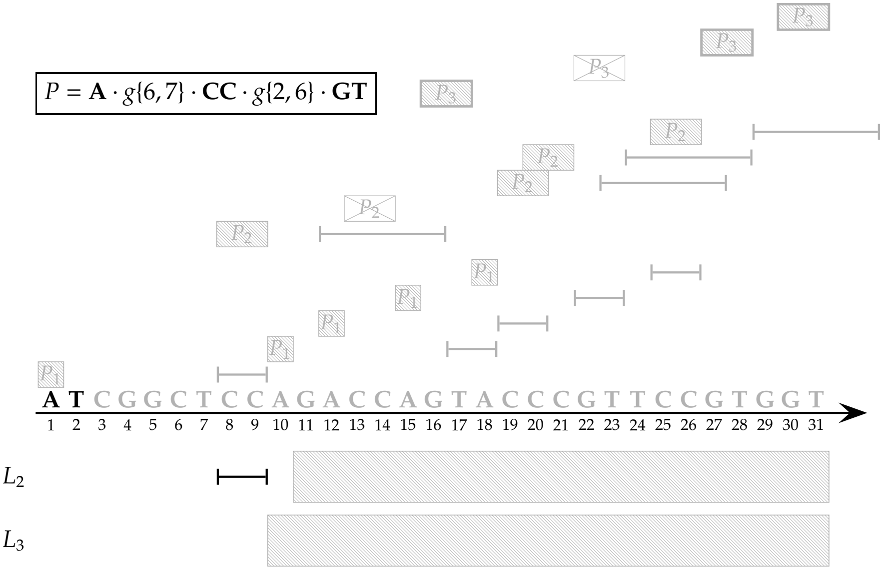
Illustrating the Algorithm

$$\boxed{P = \mathbf{A} \cdot g\{6,7\} \cdot \mathbf{CC} \cdot g\{2,6\} \cdot \mathbf{GT}}$$



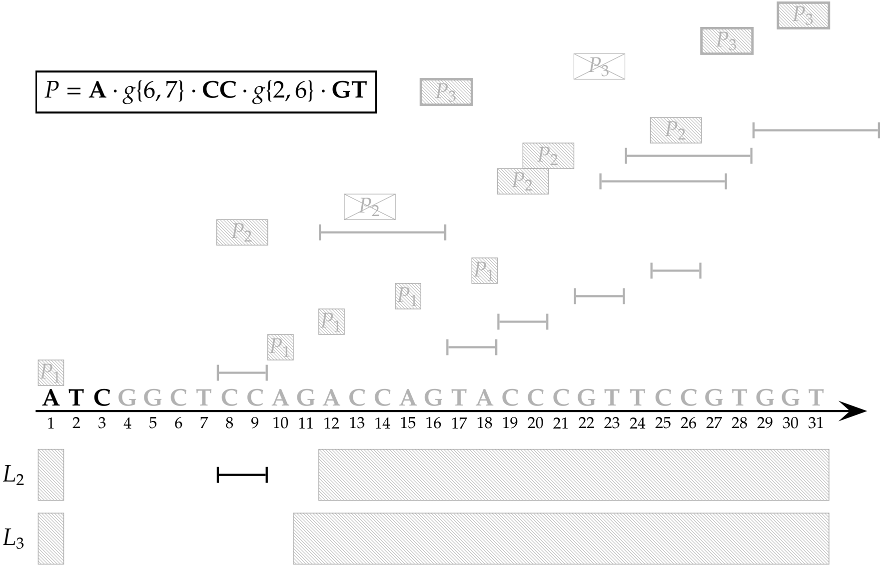
Illustrating the Algorithm

$$\boxed{P = \mathbf{A} \cdot g\{6,7\} \cdot \mathbf{CC} \cdot g\{2,6\} \cdot \mathbf{GT}}$$



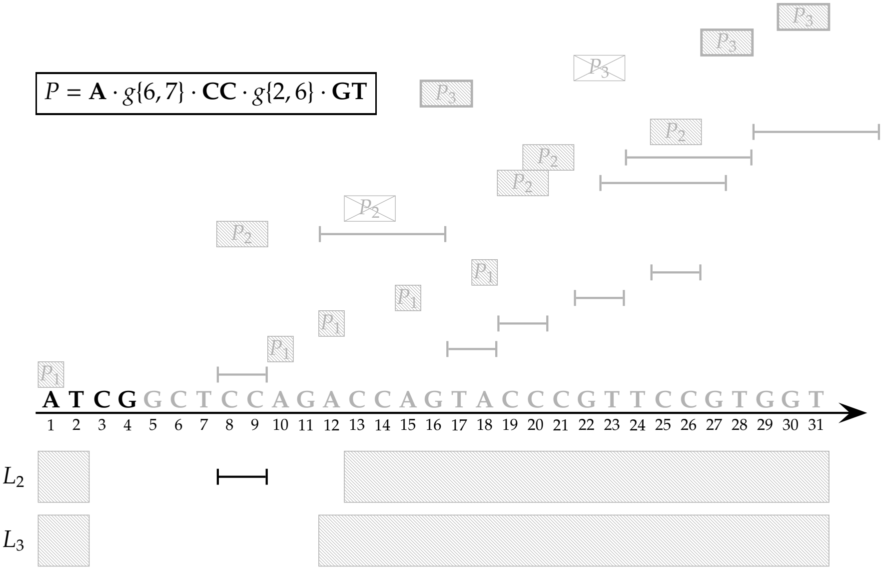
Illustrating the Algorithm

$$\boxed{P = \mathbf{A} \cdot g\{6,7\} \cdot \mathbf{CC} \cdot g\{2,6\} \cdot \mathbf{GT}}$$



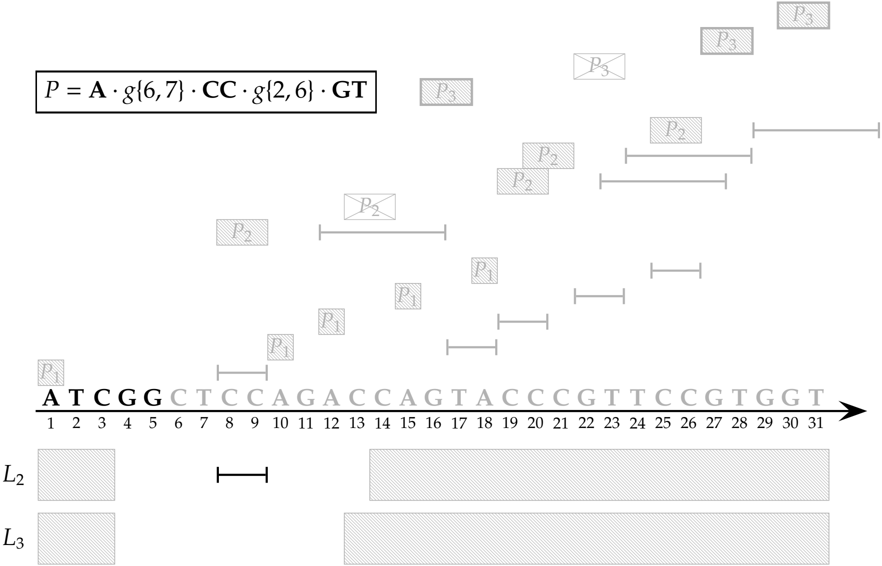
Illustrating the Algorithm

$$\boxed{P = \mathbf{A} \cdot g\{6,7\} \cdot \mathbf{CC} \cdot g\{2,6\} \cdot \mathbf{GT}}$$



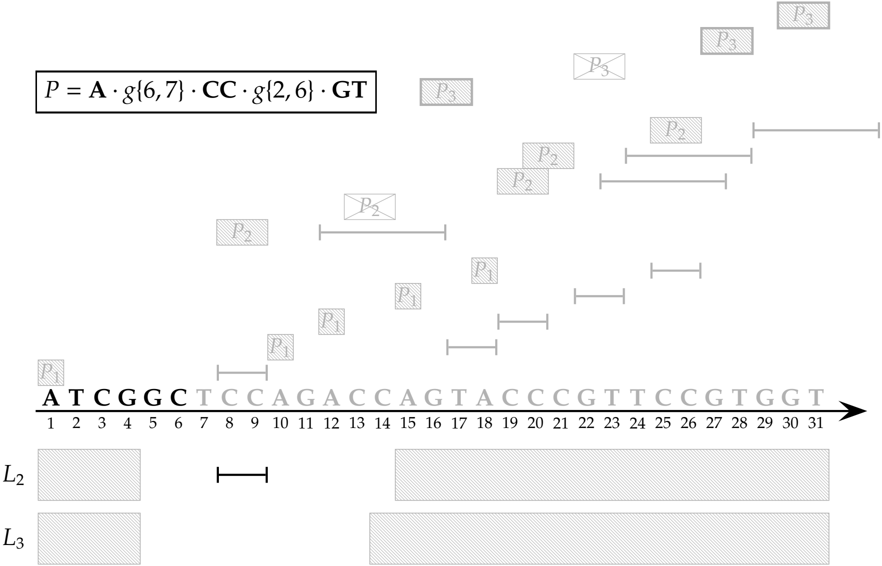
Illustrating the Algorithm

$$\boxed{P = \mathbf{A} \cdot g\{6,7\} \cdot \mathbf{CC} \cdot g\{2,6\} \cdot \mathbf{GT}}$$



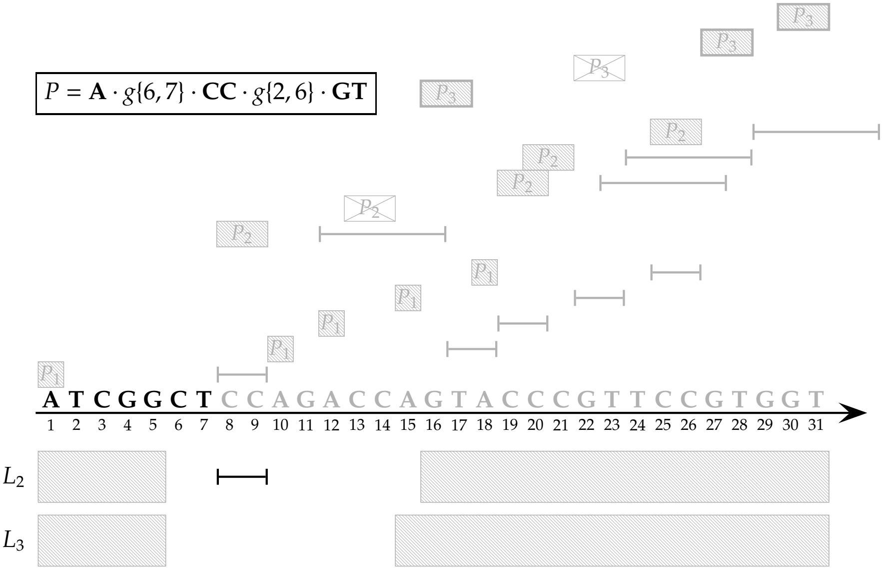
Illustrating the Algorithm

$P = \mathbf{A} \cdot g\{6,7\} \cdot \mathbf{CC} \cdot g\{2,6\} \cdot \mathbf{GT}$



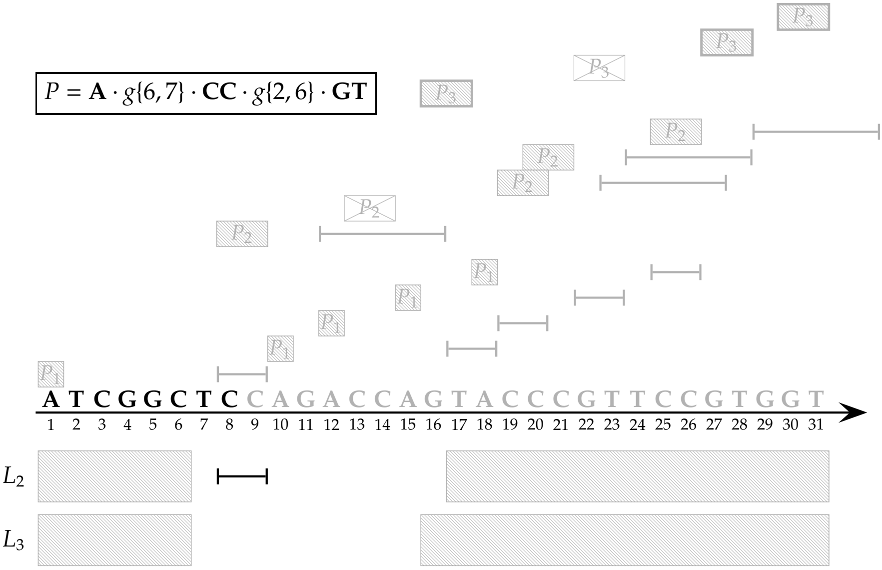
Illustrating the Algorithm

$P = \mathbf{A} \cdot g\{6,7\} \cdot \mathbf{CC} \cdot g\{2,6\} \cdot \mathbf{GT}$



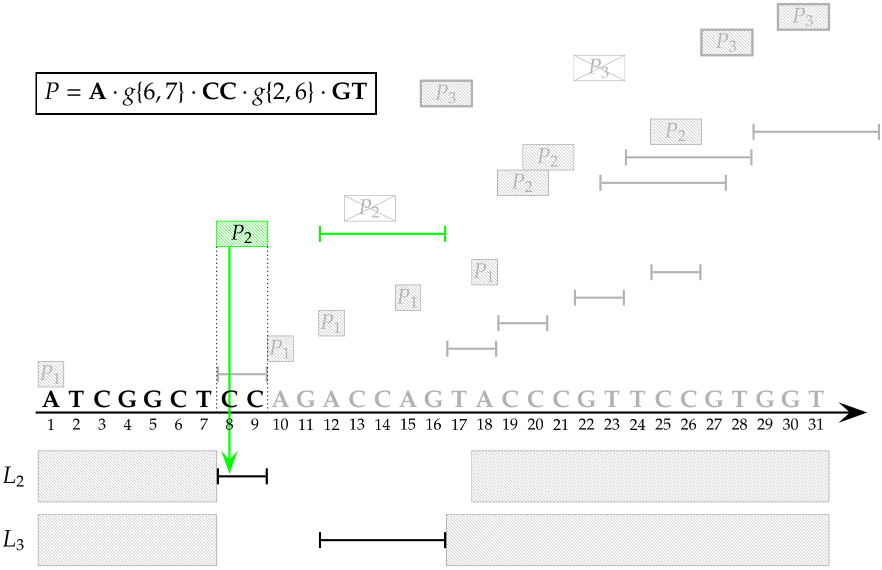
Illustrating the Algorithm

$$\boxed{P = \mathbf{A} \cdot g\{6,7\} \cdot \mathbf{CC} \cdot g\{2,6\} \cdot \mathbf{GT}}$$



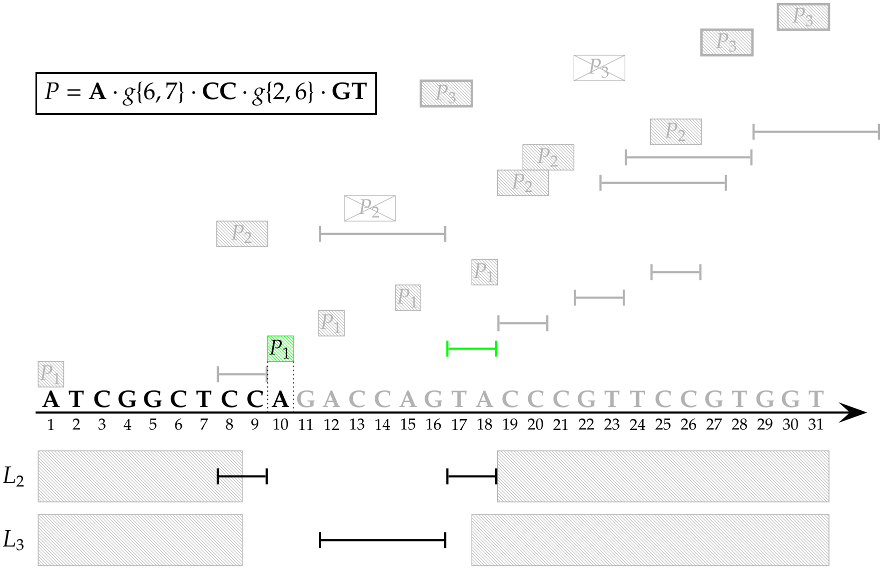
Illustrating the Algorithm

$$\boxed{P = \mathbf{A} \cdot g\{6,7\} \cdot \mathbf{CC} \cdot g\{2,6\} \cdot \mathbf{GT}}$$



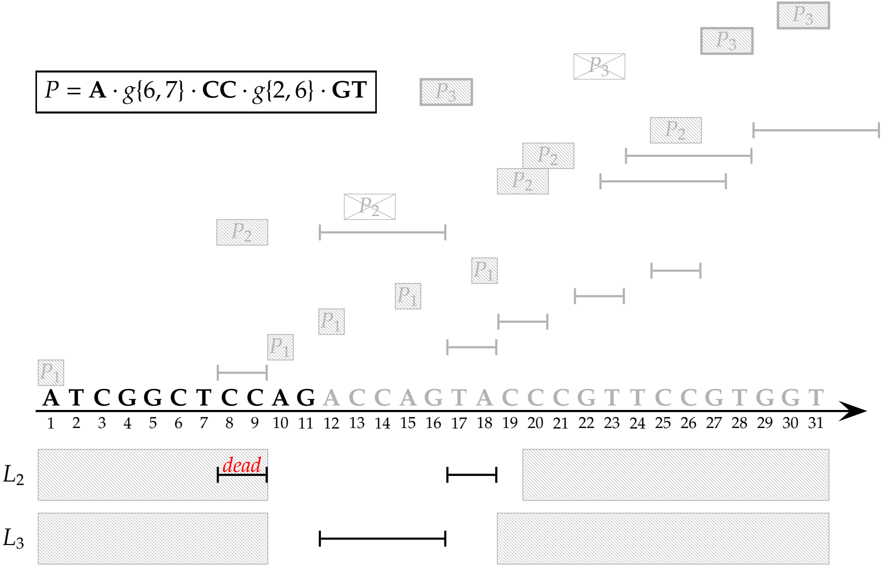
Illustrating the Algorithm

$$\boxed{P = \mathbf{A} \cdot g\{6,7\} \cdot \mathbf{CC} \cdot g\{2,6\} \cdot \mathbf{GT}}$$



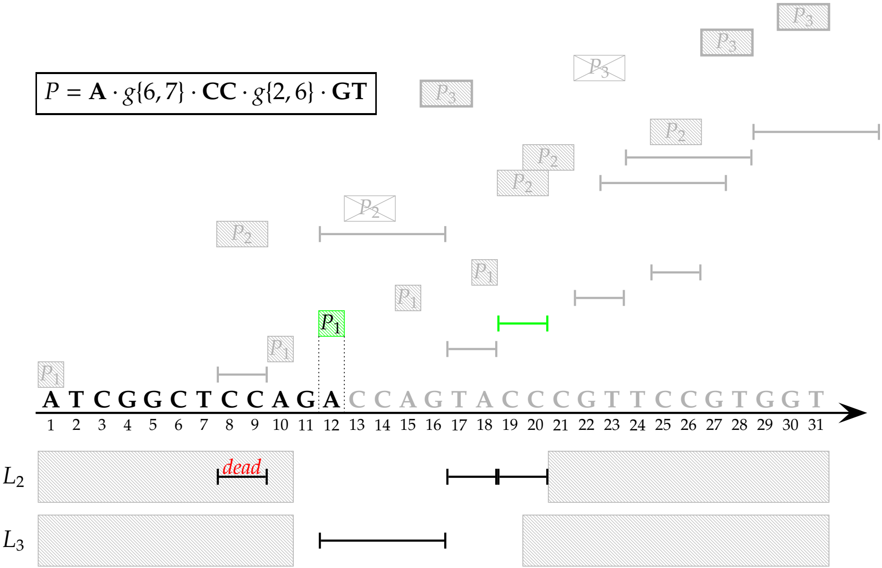
Illustrating the Algorithm

$$P = \mathbf{A} \cdot g\{6,7\} \cdot \mathbf{CC} \cdot g\{2,6\} \cdot \mathbf{GT}$$



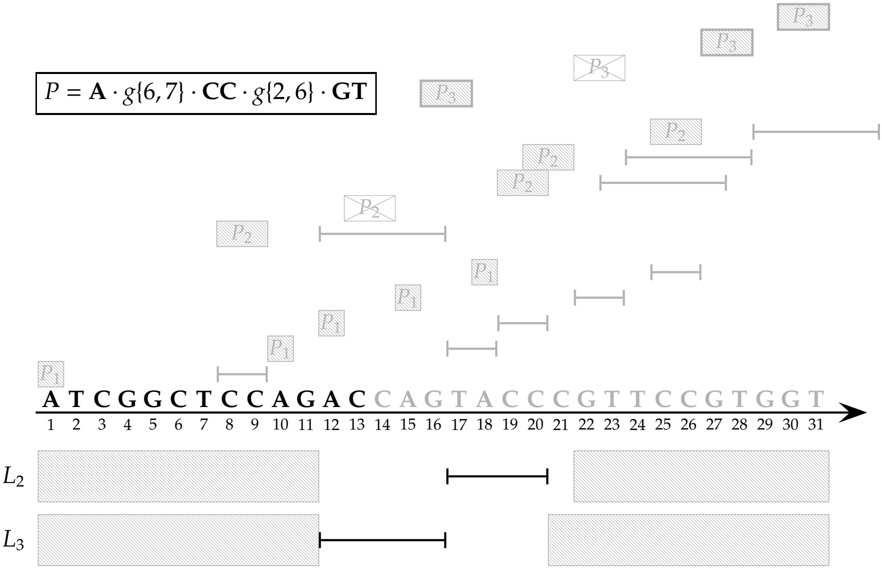
Illustrating the Algorithm

$P = \mathbf{A} \cdot g\{6,7\} \cdot \mathbf{CC} \cdot g\{2,6\} \cdot \mathbf{GT}$



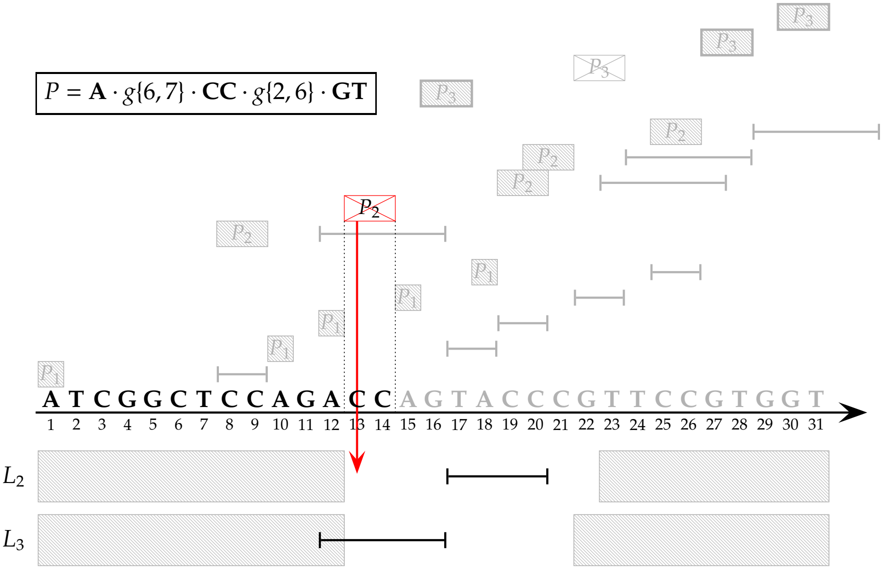
Illustrating the Algorithm

$$P = \mathbf{A} \cdot g\{6,7\} \cdot \mathbf{CC} \cdot g\{2,6\} \cdot \mathbf{GT}$$

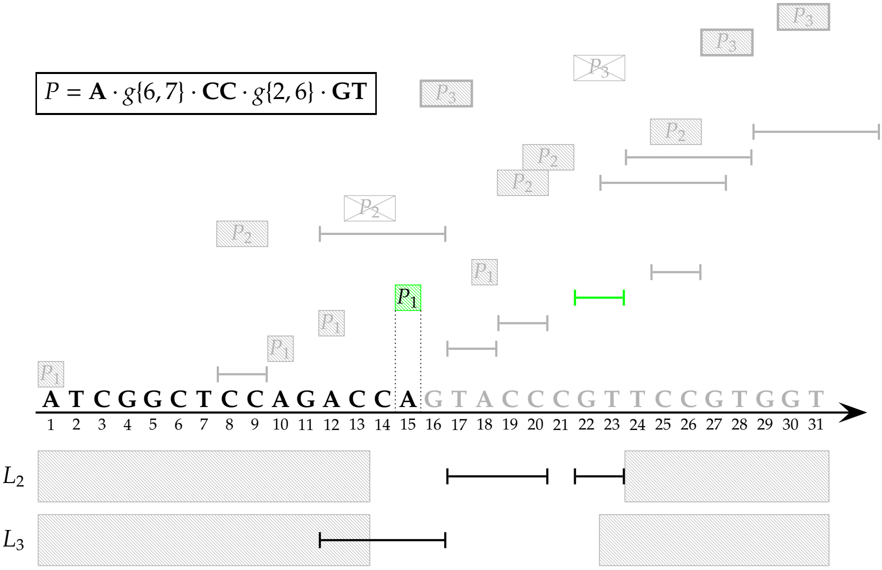


Illustrating the Algorithm

$$\boxed{P = \mathbf{A} \cdot g\{6,7\} \cdot \mathbf{CC} \cdot g\{2,6\} \cdot \mathbf{GT}}$$

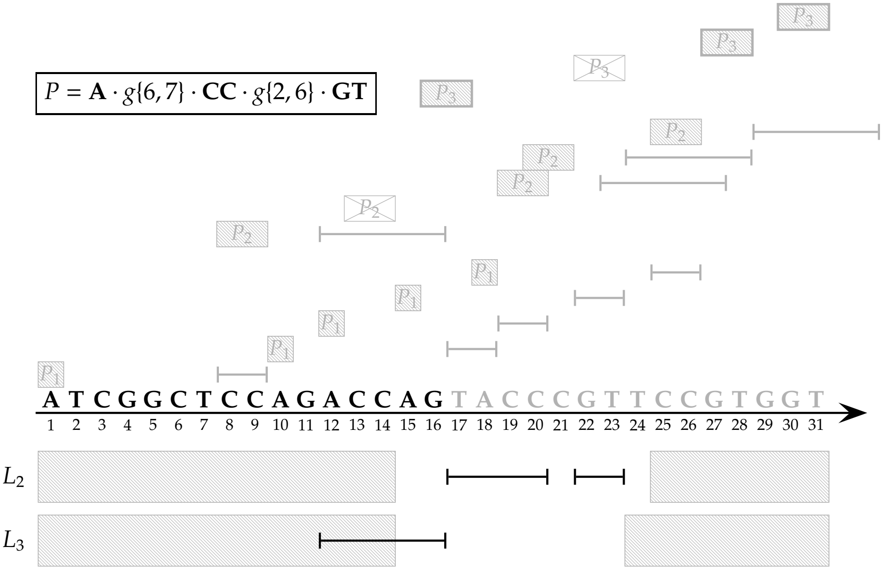


Illustrating the Algorithm



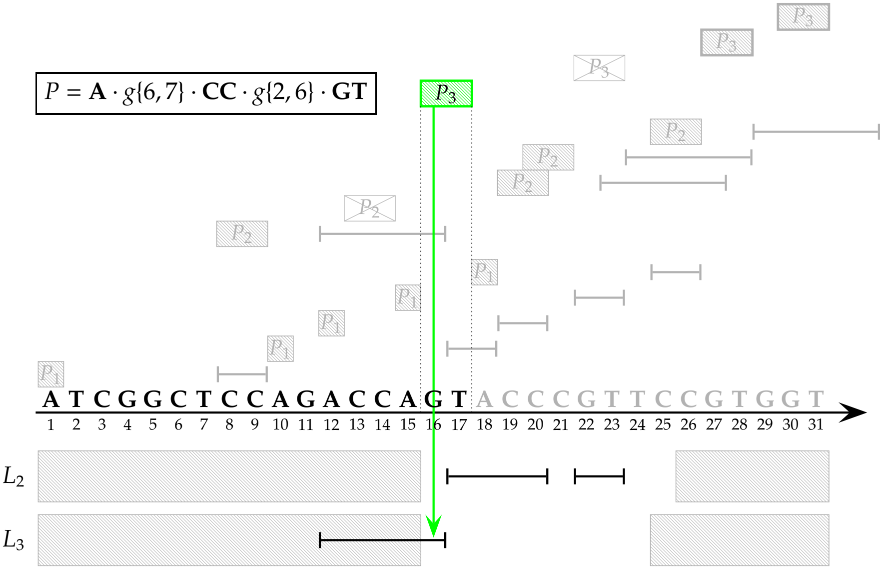
Illustrating the Algorithm

$$\boxed{P = \mathbf{A} \cdot g\{6,7\} \cdot \mathbf{CC} \cdot g\{2,6\} \cdot \mathbf{GT}}$$



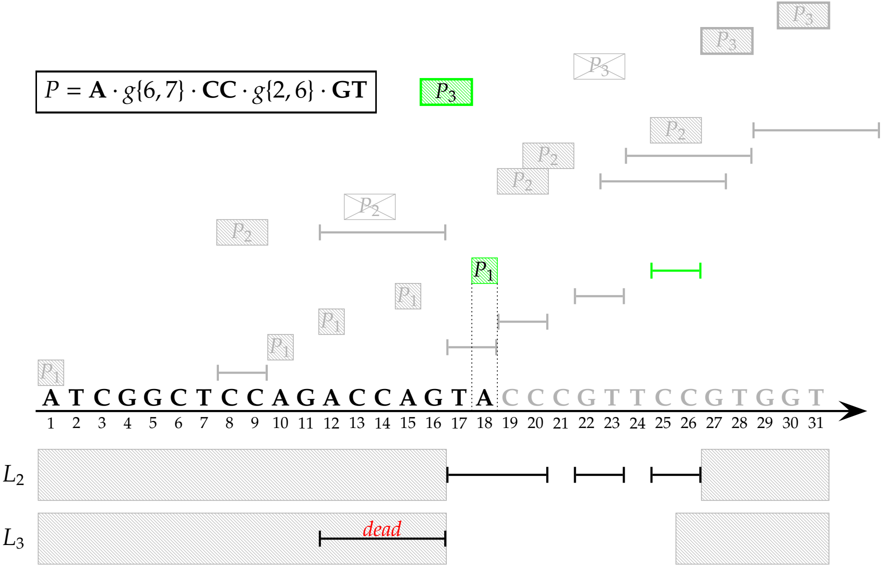
Illustrating the Algorithm

$P = \mathbf{A} \cdot g\{6,7\} \cdot \mathbf{CC} \cdot g\{2,6\} \cdot \mathbf{GT}$



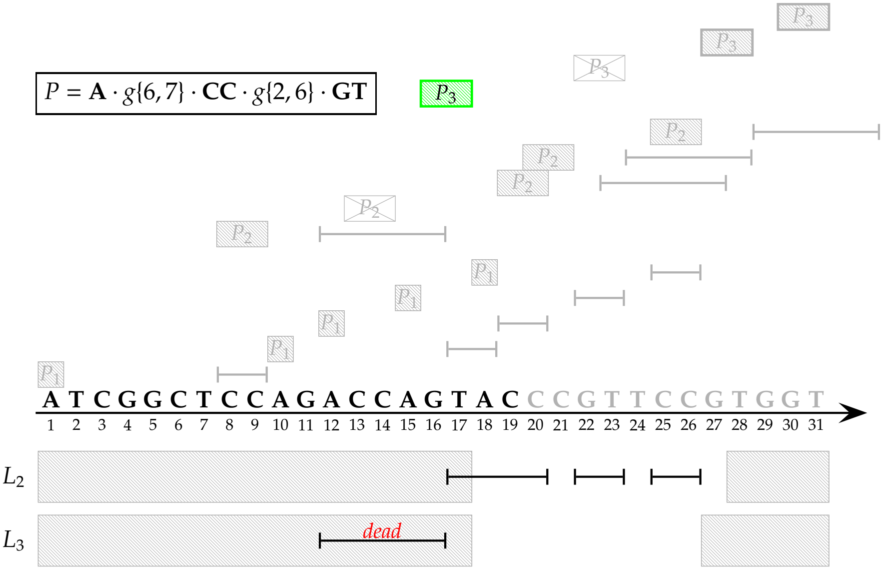
Illustrating the Algorithm

$$P = \mathbf{A} \cdot g\{6,7\} \cdot \mathbf{CC} \cdot g\{2,6\} \cdot \mathbf{GT}$$



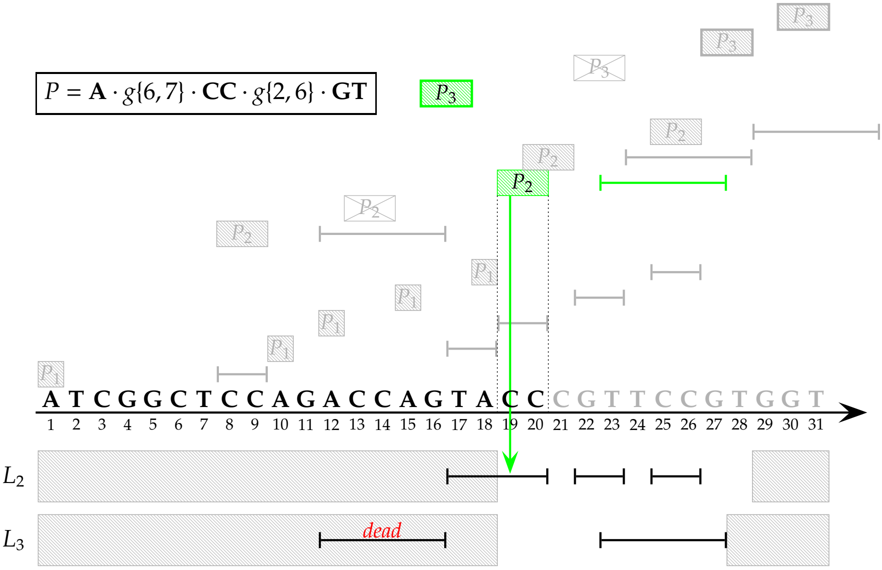
Illustrating the Algorithm

$P = \mathbf{A} \cdot g\{6,7\} \cdot \mathbf{CC} \cdot g\{2,6\} \cdot \mathbf{GT}$

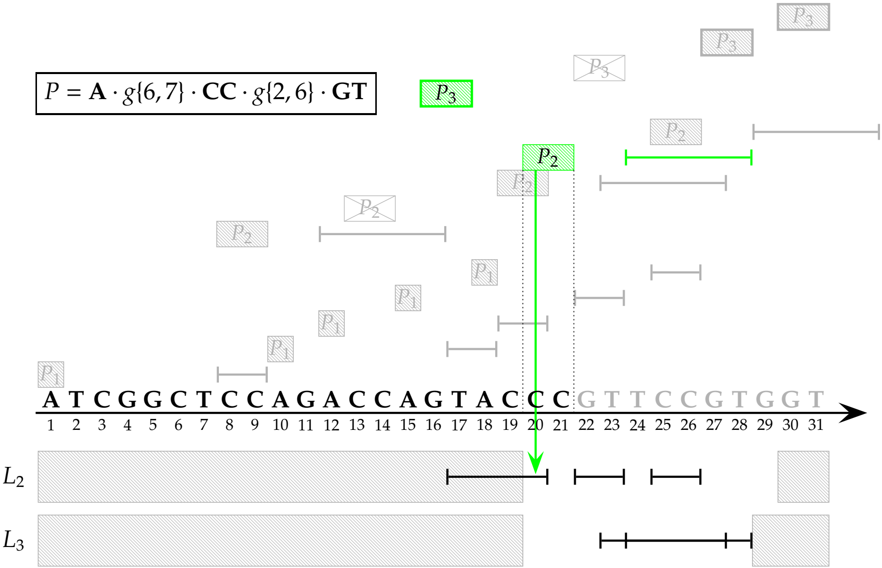


Illustrating the Algorithm

$$P = \mathbf{A} \cdot g\{6,7\} \cdot \mathbf{CC} \cdot g\{2,6\} \cdot \mathbf{GT}$$

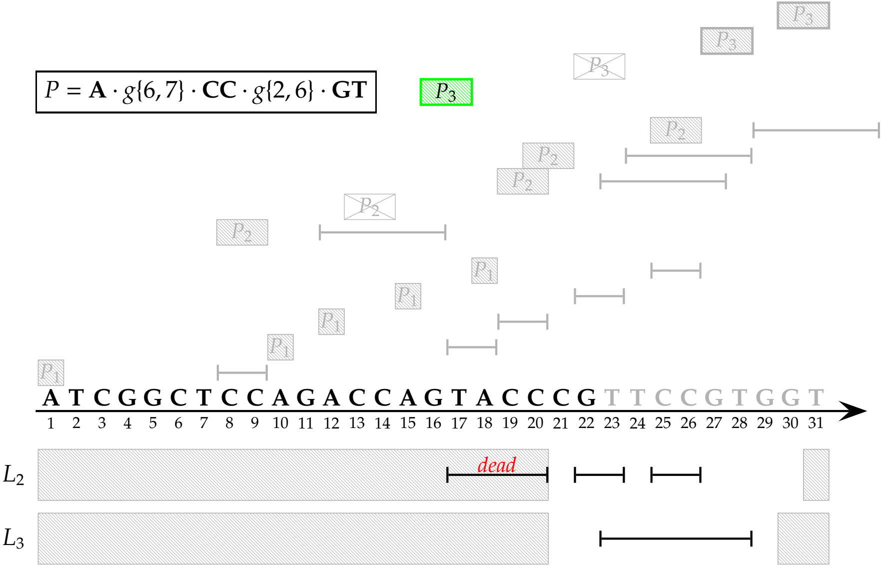


Illustrating the Algorithm

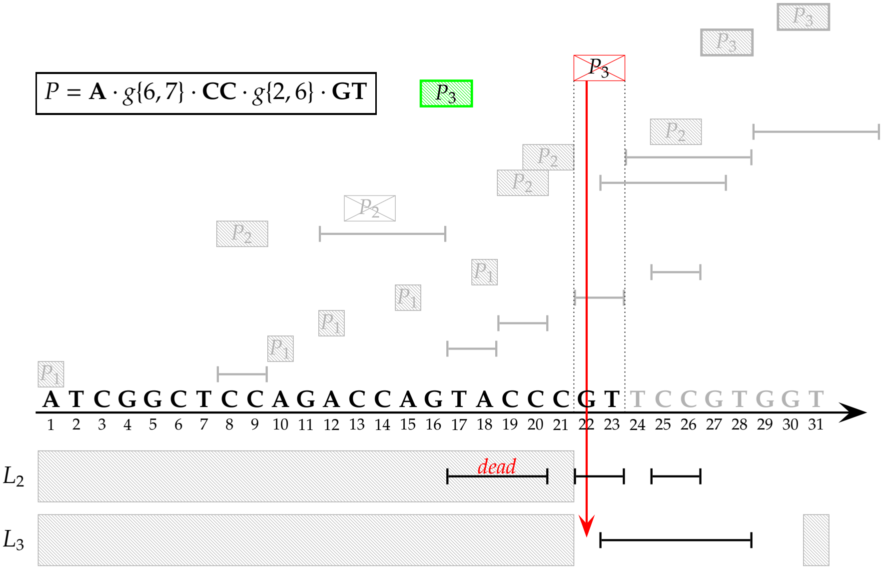


Illustrating the Algorithm

$\boxed{P = \mathbf{A} \cdot g\{6,7\} \cdot \mathbf{CC} \cdot g\{2,6\} \cdot \mathbf{GT}}$

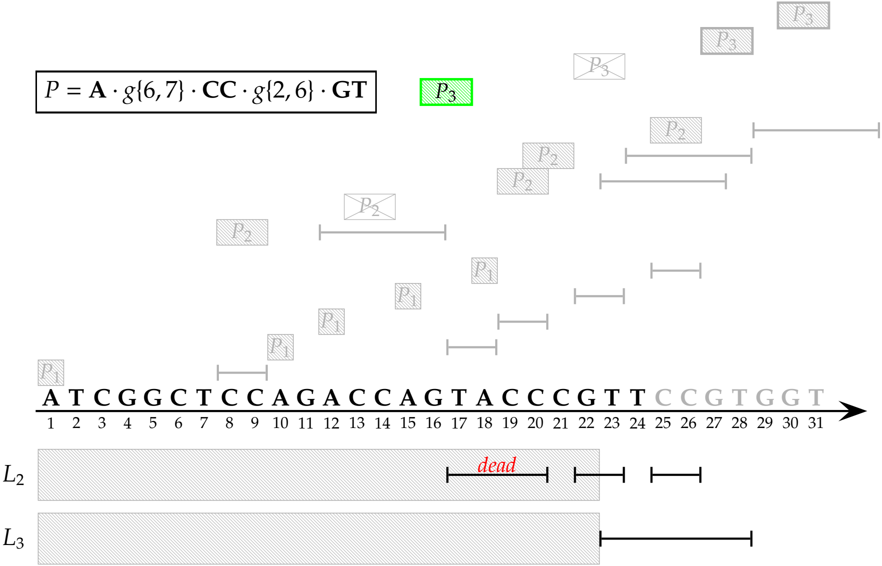


Illustrating the Algorithm



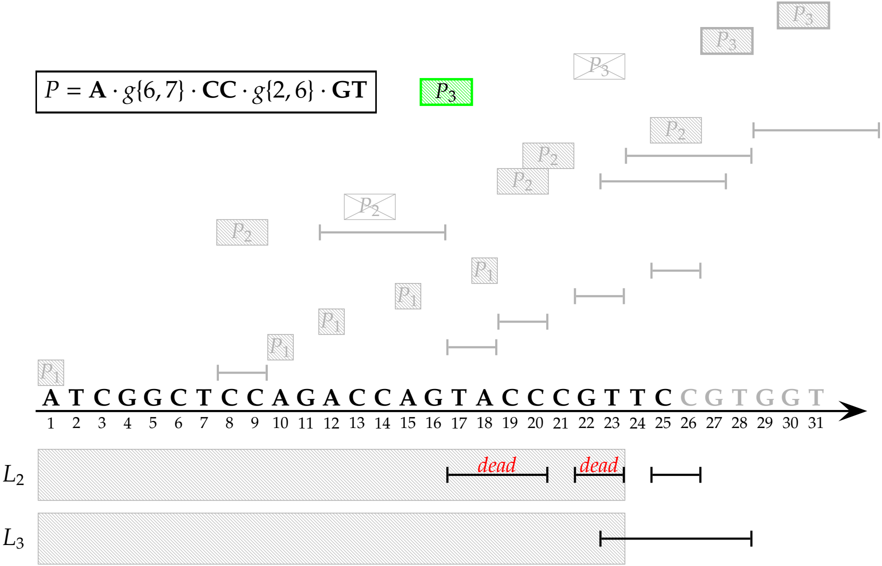
Illustrating the Algorithm

$P = \mathbf{A} \cdot g\{6,7\} \cdot \mathbf{CC} \cdot g\{2,6\} \cdot \mathbf{GT}$



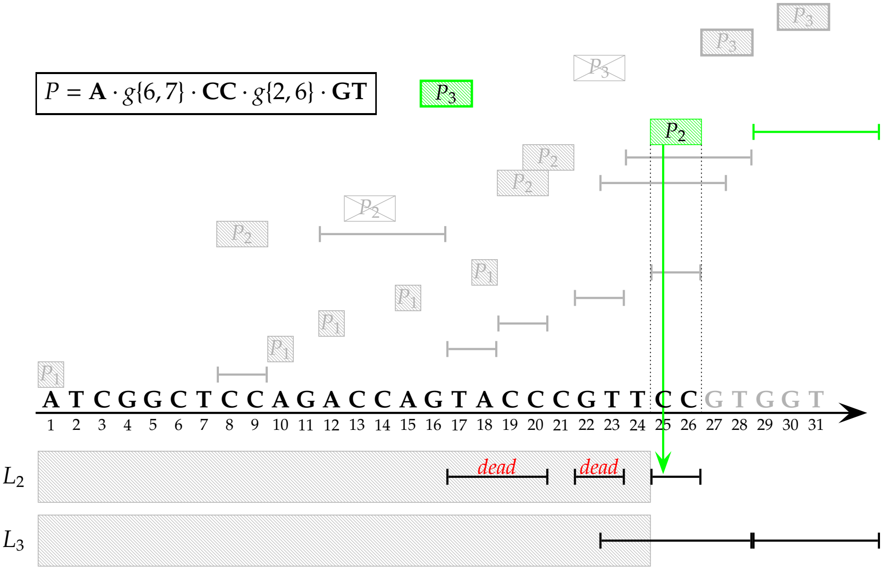
Illustrating the Algorithm

$$P = \mathbf{A} \cdot g\{6,7\} \cdot \mathbf{CC} \cdot g\{2,6\} \cdot \mathbf{GT}$$



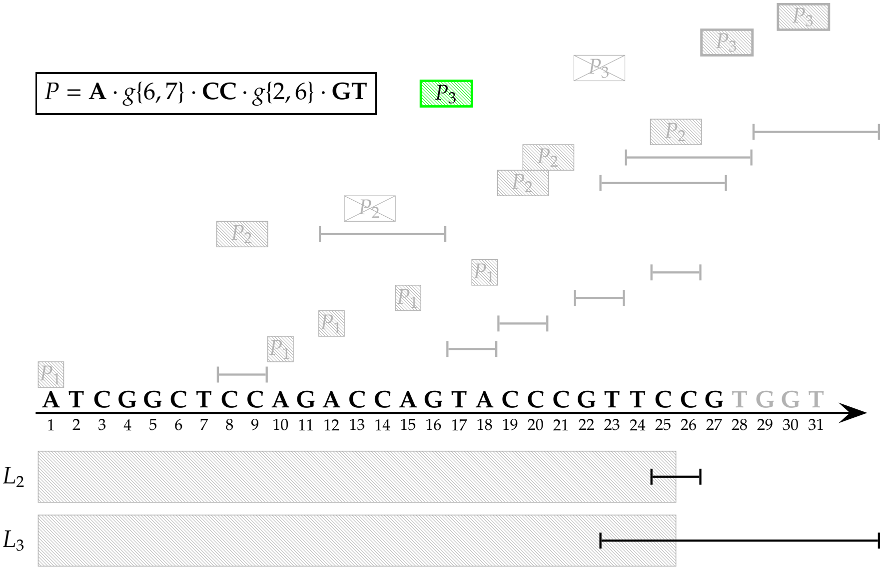
Illustrating the Algorithm

$P = \mathbf{A} \cdot g\{6,7\} \cdot \mathbf{CC} \cdot g\{2,6\} \cdot \mathbf{GT}$



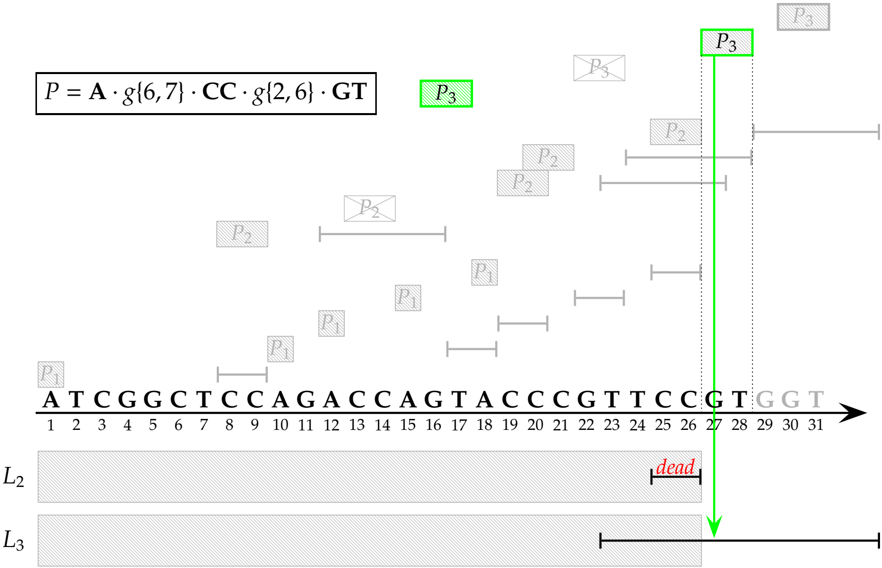
Illustrating the Algorithm

$P = \mathbf{A} \cdot g\{6,7\} \cdot \mathbf{CC} \cdot g\{2,6\} \cdot \mathbf{GT}$



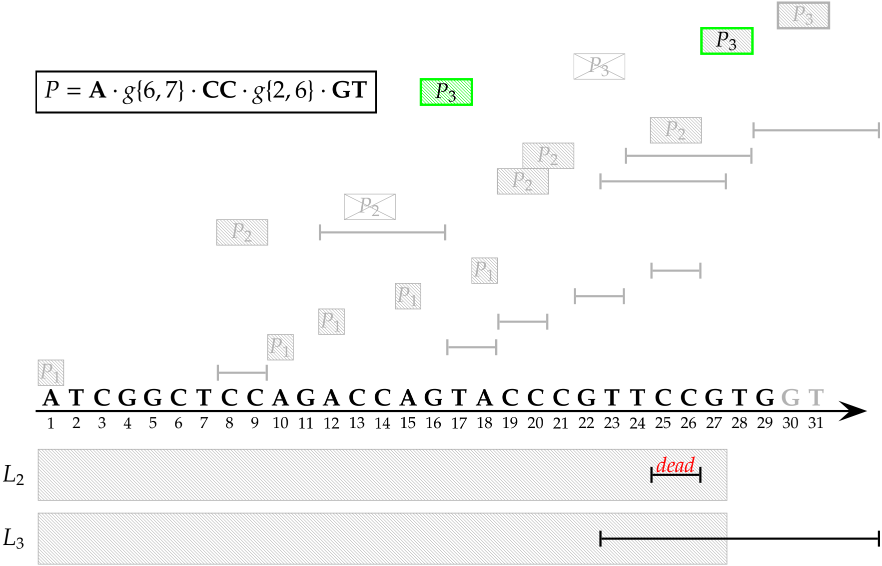
Illustrating the Algorithm

$P = \mathbf{A} \cdot g\{6,7\} \cdot \mathbf{CC} \cdot g\{2,6\} \cdot \mathbf{GT}$



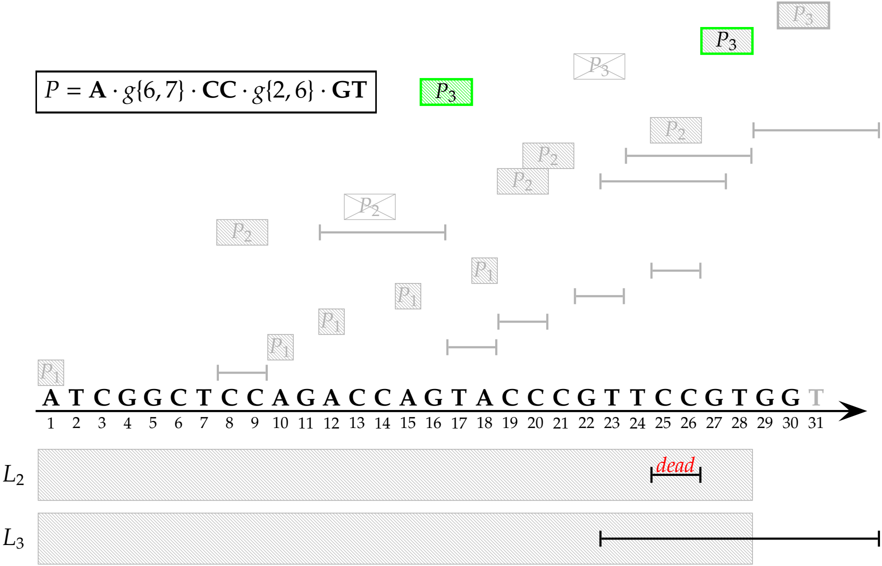
Illustrating the Algorithm

$P = \mathbf{A} \cdot g\{6,7\} \cdot \mathbf{CC} \cdot g\{2,6\} \cdot \mathbf{GT}$

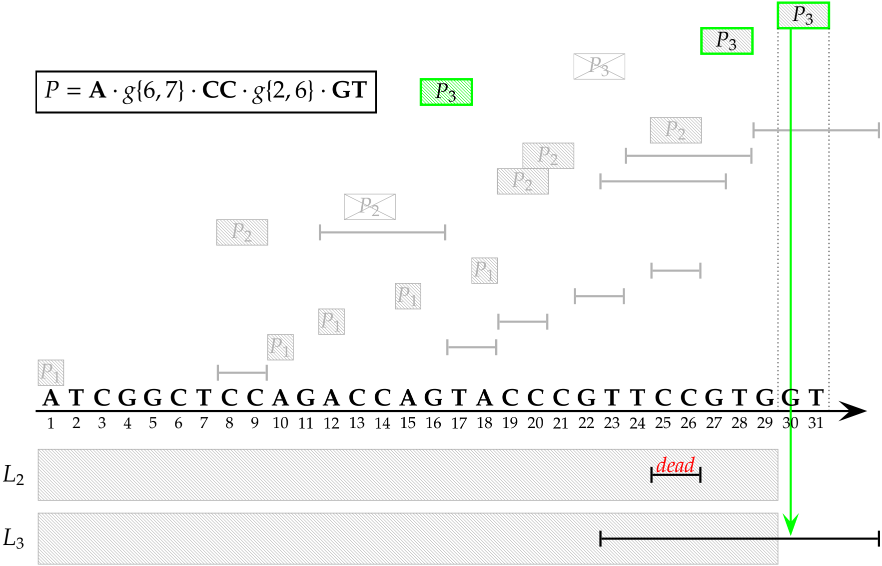


Illustrating the Algorithm

$P = \mathbf{A} \cdot g\{6,7\} \cdot \mathbf{CC} \cdot g\{2,6\} \cdot \mathbf{GT}$

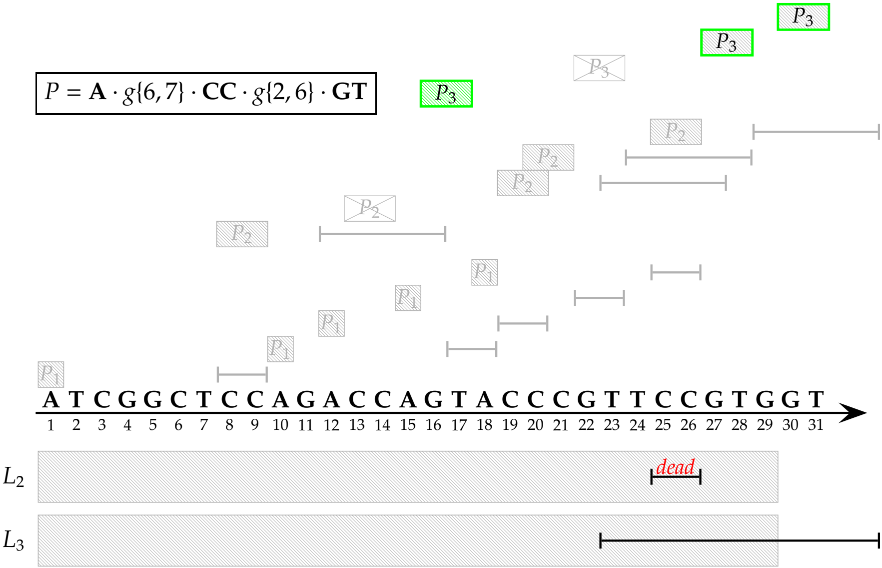


Illustrating the Algorithm



Illustrating the Algorithm

$P = \mathbf{A} \cdot g\{6,7\} \cdot \mathbf{CC} \cdot g\{2,6\} \cdot \mathbf{GT}$



Time and Space

Claim: The algorithm runs in $O((n + m)\log k + \alpha)$ time and uses $O(m + A)$ space.

Time and Space

Claim: The algorithm runs in $O((n+m)\log k + \alpha)$ time and uses $O(m + A)$ space.

Time

- Processing T using AC automaton takes $O((n+m)\log k + \alpha)$ time.
- At most α ranges are added and removed, so $O(\alpha)$ extra time is spent maintaining the lists.

Time and Space

Claim: The algorithm runs in $O((n+m)\log k + \alpha)$ time and uses $O(m + A)$ space.

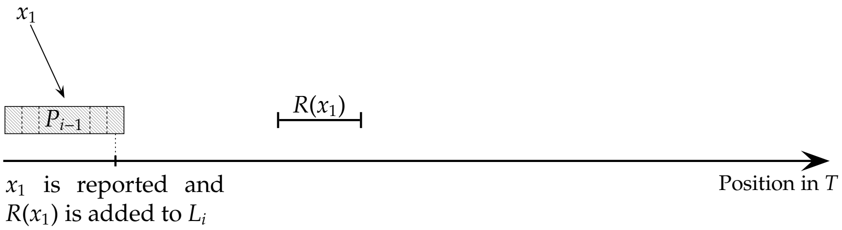
Time

- Processing T using AC automaton takes $O((n+m)\log k + \alpha)$ time.
- At most α ranges are added and removed, so $O(\alpha)$ extra time is spent maintaining the lists.

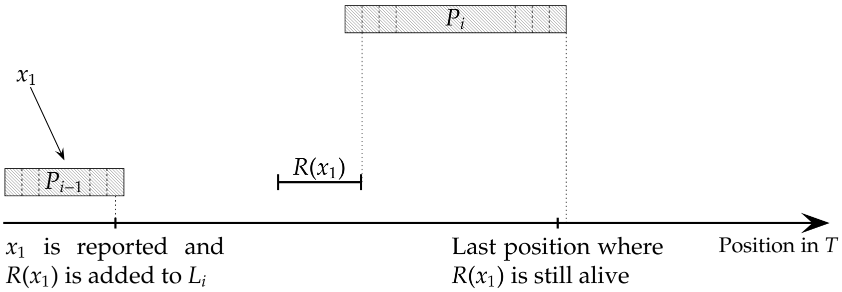
Space

- AC automaton takes $O(m)$ space.
- How much space is used by $L_2, \ldots, L_k$?

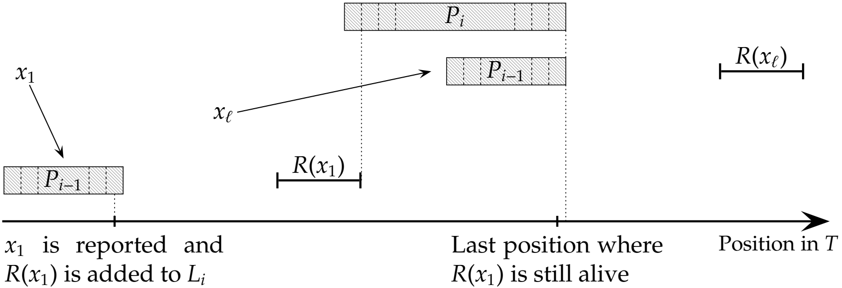
Maximum Size of L_i



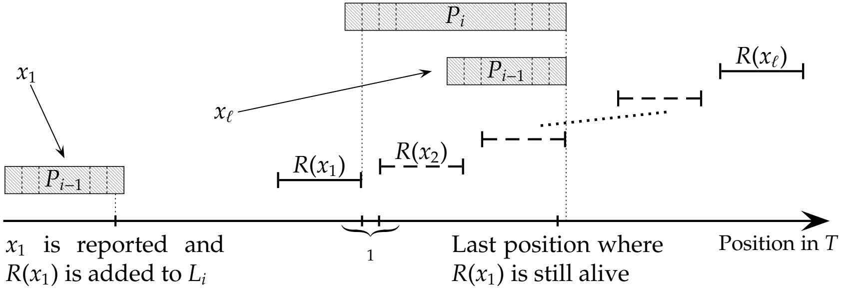
Maximum Size of L_i



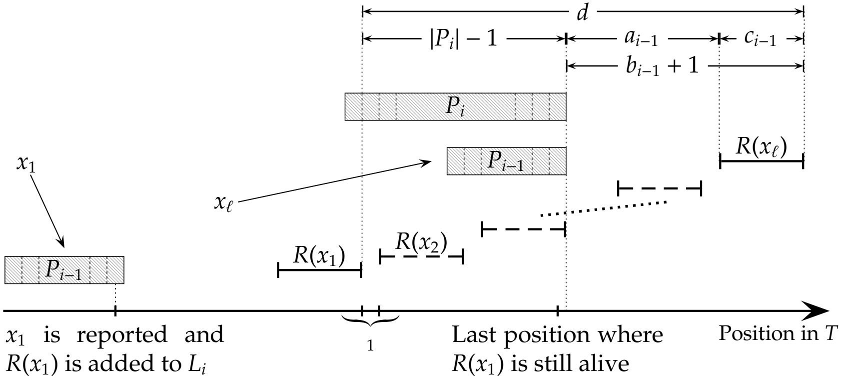
Maximum Size of L_i



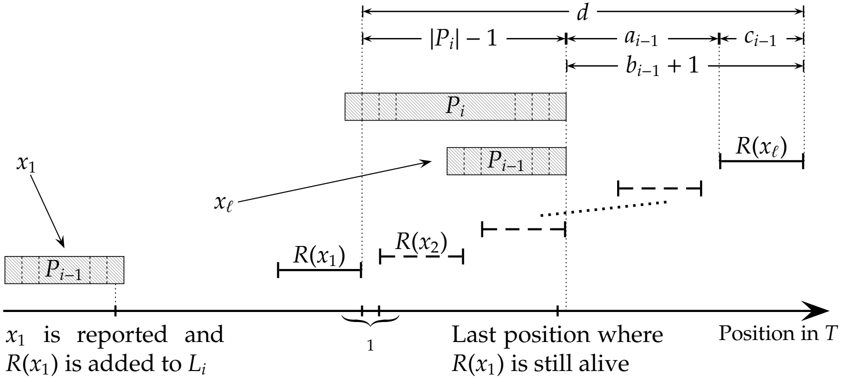
Maximum Size of L_i



Maximum Size of L_i

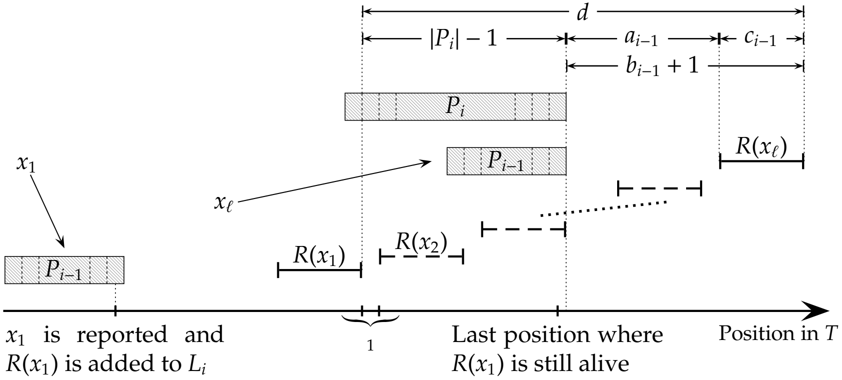


Maximum Size of L_i



$$|L_i| \leq \left\lfloor \frac{d}{c_{i-1}+1} \right\rfloor + 1 = \left\lfloor \frac{2c_{i-1} + |P_i| + a_{i-1}}{c_{i-1}+1} \right\rfloor = O(|P_i| + a_{i-1}) .$$

Maximum Size of L_i



$$|L_i| \le \left\lfloor \frac{d}{c_{i-1} + 1} \right\rfloor + 1 = \left\lfloor \frac{2c_{i-1} + |P_i| + a_{i-1}}{c_{i-1} + 1} \right\rfloor = O(|P_i| + a_{i-1}) .$$

Total space: $\displaystyle \sum_{i=2}^{k} |L_i| = O\left(\sum_{i=2}^{k} |P_i| + \sum_{i=1}^{k-1} a_i \right) = O(m + A)$