Efficient Index for Weighted Sequences

Carl Barton[1], **Tomasz Kociumaka**[2],
Solon P. Pissis[3], Jakub Radoszewski[2,3]

[1]Queen Mary University of London, UK

[2]University of Warsaw, Poland

[3]King's College London, UK

CPM 2016
Tel Aviv, Israel
June 27, 2016

Weighted Sequences

Weighted sequence (position weight matrix, PWM):
probability of occurrence $\pi_i(a)$ for each position i and letter $a \in \Sigma$.

$$\sum_{a \in \Sigma} \pi_i(a) = 1$$

a $\frac{1}{4}$	a 1	a $\frac{3}{4}$	a $\frac{1}{2}$	a 1
b $\frac{3}{4}$	b 0	b $\frac{1}{4}$	b $\frac{1}{2}$	b 0

Weighted Sequences

Weighted sequence (position weight matrix, PWM):
probability of occurrence $\pi_i(a)$ for each position i and letter $a \in \Sigma$.

$$\sum_{a \in \Sigma} \pi_i(a) = 1$$

a $\frac{1}{4}$	a 1	a $\frac{3}{4}$	a $\frac{1}{2}$	a 1
b $\frac{3}{4}$	b 0	b $\frac{1}{4}$	b $\frac{1}{2}$	b 0

Encodes probability distribution on solid strings:

Weighted Sequences

Weighted sequence (position weight matrix, PWM):
probability of occurrence $\pi_i(a)$ for each position i and letter $a \in \Sigma$.

$$\sum_{a \in \Sigma} \pi_i(a) = 1$$

a $\frac{1}{4}$	a 1	a $\frac{3}{4}$	a $\frac{1}{2}$	a 1
b $\frac{3}{4}$	b 0	b $\frac{1}{4}$	b $\frac{1}{2}$	b 0

Encodes probability distribution on solid strings:

b a b a a probability $\frac{3}{32}$

Weighted Sequences

Weighted sequence (position weight matrix, PWM):
probability of occurrence $\pi_i(a)$ for each position i and letter $a \in \Sigma$.

$$\sum_{a \in \Sigma} \pi_i(a) = 1$$

a $\frac{1}{4}$	a 1	a $\frac{3}{4}$	a $\frac{1}{2}$	a 1
b $\frac{3}{4}$	b 0	b $\frac{1}{4}$	b $\frac{1}{2}$	b 0

Encodes probability distribution on solid strings:

b a b a a probability $\frac{3}{32}$

Match specified by **threshold** (cut-off) probability $\frac{1}{z}$.

Weighted Sequences

Weighted sequence (position weight matrix, PWM):
probability of occurrence $\pi_i(a)$ for each position i and letter $a \in \Sigma$.

$$\sum_{a \in \Sigma} \pi_i(a) = 1$$

a $\frac{1}{4}$	a 1	a $\frac{3}{4}$	a $\frac{1}{2}$	a 1
b $\frac{3}{4}$	b 0	b $\frac{1}{4}$	b $\frac{1}{2}$	b 0

Encodes probability distribution on solid strings:

$\frac{1}{z} = \frac{1}{8}$ b a b a a probability $\frac{3}{32} < \frac{1}{8}$ NO

Match specified by **threshold** (cut-off) probability $\frac{1}{z}$.

Weighted Sequences

Weighted sequence (position weight matrix, PWM):
probability of occurrence $\pi_i(a)$ for each position i and letter $a \in \Sigma$.

$$\sum_{a \in \Sigma} \pi_i(a) = 1$$

a $\frac{1}{4}$	a 1	a $\frac{3}{4}$	a $\frac{1}{2}$	a 1
b $\frac{3}{4}$	b 0	b $\frac{1}{4}$	b $\frac{1}{2}$	b 0

Encodes probability distribution on solid strings:

$\frac{1}{z} = \frac{1}{8}$ b a a b a probability $\frac{9}{32} \geq \frac{1}{8}$ YES

Match specified by **threshold** (cut-off) probability $\frac{1}{z}$.

Weighted Sequences: Motivation

Origin

Computational biology:

1982 Stormo, Schneider, Gold, and Ehrenfeucht
Nucleic Acids Research.

Weighted Sequences: Motivation

Origin

Computational biology:

1982 Stormo, Schneider, Gold, and Ehrenfeucht
Nucleic Acids Research.

Primary use

Motifs in biological sequences.

Weighted Sequences: Motivation

Origin

Computational biology:

1982 Stormo, Schneider, Gold, and Ehrenfeucht
Nucleic Acids Research.

Primary use

Motifs in biological sequences.

Related notion

Profiles (scoring matrices):

- additive (as $\log \pi_i(a)$),
- arbitrary scores (no equivalent of $\sum_{a \in \Sigma} \pi_i(a) = 1$ constraint).

Problems Studied

Pattern matching and indexing

Christodoulakis et al., 2004
Iliopoulos et al., 2006
Amir et al., 2008

Approximate and gapped pattern matching

Zhang et al., 2004
Amir et al., 2006
Zhang et al., 2010

Repetitions and regularities discovery

Iliopoulos et al., 2005
Christodoulakis et al., 2006
Zhang et al., 2013
Barton and Pissis, 2014

Longest common subsequence problem

Amir et al., 2009
Cygan et al., 2011

Alignment of weighted sequences

Na et al., 2009

Problems Studied

Pattern matching and indexing ← **this work**

Christodoulakis et al., 2004
Iliopoulos et al., 2006
Amir et al., 2008

Approximate and gapped pattern matching

Zhang et al., 2004
Amir et al., 2006
Zhang et al., 2010

Repetitions and regularities discovery

Iliopoulos et al., 2005
Christodoulakis et al., 2006
Zhang et al., 2013
Barton and Pissis, 2014

Longest common subsequence problem

Amir et al., 2009
Cygan et al., 2011

Alignment of weighted sequences

Na et al., 2009

Indexing for Weighted Pattern Matching

Input

T – weighted sequence of length n over alphabet of size σ

($\sigma = \mathcal{O}(1)$ in this talk)

$\frac{1}{z}$ – threshold probability

Query Input

P – string pattern of length m

Query Output

$\mathrm{Occ} = \{i : P \text{ matches } T[i, i+m-1] \text{ with probability } \geq \frac{1}{z}\}$

Indexing for Weighted Pattern Matching

Input

T – weighted sequence of length n over alphabet of size σ

($\sigma = \mathcal{O}(1)$ in this talk)

$\frac{1}{z}$ – threshold probability

Query Input

P – string pattern of length m

Query Output

$\mathrm{Occ} = \{i : P \text{ matches } T[i, i+m-1] \text{ with probability} \geq \frac{1}{z}\}$

a $\frac{1}{4}$	a 1	a $\frac{3}{4}$	a $\frac{1}{2}$	a 1
b $\frac{3}{4}$	b 0	b $\frac{1}{4}$	b $\frac{1}{2}$	b 0

Indexing for Weighted Pattern Matching

Input

T – weighted sequence of length n over alphabet of size σ

($\sigma = \mathcal{O}(1)$ in this talk)

$\frac{1}{z}$ – threshold probability

Query Input

P – string pattern of length m

Query Output

$\mathrm{Occ} = \{i : P \text{ matches } T[i, i+m-1] \text{ with probability} \geq \frac{1}{z}\}$

a $\frac{1}{4}$	a 1	a $\frac{3}{4}$	a $\frac{1}{2}$	a 1
b $\frac{3}{4}$	b 0	b $\frac{1}{4}$	b $\frac{1}{2}$	b 0

$\frac{1}{z} = \frac{1}{8}$ $\quad P = \mathtt{aba}$

Indexing for Weighted Pattern Matching

Input

T – weighted sequence of length n over alphabet of size σ

($\sigma = \mathcal{O}(1)$ in this talk)

$\frac{1}{z}$ – threshold probability

Query Input

P – string pattern of length m

Query Output

$\mathrm{Occ} = \{i : P \text{ matches } T[i, i+m-1] \text{ with probability} \geq \frac{1}{z}\}$

a $\frac{1}{4}$	a 1	a $\frac{3}{4}$	a $\frac{1}{2}$	a 1
b $\frac{3}{4}$	b 0	b $\frac{1}{4}$	b $\frac{1}{2}$	b 0

a b a

$\frac{1}{z} = \frac{1}{8}$ $P = \mathtt{aba}$ probability $0 < \frac{1}{8}$ NO $\mathrm{Occ} = \{\}$

Indexing for Weighted Pattern Matching

Input

T – weighted sequence of length n over alphabet of size σ

($\sigma = \mathcal{O}(1)$ in this talk)

$\frac{1}{z}$ – threshold probability

Query Input

P – string pattern of length m

Query Output

$\mathrm{Occ} = \{i : P \text{ matches } T[i, i+m-1] \text{ with probability} \geq \frac{1}{z}\}$

a $\frac{1}{4}$	a 1	a $\frac{3}{4}$	a $\frac{1}{2}$	a 1
b $\frac{3}{4}$	b 0	b $\frac{1}{4}$	b $\frac{1}{2}$	b 0

a b a

$\frac{1}{z} = \frac{1}{8}$ $P = \mathtt{aba}$ probability $\frac{1}{8} \geq \frac{1}{8}$ YES $\mathrm{Occ} = \{2\}$

Indexing for Weighted Pattern Matching

Input

T – weighted sequence of length n over alphabet of size σ

($\sigma = \mathcal{O}(1)$ in this talk)

$\frac{1}{z}$ – threshold probability

Query Input

P – string pattern of length m

Query Output

$\mathrm{Occ} = \{i : P \text{ matches } T[i, i+m-1] \text{ with probability} \geq \frac{1}{z}\}$

a $\frac{1}{4}$	a 1	a $\frac{3}{4}$	a $\frac{1}{2}$	a 1
b $\frac{3}{4}$	b 0	b $\frac{1}{4}$	b $\frac{1}{2}$	b 0

a b a

$\frac{1}{z} = \frac{1}{8}$ $P = \mathtt{aba}$ probability $\frac{3}{8} \geq \frac{1}{8}$ YES $\mathrm{Occ} = \{2, 3\}$

Indexing for Weighted Pattern Matching

Input

T – weighted sequence of length n over alphabet of size σ

($\sigma = \mathcal{O}(1)$ in this talk)

$\frac{1}{z}$ – threshold probability

Query Input

P – string pattern of length m

Query Output

$\mathrm{Occ} = \{i : P \text{ matches } T[i, i+m-1] \text{ with probability} \geq \frac{1}{z}\}$

a $\frac{1}{4}$	a 1	a $\frac{3}{4}$	a $\frac{1}{2}$	a 1
b $\frac{3}{4}$	b 0	b $\frac{1}{4}$	b $\frac{1}{2}$	b 0

$\frac{1}{z} = \frac{1}{8}$ $\quad P = \mathtt{aba}$ $\quad \mathrm{Occ} = \{2, 3\}$

Results

Iliopoulos et al., 2006:

- $\mathcal{O}(n \cdot f(z))$ space
- $\mathcal{O}(n \cdot f(z))$ construction time
- $\mathcal{O}(m + |\mathrm{Occ}|)$ query time

Amir et al., 2008:

- $\mathcal{O}(nz^2 \log z)$ space
- $\mathcal{O}(nz^2 \log z(\log\log z + \log\log n))$ construction time
- $\mathcal{O}(nz^2 \log z)$ construction time follows from later results: Iliopoulos & Rahman, 2008; Juan et al. 2009
- $\mathcal{O}(m + |\mathrm{Occ}|)$ query time

Results

Iliopoulos et al., 2006:

- $\mathcal{O}(n \cdot f(z))$ space
- $\mathcal{O}(n \cdot f(z))$ construction time
- $\mathcal{O}(m + |\mathrm{Occ}|)$ query time

Amir et al., 2008:

- $\mathcal{O}(nz^2 \log z)$ space
- $\mathcal{O}(nz^2 \log z(\log\log z + \log\log n))$ construction time
- $\mathcal{O}(nz^2 \log z)$ construction time follows from later results: Iliopoulos & Rahman, 2008; Juan et al. 2009
- $\mathcal{O}(m + |\mathrm{Occ}|)$ query time

Our results:

- $\mathcal{O}(nz)$ space
- $\mathcal{O}(nz)$ construction time
- $\mathcal{O}(m + |\mathrm{Occ}|)$ query time

Maximal solid factors

Solid factor: an occurrence of a string (at a particular position)

Maximal solid factor: cannot be extended to an occurrence of a superstring

Maximal solid factors

Solid factor: an occurrence of a string (at a particular position)

Maximal solid factor: cannot be extended to an occurrence of a superstring

a $\frac{1}{4}$	a 1	a $\frac{3}{4}$	a $\frac{1}{2}$	a 1
b $\frac{3}{4}$	b 0	b $\frac{1}{4}$	b $\frac{1}{2}$	b 0

Maximal solid factors

Solid factor: an occurrence of a string (at a particular position)

Maximal solid factor: cannot be extended to an occurrence of a superstring

a $\frac{1}{4}$	a 1	a $\frac{3}{4}$	a $\frac{1}{2}$	a 1
b $\frac{3}{4}$	b 0	b $\frac{1}{4}$	b $\frac{1}{2}$	b 0

Probability: $\frac{1}{4} \geq \frac{1}{8}$

a

Maximal solid factors

Solid factor: an occurrence of a string (at a particular position)

Maximal solid factor: cannot be extended to an occurrence of a superstring

a $\frac{1}{4}$	a 1	a $\frac{3}{4}$	a $\frac{1}{2}$	a 1
b $\frac{3}{4}$	b 0	b $\frac{1}{4}$	b $\frac{1}{2}$	b 0

Probability: $\frac{1}{4} \geq \frac{1}{8}$

a a

Maximal solid factors

Solid factor: an occurrence of a string (at a particular position)

Maximal solid factor: cannot be extended to an occurrence of a superstring

a $\frac{1}{4}$	a 1	a $\frac{3}{4}$	a $\frac{1}{2}$	a 1
b $\frac{3}{4}$	b 0	b $\frac{1}{4}$	b $\frac{1}{2}$	b 0

Probability: $\frac{3}{16} \geq \frac{1}{8}$

a a a

Maximal solid factors

Solid factor: an occurrence of a string (at a particular position)

Maximal solid factor: cannot be extended to an occurrence of a superstring

a $\frac{1}{4}$	a 1	a $\frac{3}{4}$	a $\frac{1}{2}$	a 1
b $\frac{3}{4}$	b 0	b $\frac{1}{4}$	b $\frac{1}{2}$	b 0

Probability: $\frac{3}{32} < \frac{1}{8}$

a a a a

Maximal solid factors

Solid factor: an occurrence of a string (at a particular position)

Maximal solid factor: cannot be extended to an occurrence of a superstring

a $\frac{1}{4}$	a 1	a $\frac{3}{4}$	a $\frac{1}{2}$	a 1
b $\frac{3}{4}$	b 0	b $\frac{1}{4}$	b $\frac{1}{2}$	b 0

Probability: $\frac{3}{32} < \frac{1}{8}$

a a a b

Maximal solid factors

Solid factor: an occurrence of a string (at a particular position)

Maximal solid factor: cannot be extended to an occurrence of a superstring

a $\frac{1}{4}$	a 1	a $\frac{3}{4}$	a $\frac{1}{2}$	a 1
b $\frac{3}{4}$	b 0	b $\frac{1}{4}$	b $\frac{1}{2}$	b 0

Probability: $\frac{3}{16} \geq \frac{1}{8}$

a a a

Maximal solid factors

Solid factor: an occurrence of a string (at a particular position)

Maximal solid factor: cannot be extended to an occurrence of a superstring

a $\frac{1}{4}$	a 1	a $\frac{3}{4}$	a $\frac{1}{2}$	a 1
b $\frac{3}{4}$	b 0	b $\frac{1}{4}$	b $\frac{1}{2}$	b 0

Probability: $\frac{1}{16} < \frac{1}{8}$

a a a

a a b

Maximal solid factors

Solid factor: an occurrence of a string (at a particular position)

Maximal solid factor: cannot be extended to an occurrence of a superstring

a $\frac{1}{4}$	a 1	a $\frac{3}{4}$	a $\frac{1}{2}$	a 1
b $\frac{3}{4}$	b 0	b $\frac{1}{4}$	b $\frac{1}{2}$	b 0

Probability: $\frac{3}{4} \geq \frac{1}{8}$

a a a

b

Maximal solid factors

Solid factor: an occurrence of a string (at a particular position)

Maximal solid factor: cannot be extended to an occurrence of a superstring

a $\frac{1}{4}$	a 1	a $\frac{3}{4}$	a $\frac{1}{2}$	a 1
b $\frac{3}{4}$	b 0	b $\frac{1}{4}$	b $\frac{1}{2}$	b 0

Probability: $\frac{3}{4} \geq \frac{1}{8}$

a a a

b a

Maximal solid factors

Solid factor: an occurrence of a string (at a particular position)

Maximal solid factor: cannot be extended to an occurrence of a superstring

a $\frac{1}{4}$	a 1	a $\frac{3}{4}$	a $\frac{1}{2}$	a 1
b $\frac{3}{4}$	b 0	b $\frac{1}{4}$	b $\frac{1}{2}$	b 0

Probability: $\frac{9}{16} \geq \frac{1}{8}$

a a a

b a a

Maximal solid factors

Solid factor: an occurrence of a string (at a particular position)

Maximal solid factor: cannot be extended to an occurrence of a superstring

a $\frac{1}{4}$	a 1	a $\frac{3}{4}$	a $\frac{1}{2}$	a 1
b $\frac{3}{4}$	b 0	b $\frac{1}{4}$	b $\frac{1}{2}$	b 0

Probability: $\frac{9}{32} \geq \frac{1}{8}$

a a a

b a a a

Maximal solid factors

Solid factor: an occurrence of a string (at a particular position)

Maximal solid factor: cannot be extended to an occurrence of a superstring

a $\frac{1}{4}$	a 1	a $\frac{3}{4}$	a $\frac{1}{2}$	a 1
b $\frac{3}{4}$	b 0	b $\frac{1}{4}$	b $\frac{1}{2}$	b 0

Probability: $\frac{9}{32} \geq \frac{1}{8}$

a a a

b a a a a

Maximal solid factors

Solid factor: an occurrence of a string (at a particular position)

Maximal solid factor: cannot be extended to an occurrence of a superstring

a $\frac{1}{4}$	a 1	a $\frac{3}{4}$	a $\frac{1}{2}$	a 1
b $\frac{3}{4}$	b 0	b $\frac{1}{4}$	b $\frac{1}{2}$	b 0

Probability: $\frac{9}{32} \geq \frac{1}{8}$

a	a	a		
b	a	a	a	a
b	a	a	b	

Maximal solid factors

Solid factor: an occurrence of a string (at a particular position)

Maximal solid factor: cannot be extended to an occurrence of a superstring

a $\frac{1}{4}$	a 1	a $\frac{3}{4}$	a $\frac{1}{2}$	a 1
b $\frac{3}{4}$	b 0	b $\frac{1}{4}$	b $\frac{1}{2}$	b 0

Probability: $\frac{9}{32} \geq \frac{1}{8}$

a	a	a		
b	a	a	a	a
b	a	a	b	a

Maximal solid factors

Solid factor: an occurrence of a string (at a particular position)

Maximal solid factor: cannot be extended to an occurrence of a superstring

a $\frac{1}{4}$	a 1	a $\frac{3}{4}$	a $\frac{1}{2}$	a 1
b $\frac{3}{4}$	b 0	b $\frac{1}{4}$	b $\frac{1}{2}$	b 0

Probability: $\frac{3}{16} \geq \frac{1}{8}$

a	a	a		
b	a	a	a	a
b	a	a	b	a
b	a	b		

Maximal solid factors

Solid factor: an occurrence of a string (at a particular position)

Maximal solid factor: cannot be extended to an occurrence of a superstring

a $\frac{1}{4}$	a 1	a $\frac{3}{4}$	a $\frac{1}{2}$	a 1
b $\frac{3}{4}$	b 0	b $\frac{1}{4}$	b $\frac{1}{2}$	b 0

Probability: $\frac{3}{32} < \frac{1}{8}$

a	a	a		
b	a	a	a	a
b	a	a	b	a
b	a	b	a	

Maximal solid factors

Solid factor: an occurrence of a string (at a particular position)

Maximal solid factor: cannot be extended to an occurrence of a superstring

a $\frac{1}{4}$	a 1	a $\frac{3}{4}$	a $\frac{1}{2}$	a 1
b $\frac{3}{4}$	b 0	b $\frac{1}{4}$	b $\frac{1}{2}$	b 0

Probability: $\frac{3}{32} < \frac{1}{8}$

a	a	a		
b	a	a	a	a
b	a	a	b	a
b	a	b	b	

Maximal solid factors

Solid factor: an occurrence of a string (at a particular position)

Maximal solid factor: cannot be extended to an occurrence of a superstring

a $\frac{1}{4}$	a 1	a $\frac{3}{4}$	a $\frac{1}{2}$	a 1
b $\frac{3}{4}$	b 0	b $\frac{1}{4}$	b $\frac{1}{2}$	b 0

Probability: $\frac{3}{16} \geq \frac{1}{8}$

a	a	a		
b	a	a	a	a
b	a	a	b	a
b	a	b		

Maximal solid factors

Solid factor: an occurrence of a string (at a particular position)

Maximal solid factor: cannot be extended to an occurrence of a superstring

a $\frac{1}{4}$	a 1	a $\frac{3}{4}$	a $\frac{1}{2}$	a 1
b $\frac{3}{4}$	b 0	b $\frac{1}{4}$	b $\frac{1}{2}$	b 0

Probability: $\frac{1}{8} \geq \frac{1}{8}$

a	a	a		
b	a	a	a	a
b	a	a	b	a
b	a	b		
	a	b	a	a

Maximal solid factors

Solid factor: an occurrence of a string (at a particular position)

Maximal solid factor: cannot be extended to an occurrence of a superstring

a $\frac{1}{4}$	a 1	a $\frac{3}{4}$	a $\frac{1}{2}$	a 1
b $\frac{3}{4}$	b 0	b $\frac{1}{4}$	b $\frac{1}{2}$	b 0

Probability: $\frac{1}{8} \geq \frac{1}{8}$

a	a	a		
b	a	a	a	a
b	a	a	b	a
b	a	b		
	a	b	a	a
	a	b	b	a

Maximal solid factors

Solid factor: an occurrence of a string (at a particular position)

Maximal solid factor: cannot be extended to an occurrence of a superstring

a $\frac{1}{4}$	a 1	a $\frac{3}{4}$	a $\frac{1}{2}$	a 1
b $\frac{3}{4}$	b 0	b $\frac{1}{4}$	b $\frac{1}{2}$	b 0

Probability:

a	a	a		
b	a	a	a	a
b	a	a	b	a
b	a	b		
	a	b	a	a
	a	b	b	a

Maximal solid factors

Solid factor: an occurrence of a string (at a particular position)

Maximal solid factor: cannot be extended to an occurrence of a superstring

a $\frac{1}{4}$	a 1	a $\frac{3}{4}$	a $\frac{1}{2}$	a 1
b $\frac{3}{4}$	b 0	b $\frac{1}{4}$	b $\frac{1}{2}$	b 0

Probability:

a	a	a		
b	a	a	a	a
b	a	a	b	a
b	a	b		
	a	b	a	a
	a	b	b	a

Naive solution: Pattern matching in all maximal solid factors.

- Might be of length $\Theta(n^2 z)$ in total.

Previous approach

Heavy character: highest probability at the given position.

Previous approach

Heavy character: highest probability at the given position.

a $\frac{1}{4}$	a 1	a $\frac{3}{4}$	a $\frac{1}{2}$	a 1
b $\frac{3}{4}$	b 0	b $\frac{1}{4}$	b $\frac{1}{2}$	b 0

Previous approach

Heavy character: highest probability at the given position.

a $\frac{1}{4}$	a 1	a $\frac{3}{4}$	a $\frac{1}{2}$	a 1
b $\frac{3}{4}$	b 0	b $\frac{1}{4}$	b $\frac{1}{2}$	b 0

Previous approach

Heavy character: highest probability at the given position.

a $\frac{1}{4}$	a 1	a $\frac{3}{4}$	a $\frac{1}{2}$	a 1
b $\frac{3}{4}$	b 0	b $\frac{1}{4}$	b $\frac{1}{2}$	b 0

a	a	a		
b	a	a	a	a
b	a	a	b	a
b	a	b		
	a	b	a	a
	a	b	b	a

Amir et al. (reinterpreted):

1. Merge overlapping solid factors with equal heavy extensions.

Previous approach

Heavy character: highest probability at the given position.

a $\frac{1}{4}$	a 1	a $\frac{3}{4}$	a $\frac{1}{2}$	a 1
b $\frac{3}{4}$	b 0	b $\frac{1}{4}$	b $\frac{1}{2}$	b 0

a	a	a	b	a
b	a	a	a	a
b	a	a	b	a
b	a	b	b	a
b	a	b	a	a
b	a	b	b	a

Amir et al. (reinterpreted):

1. Merge overlapping solid factors with equal heavy extensions.

Previous approach

Heavy character: highest probability at the given position.

a $\frac{1}{4}$	a 1	a $\frac{3}{4}$	a $\frac{1}{2}$	a 1
b $\frac{3}{4}$	b 0	b $\frac{1}{4}$	b $\frac{1}{2}$	b 0

a	a	a	b	a
b	a	a	a	a
b	a	a	b	a
b	a	b	b	a
b	a	b	a	a

Amir et al. (reinterpreted):

1. Merge overlapping solid factors with equal heavy extensions.

Previous approach

Heavy character: highest probability at the given position.

a $\frac{1}{4}$	a 1	a $\frac{3}{4}$	a $\frac{1}{2}$	a 1
b $\frac{3}{4}$	b 0	b $\frac{1}{4}$	b $\frac{1}{2}$	b 0

a	a	a		
b	a	a	a	a
b	a	a	b	a
b	a	b	b	a
	a	b	a	a

Amir et al. (reinterpreted):

1. Merge overlapping solid factors with equal heavy extensions.
2. Each position covered by $\mathcal{O}(z^2 \log z)$ factors.

Previous approach

Heavy character: highest probability at the given position.

a $\frac{1}{4}$	a 1	a $\frac{3}{4}$	a $\frac{1}{2}$	a 1
b $\frac{3}{4}$	b 0	b $\frac{1}{4}$	b $\frac{1}{2}$	b 0

a	a	a		
b	a	a	a	a
b	a	a	b	a
b	a	b	b	a
	a	b	a	a

Amir et al. (reinterpreted):

1. Merge overlapping solid factors with equal heavy extensions.
2. Each position covered by $\mathcal{O}(z^2 \log z)$ factors.
3. **Property matching** in the resulting factors:
   upper bound on fragment length for each starting position.

Our approach

a $\frac{1}{4}$	a 1	a $\frac{3}{4}$	a $\frac{1}{2}$	a 1
b $\frac{3}{4}$	b 0	b $\frac{1}{4}$	b $\frac{1}{2}$	b 0

a	a	a		
b	a	a	a	a
b	a	a	b	a
b	a	b		
	a	b	a	a
	a	b	b	a

Our approach:

Our approach

a $\frac{1}{4}$	a 1	a $\frac{3}{4}$	a $\frac{1}{2}$	a 1
b $\frac{3}{4}$	b 0	b $\frac{1}{4}$	b $\frac{1}{2}$	b 0

a	a	a	b	a
b	a	a	a	a
b	a	a	b	a
b	a	b	b	a
	a	b	a	a
	a	b	b	a

Our approach:

1. Extend each maximal solid factor to the right.

Our approach

a $\frac{1}{4}$	a 1	a $\frac{3}{4}$	a $\frac{1}{2}$	a 1
b $\frac{3}{4}$	b 0	b $\frac{1}{4}$	b $\frac{1}{2}$	b 0

a	a	a	b	a
b	a	a	a	a
b	a	a	b	a
b	a	b	b	a
	a	b	a	a
	a	b	b	a

Our approach:

1. Extend each maximal solid factor to the right.
2. Build a trie of (the reverses of) the obtained suffixes.

Our approach

a $\frac{1}{4}$	a 1	a $\frac{3}{4}$	a $\frac{1}{2}$	a 1
b $\frac{3}{4}$	b 0	b $\frac{1}{4}$	b $\frac{1}{2}$	b 0

b	a	a	b	a
a	a	a	b	a
b	a	b	b	a
	a	b	b	a
b	a	a	a	a
	a	b	a	a

Our approach:

1. Extend each maximal solid factor to the right.
2. Build a trie of (the reverses of) the obtained suffixes.

Our approach

a $\frac{1}{4}$	a 1	a $\frac{3}{4}$	a $\frac{1}{2}$	a 1
b $\frac{3}{4}$	b 0	b $\frac{1}{4}$	b $\frac{1}{2}$	b 0

b	a	a	b	a
a	a	a	b	a
b	a	b	b	a
b	a	a	a	a
	a	b	a	a

Our approach:

1. Extend each maximal solid factor to the right.
2. Build a trie of (the reverses of) the obtained suffixes.

Our approach

a $\frac{1}{4}$	a 1	a $\frac{3}{4}$	a $\frac{1}{2}$	a 1
b $\frac{3}{4}$	b 0	b $\frac{1}{4}$	b $\frac{1}{2}$	b 0

b	a	a	b	a
a				
b	a	b		
b	a	a	a	
	a	b		

Our approach:

1. Extend each maximal solid factor to the right.
2. Build a trie of (the reverses of) the obtained suffixes.

Our approach

a $\frac{1}{4}$	a 1	a $\frac{3}{4}$	a $\frac{1}{2}$	a 1
b $\frac{3}{4}$	b 0	b $\frac{1}{4}$	b $\frac{1}{2}$	b 0

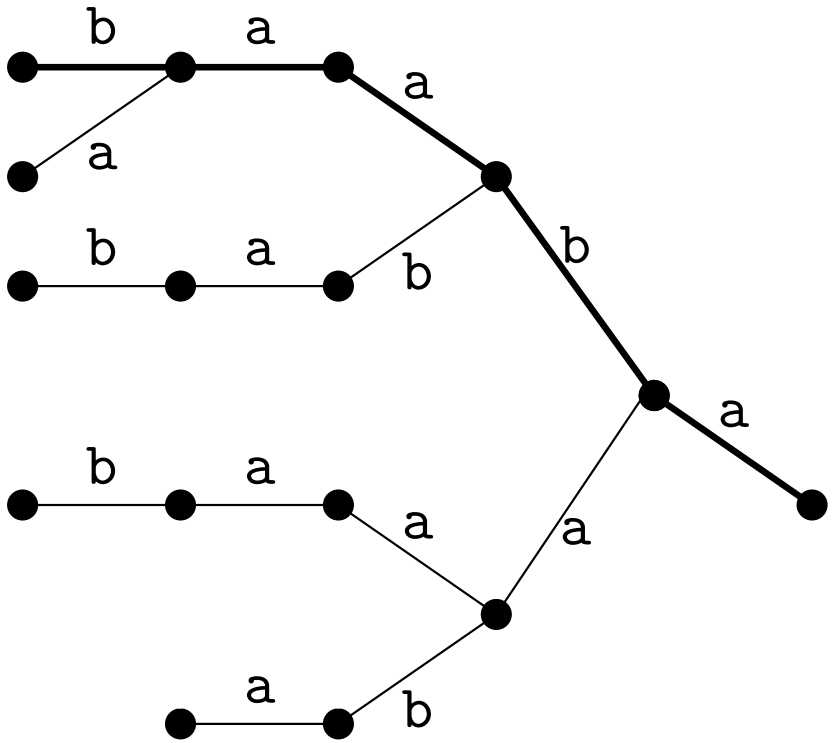
b	a	a	b	a
a				
b	a	b		
b	a	a	a	
	a	b		

Our approach:

1. Extend each maximal solid factor to the right.
2. Build a trie of (the reverses of) the obtained suffixes.
3. Property matching in a trie of $\mathcal{O}(nz)$ nodes.

Trie size

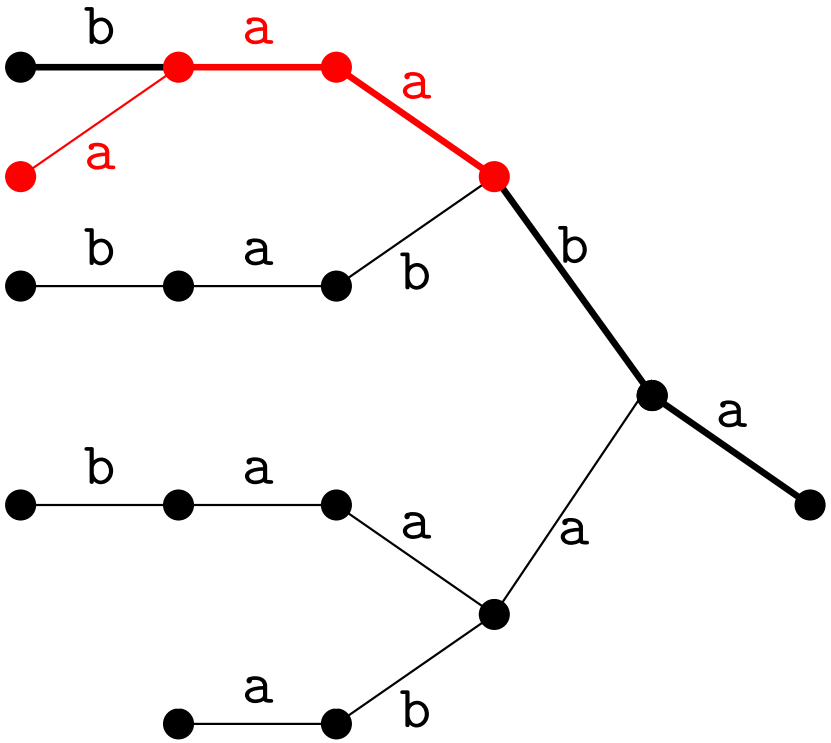
a $\frac{1}{4}$	a 1	a $\frac{3}{4}$	a $\frac{1}{2}$	a 1
b $\frac{3}{4}$	b 0	b $\frac{1}{4}$	b $\frac{1}{2}$	b 0



Bijection: nodes $\leftrightarrow$ right-maximal solid factors.

Trie size

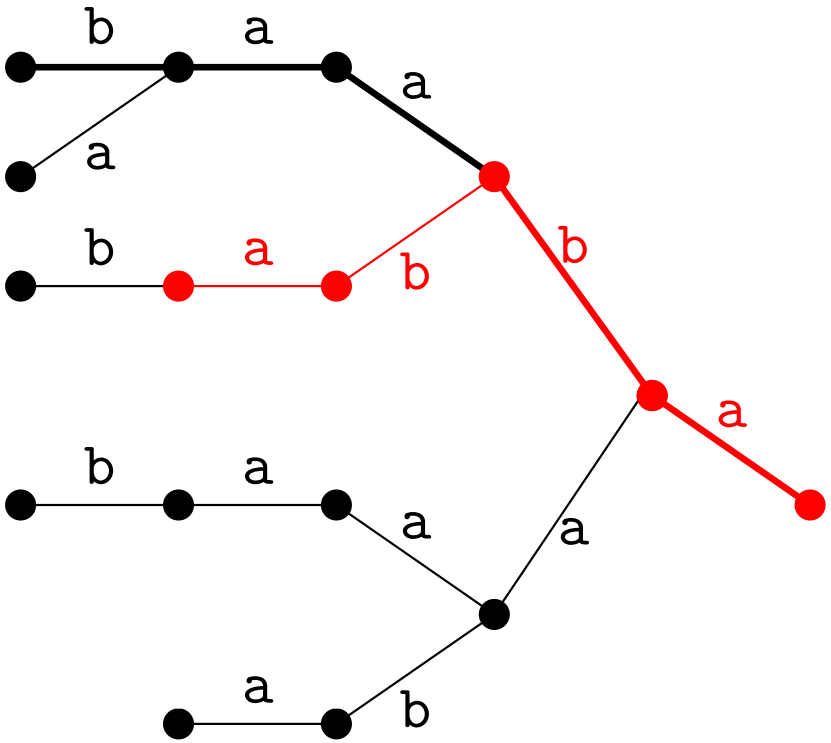
a $\frac{1}{4}$	a 1	a $\frac{3}{4}$	a $\frac{1}{2}$	a 1
b $\frac{3}{4}$	b 0	b $\frac{1}{4}$	b $\frac{1}{2}$	b 0



Bijection: nodes $\leftrightarrow$ right-maximal solid factors.

Trie size

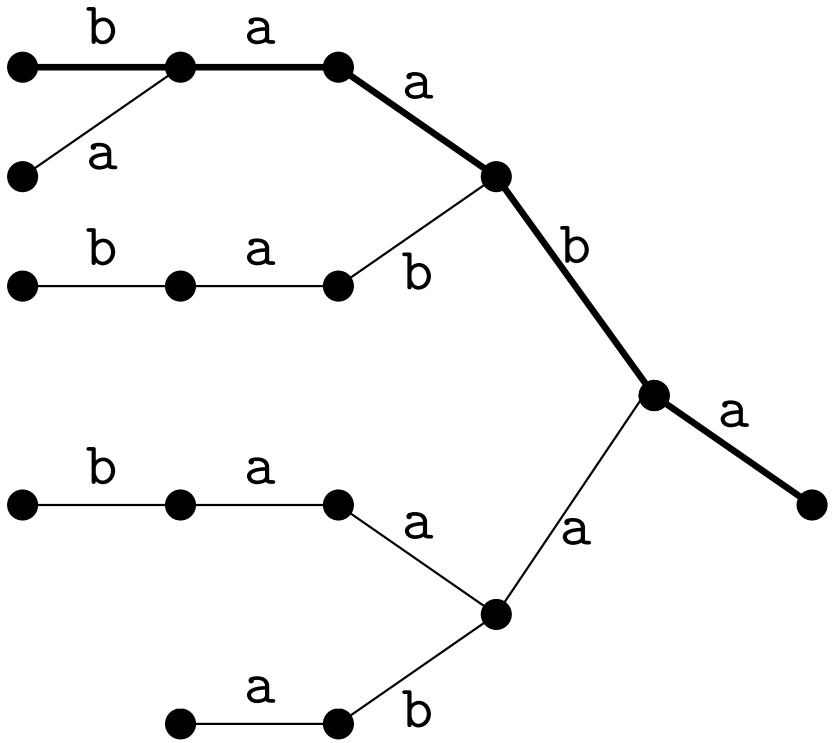
a $\frac{1}{4}$	a 1	a $\frac{3}{4}$	a $\frac{1}{2}$	a 1
b $\frac{3}{4}$	b 0	b $\frac{1}{4}$	b $\frac{1}{2}$	b 0



Bijection: nodes $\leftrightarrow$ right-maximal solid factors.

Trie size

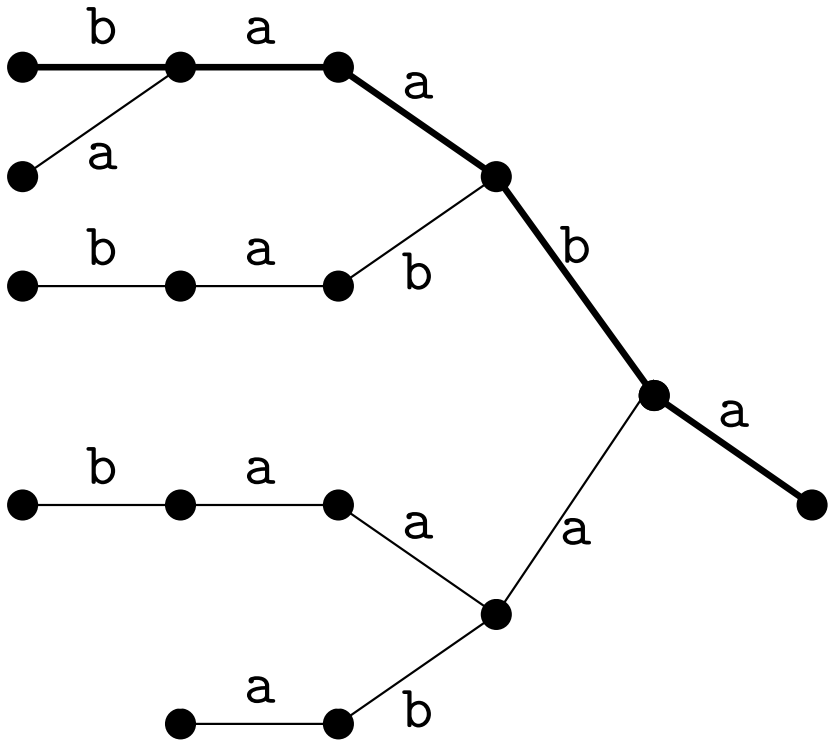
a $\frac{1}{4}$	a 1	a $\frac{3}{4}$	a $\frac{1}{2}$	a 1
b $\frac{3}{4}$	b 0	b $\frac{1}{4}$	b $\frac{1}{2}$	b 0



Bijection: nodes $\leftrightarrow$ right-maximal solid factors.

Trie size

a $\frac{1}{4}$	a 1	a $\frac{3}{4}$	a $\frac{1}{2}$	a 1
b $\frac{3}{4}$	b 0	b $\frac{1}{4}$	b $\frac{1}{2}$	b 0



Bijection: nodes $\leftrightarrow$ right-maximal solid factors.

Fact [Amir et al.]: $\leq z$ right-maximal solid factors start at any i.

Trie construction

a $\frac{1}{4}$	a 1	a $\frac{3}{4}$	a $\frac{1}{2}$	a 1
b $\frac{3}{4}$	b 0	b $\frac{1}{4}$	b $\frac{1}{2}$	b 0

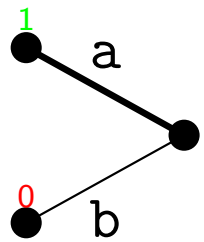
Level-by-level **construction**: for each node:

Trie construction

a $\frac{1}{4}$	a 1	a $\frac{3}{4}$	a $\frac{1}{2}$	a 1
b $\frac{3}{4}$	b 0	b $\frac{1}{4}$	b $\frac{1}{2}$	b 0

Level-by-level **construction**: for each node:

1. extend to the left using every possible character,

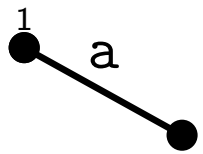


Trie construction

a $\frac{1}{4}$	a 1	a $\frac{3}{4}$	a $\frac{1}{2}$	a 1
b $\frac{3}{4}$	b 0	b $\frac{1}{4}$	b $\frac{1}{2}$	b 0

Level-by-level **construction**: for each node:

1. extend to the left using every possible character,
2. trim leaves with prob. $< \frac{1}{z}$ counting from the heavy path.

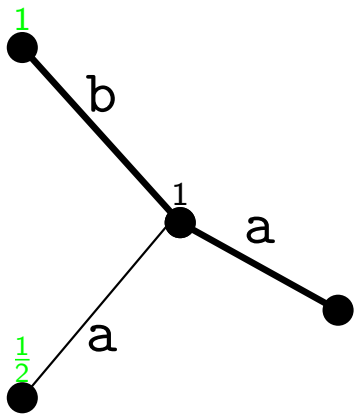


Trie construction

a $\frac{1}{4}$	a 1	a $\frac{3}{4}$	a $\frac{1}{2}$	a 1
b $\frac{3}{4}$	b 0	b $\frac{1}{4}$	b $\frac{1}{2}$	b 0

Level-by-level **construction**: for each node:

1. extend to the left using every possible character,
2. trim leaves with prob. $< \frac{1}{z}$ counting from the heavy path.

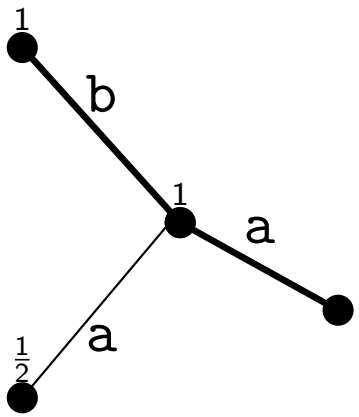


Trie construction

a $\frac{1}{4}$	a 1	a $\frac{3}{4}$	a $\frac{1}{2}$	a 1
b $\frac{3}{4}$	b 0	b $\frac{1}{4}$	b $\frac{1}{2}$	b 0

Level-by-level **construction**: for each node:

1. extend to the left using every possible character,
2. trim leaves with prob. $< \frac{1}{z}$ counting from the heavy path.

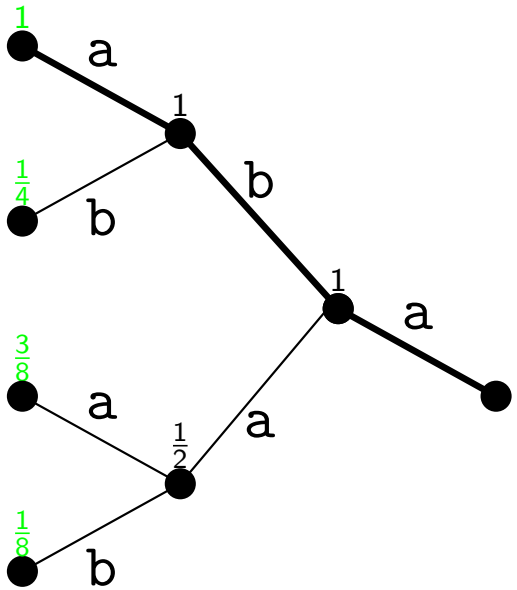


Trie construction

a $\frac{1}{4}$	a 1	a $\frac{3}{4}$	a $\frac{1}{2}$	a 1
b $\frac{3}{4}$	b 0	b $\frac{1}{4}$	b $\frac{1}{2}$	b 0

Level-by-level **construction**: for each node:

1. extend to the left using every possible character,
2. trim leaves with prob. $< \frac{1}{z}$ counting from the heavy path.

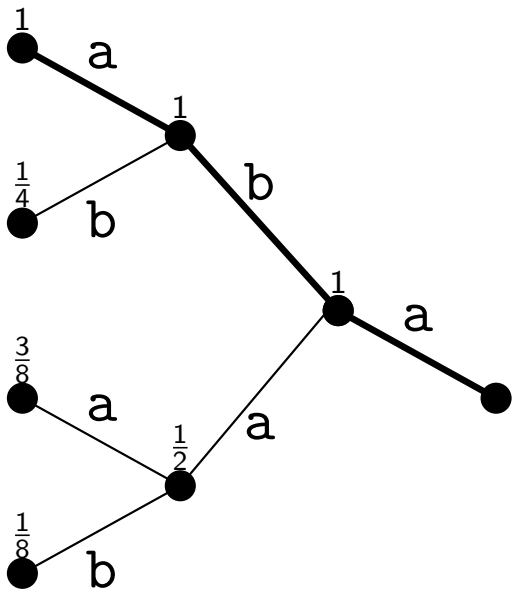


Trie construction

a $\frac{1}{4}$	a 1	a $\frac{3}{4}$	a $\frac{1}{2}$	a 1
b $\frac{3}{4}$	b 0	b $\frac{1}{4}$	b $\frac{1}{2}$	b 0

Level-by-level **construction**: for each node:

1. extend to the left using every possible character,
2. trim leaves with prob. $< \frac{1}{z}$ counting from the heavy path.

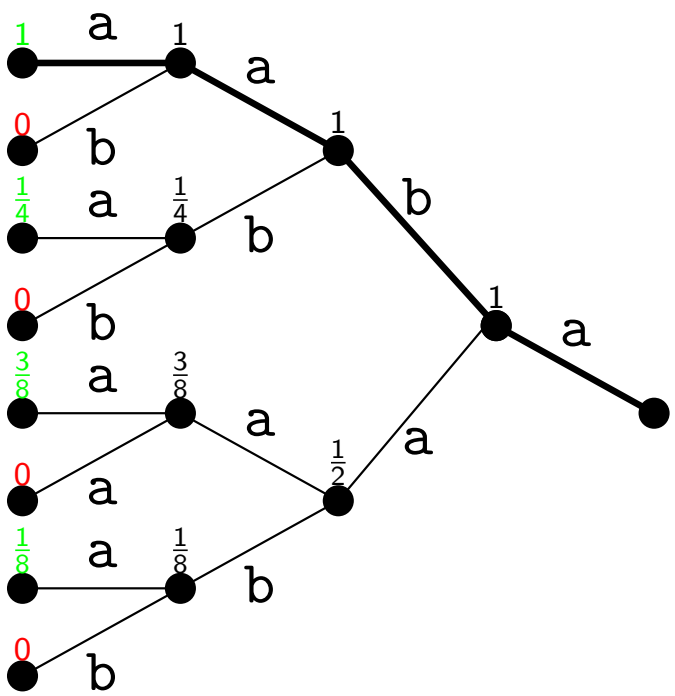


Trie construction

a $\frac{1}{4}$	a 1	a $\frac{3}{4}$	a $\frac{1}{2}$	a 1
b $\frac{3}{4}$	b 0	b $\frac{1}{4}$	b $\frac{1}{2}$	b 0

Level-by-level **construction**: for each node:

1. extend to the left using every possible character,
2. trim leaves with prob. $< \frac{1}{z}$ counting from the heavy path.

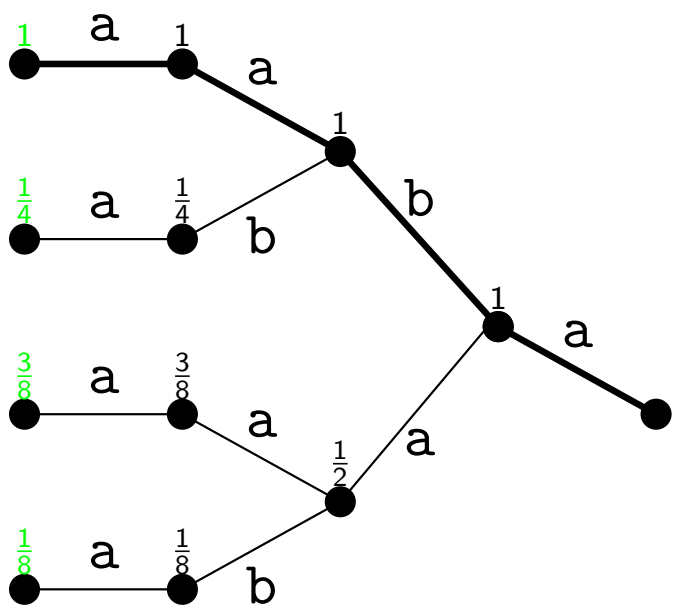


Trie construction

a $\frac{1}{4}$	a 1	a $\frac{3}{4}$	a $\frac{1}{2}$	a 1
b $\frac{3}{4}$	b 0	b $\frac{1}{4}$	b $\frac{1}{2}$	b 0

Level-by-level **construction**: for each node:

1. extend to the left using every possible character,
2. trim leaves with prob. $< \frac{1}{z}$ counting from the heavy path.

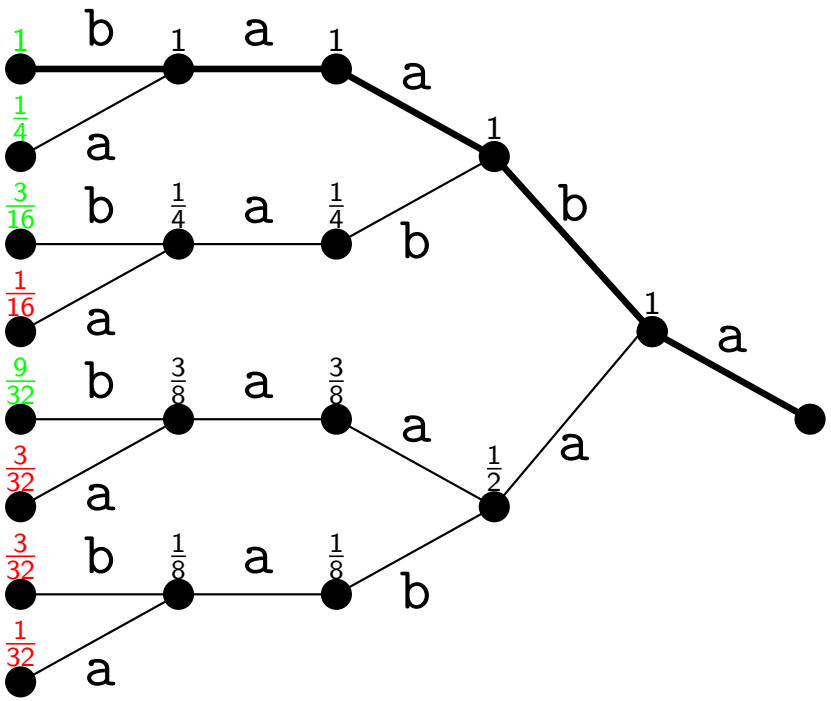


Trie construction

a $\frac{1}{4}$	a 1	a $\frac{3}{4}$	a $\frac{1}{2}$	a 1
b $\frac{3}{4}$	b 0	b $\frac{1}{4}$	b $\frac{1}{2}$	b 0

Level-by-level **construction**: for each node:

1. extend to the left using every possible character,
2. trim leaves with prob. $< \frac{1}{z}$ counting from the heavy path.

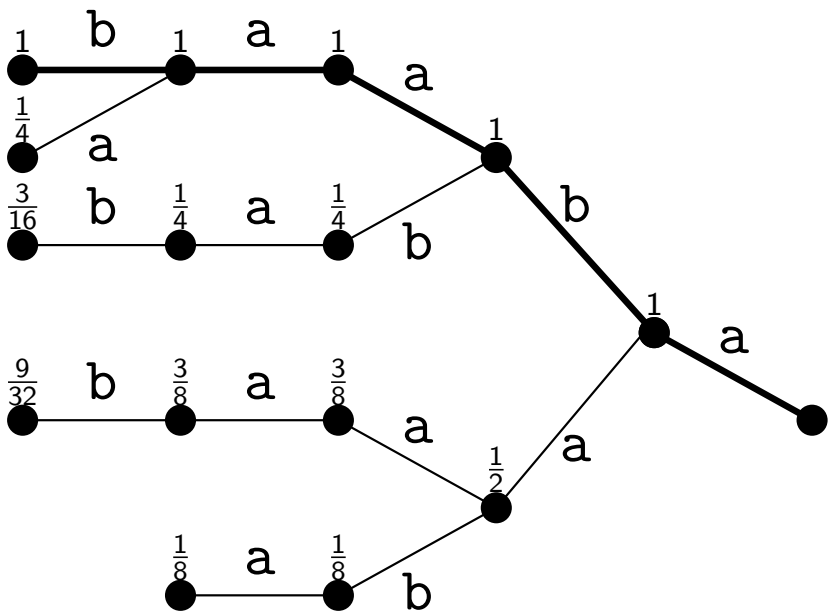


Trie construction

a $\frac{1}{4}$	a 1	a $\frac{3}{4}$	a $\frac{1}{2}$	a 1
b $\frac{3}{4}$	b 0	b $\frac{1}{4}$	b $\frac{1}{2}$	b 0

Level-by-level **construction**: for each node:

1. extend to the left using every possible character,
2. trim leaves with prob. $< \frac{1}{z}$ counting from the heavy path.

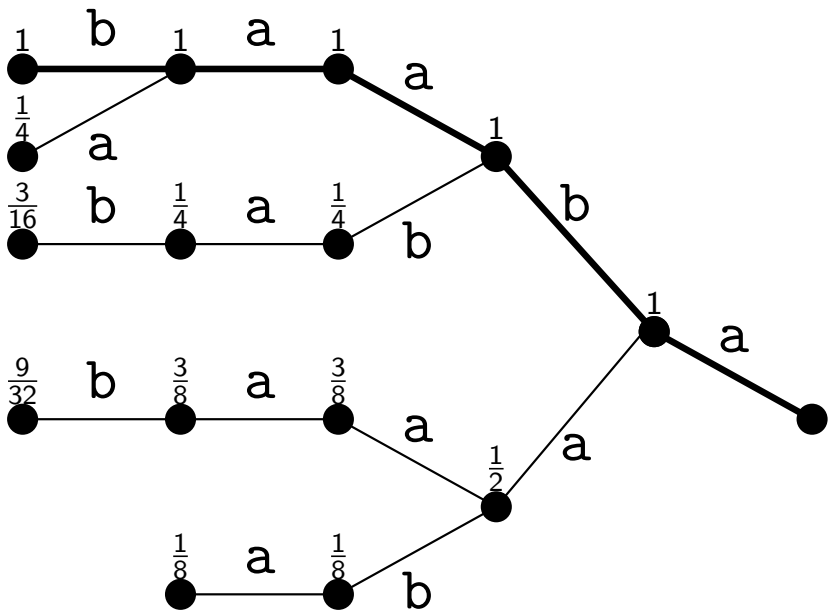


Trie construction

a $\frac{1}{4}$	a 1	a $\frac{3}{4}$	a $\frac{1}{2}$	a 1
b $\frac{3}{4}$	b 0	b $\frac{1}{4}$	b $\frac{1}{2}$	b 0

Level-by-level **construction**: for each node:

1. extend to the left using every possible character,
2. trim leaves with prob. $< \frac{1}{z}$ counting from the heavy path.



Upper bounds for property matching:
generate the bijection via a left-to-right traversal.

Weighted index construction

Suffix tree of a trie:
Compressed trie of paths from each node of the trie to the root.

Weighted index construction

Suffix tree of a trie:
Compressed trie of paths from each node of the trie to the root.

Theorem [Shibuya, 2003]:
Suffix tree of a trie can be computed in linear time.

Weighted index construction

Suffix tree of a trie:
Compressed trie of paths from each node of the trie to the root.

Theorem [Shibuya, 2003]:
Suffix tree of a trie can be computed in linear time.

Property suffix tree of a trie:
Compressed trie of trimmed paths from each node of the trie.

Weighted index construction

Suffix tree of a trie:
Compressed trie of paths from each node of the trie to the root.

Theorem [Shibuya, 2003]:
Suffix tree of a trie can be computed in linear time.

Property suffix tree of a trie:
Compressed trie of trimmed paths from each node of the trie.

Suffix tree (of a trie) $\rightarrow$ property suffix tree (of a trie)
Each terminal node needs to be moved up.

Weighted index construction

Suffix tree of a trie:
Compressed trie of paths from each node of the trie to the root.

Theorem [Shibuya, 2003]:
Suffix tree of a trie can be computed in linear time.

Property suffix tree of a trie:
Compressed trie of trimmed paths from each node of the trie.

Suffix tree (of a trie) $\rightarrow$ property suffix tree (of a trie)
Each terminal node needs to be moved up.
Many solutions, e.g.:

1. Use off-line weighted level ancestors
   - linear-time using the special case of union-find.
2. Trim everything without a new terminal node in the subtree.

Answering queries

Queries for a pattern P of length m.

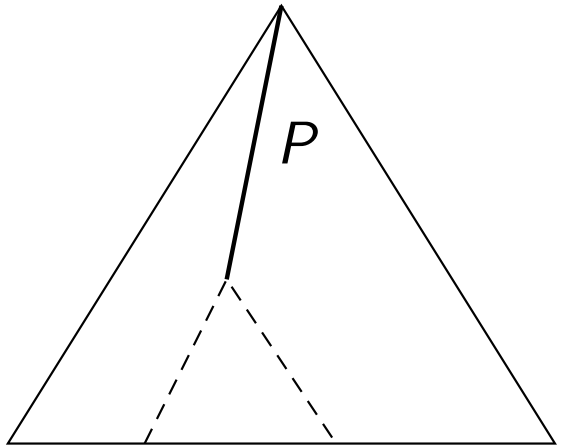
Answering queries

Queries for a pattern P of length m.

Decision version: $\mathcal{O}(m)$ time

- traverse the suffix tree with P.

Property suffix tree:



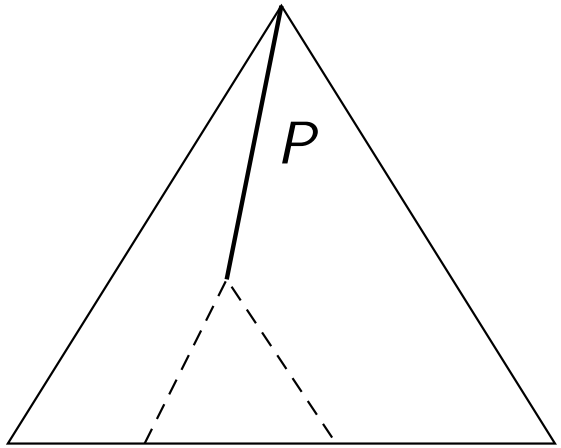
Answering queries

Queries for a pattern P of length m.

Decision version: $\mathcal{O}(m)$ time

- traverse the suffix tree with P.

Property suffix tree:



Reporting: $\mathcal{O}(m + |\mathrm{Occ}|)$ time

- report all terminal nodes in the subtree?

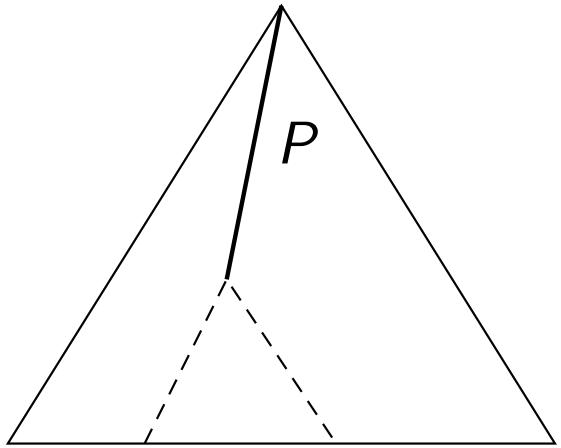
Answering queries

Queries for a pattern P of length m.

Decision version: $\mathcal{O}(m)$ time

- traverse the suffix tree with P.

Property suffix tree:

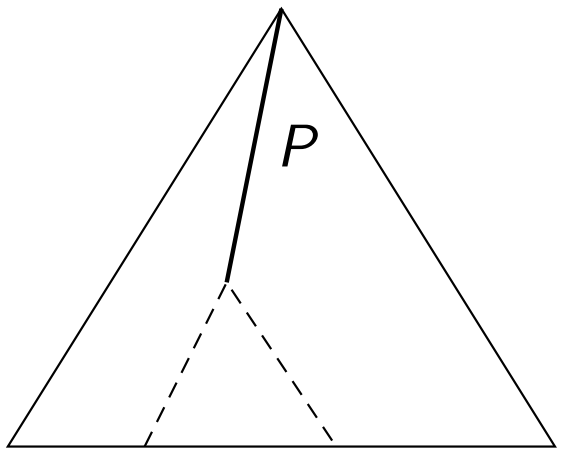


Reporting: $\mathcal{O}(m + |\mathrm{Occ}|)$ time

- report all terminal nodes in the subtree? Repeating occs.

Answering queries

Property suffix tree:



Queries for a pattern P of length m.

Decision version: $\mathcal{O}(m)$ time

- traverse the suffix tree with P.

Reporting: $\mathcal{O}(m + |\mathrm{Occ}|)$ time

- report terminal nodes with distinct starting positions.

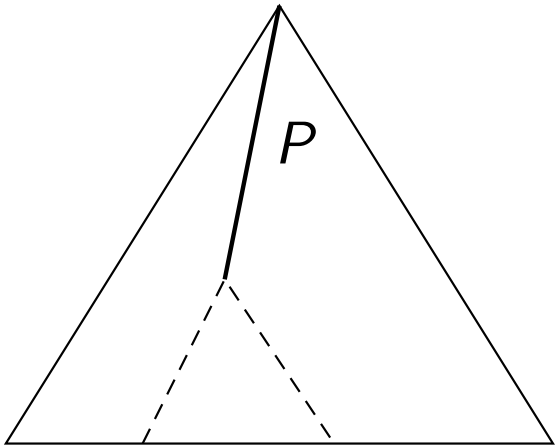
Answering queries

Queries for a pattern P of length m.

Decision version: $\mathcal{O}(m)$ time

- traverse the suffix tree with P.

Property suffix tree:



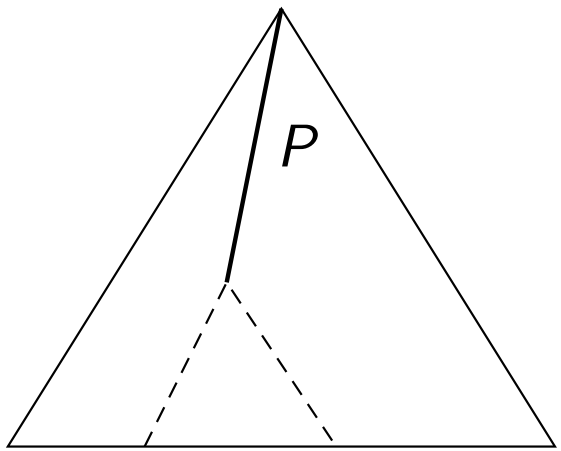
Reporting: $\mathcal{O}(m + |\mathrm{Occ}|)$ time

- report terminal nodes with distinct starting positions.

Lemma [Muthukrishnan] Colored range listing:
Report distinct elements in a given range of a preprocessed array.

Answering queries

Property suffix tree:



Queries for a pattern P of length m.

Decision version: $\mathcal{O}(m)$ time

- traverse the suffix tree with P.

Reporting: $\mathcal{O}(m + |\mathrm{Occ}|)$ time

- report terminal nodes with distinct starting positions.

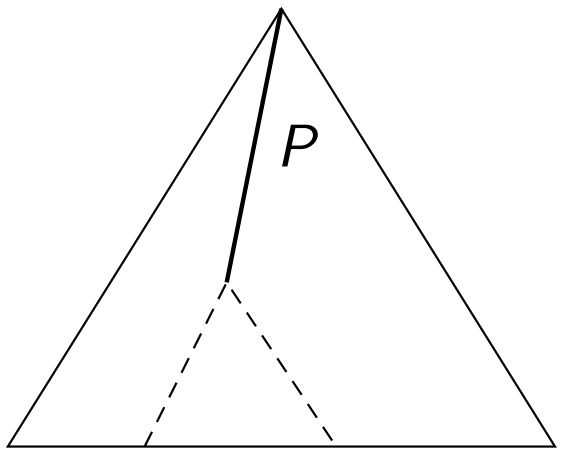
Lemma [Muthukrishnan] Colored range listing:
Report distinct elements in a given range of a preprocessed array.

Counting: $\mathcal{O}(m)$ time

- count distinct starting positions in the subtree

Answering queries

Property suffix tree:



Queries for a pattern P of length m.

Decision version: $\mathcal{O}(m)$ time

- traverse the suffix tree with P.

Reporting: $\mathcal{O}(m + |\mathrm{Occ}|)$ time

- report terminal nodes with distinct starting positions.

Lemma [Muthukrishnan] Colored range listing:
Report distinct elements in a given range of a preprocessed array.

Counting: $\mathcal{O}(m)$ time

- count distinct starting positions in the subtree

Lemma [Hui] Colored set size:
Count the number of distinct leaf colors for each node (offline).

Applications

Longest common extension:

$\mathrm{LCE}(i, j) =$ length of the longest solid factor occurring in i and j

Longest common extension:

$\mathrm{LCE}(i, j) =$ length of the longest solid factor occurring in i and j

Example: For $\frac{1}{z} = \frac{1}{8}$, $\mathrm{LCE}(1, 2) = 3$ (aaa).

a $\frac{1}{4}$	a 1	a $\frac{3}{4}$	a $\frac{1}{2}$	a 1
b $\frac{3}{4}$	b 0	b $\frac{1}{4}$	b $\frac{1}{2}$	b 0

Applications

Longest common extension:
$\mathrm{LCE}(i, j) =$ length of the longest solid factor occurring in i and j

Example: For $\frac{1}{z} = \frac{1}{8}$, $\mathrm{LCE}(1, 2) = 3$ (aaa).

a $\frac{1}{4}$	a 1	a $\frac{3}{4}$	a $\frac{1}{2}$	a 1
b $\frac{3}{4}$	b 0	b $\frac{1}{4}$	b $\frac{1}{2}$	b 0

Solution: $\mathcal{O}(z)$-time queries after $\mathcal{O}(nz)$-time preprocessing

1. extract terminal nodes with labels i and j.
2. compute LCA of every adjacent two with distinct labels.

Applications

Longest common extension:
$\mathrm{LCE}(i, j) =$ length of the longest solid factor occurring in i and j

Example: For $\frac{1}{z} = \frac{1}{8}$, $\mathrm{LCE}(1, 2) = 3$ (aaa).

a $\frac{1}{4}$	a 1	a $\frac{3}{4}$	a $\frac{1}{2}$	a 1
b $\frac{3}{4}$	b 0	b $\frac{1}{4}$	b $\frac{1}{2}$	b 0

Solution: $\mathcal{O}(z)$-time queries after $\mathcal{O}(nz)$-time preprocessing

1. extract terminal nodes with labels i and j.
2. compute LCA of every adjacent two with distinct labels.

Weighted prefix table: $\mathcal{O}(nz)$ time
$\mathrm{PREF}[i] = \mathrm{LCE}(1, i)$.

Applications

Longest common extension:
$\mathrm{LCE}(i, j) =$ length of the longest solid factor occurring in i and j

Example: For $\frac{1}{z} = \frac{1}{8}$, $\mathrm{LCE}(1, 2) = 3$ (aaa).

a $\frac{1}{4}$	a 1	a $\frac{3}{4}$	a $\frac{1}{2}$	a 1
b $\frac{3}{4}$	b 0	b $\frac{1}{4}$	b $\frac{1}{2}$	b 0

Solution: $\mathcal{O}(z)$-time queries after $\mathcal{O}(nz)$-time preprocessing

1. extract terminal nodes with labels i and j.
2. compute LCA of every adjacent two with distinct labels.

Weighted prefix table: $\mathcal{O}(nz)$ time
$\mathrm{PREF}[i] = \mathrm{LCE}(1, i)$.

Covers: $\mathcal{O}(nz)$ time
solid strings whose occurrences cover all positions of the text.

Summary and open problems

Index for weighted pattern matching:

- $\mathcal{O}(nz)$ space,
- $\mathcal{O}(nz)$ construction time,
- $\mathcal{O}(m + |\mathrm{Occ}|)$ query time (reporting),
- $\mathcal{O}(m)$ query time (counting).

Summary and open problems

Index for weighted pattern matching:

- $\mathcal{O}(nz)$ space,
- $\mathcal{O}(nz)$ construction time,
- $\mathcal{O}(m + |\mathrm{Occ}|)$ query time (reporting),
- $\mathcal{O}(m)$ query time (counting).

Open problems:

- support queries with $z' < z$,
  - Biswas et al. (EDBT'16) achieve this for $z = \mathcal{O}(1)$.
- support weighted patterns in $o(z(m + |\mathrm{Occ}|))$ time.

Questions?

Thank you for your attention!