Faster Approximate Pattern Matching: A Unified Approach

Panagiotis Charalampopoulos
King's College London
University of Warsaw

Tomasz Kociumaka
Bar-Ilan University
↓
UC Berkeley

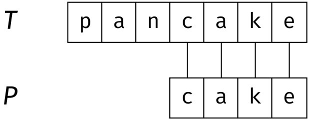
Philip Wellnitz
MPII, SIC

FOCS 2020

Approximate Pattern Matching

Pattern Matching

Given a text T and a pattern P, identify the occurrences of P as a substring of T.

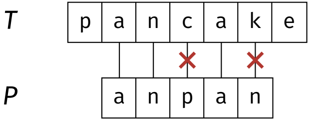


Finding cake

Approximate Pattern Matching

Pattern Matching with Mismatches

Given a text T, a pattern P, and an integer k, identify the length-$|P|$ substrings of T with **Hamming distance** at most k to P.

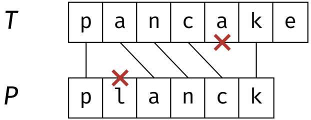


Finding anpan, $k = 2$

Approximate Pattern Matching

Pattern Matching with Errors

Given a text T, a pattern P, and an integer k, identify the (starting positions of) substrings of T that are at **edit distance** of at most k to P.



Finding `planck`, $k = 2$

What if the text and the pattern are given
in a compressed representation?

anpanisoneofthepopularjapanesesweetbunwithsweetbeaninthecentertodaytherearemanytypesofsweetbeencenteredintheanpanforexamplegomaanshiroanuguisuankurianandetcbuttheoriginalofitisthenormalankomadewithredbean

anpanisajapanesesweetrollmostcommonlyfilledwithredbeanpasteanpancanalsobepreparedwithotherfillingsincludingwhitebeansgreenbeanssesameandchestnut

Grammar Compression

Grammar Compression

For a string T, a grammar compression of T is a context-free grammar G_T that generates $\{T\}$. The grammar G_T is wlog. a straight-line program or SLP.

Known Results

Problem	uncompressed text and pattern	SLP text and pattern $n = \Omega(\log N)$, $m = \Omega(\log M)$
Pattern Matching	$O(N+M)$ [KMP'77]	$\tilde{O}(n+m)$ [Jeż'15]
PM with k Mismatches	$\tilde{O}(N + k^2 \cdot N/M)$, $\tilde{O}(N + kN/\sqrt{M})$ [CFPSS'16] [GU'18]	$\tilde{O}(nk^4 + Mk)$ [BKW'19]
PM with k Errors	$O(N + k^4 \cdot N/M)$ [CH'02]	$O(nm \operatorname{poly}(k))$ [BLRSSW'15]

N: length of uncompressed text
n: size of compressed text
k: number of mismatches/errors
M: length of uncompressed pattern
m: size of compressed pattern

Known Results

Problem	uncompressed text and pattern	SLP text and pattern $n = \Omega(\log N)$, $m = \Omega(\log M)$
Pattern Matching	$O(N + M)$ [KMP'77]	$\tilde{O}(n + m)$ [Jeż'15]
PM with k Mismatches	$\tilde{O}(N + k^2 \cdot N/M)$, $\tilde{O}(N + kN/\sqrt{M})$ [CFPSS'16] [GU'18]	$\tilde{O}(nk^4 + Mk)$ [BK**W**'19]
PM with k Errors	$O(N + k^4 \cdot N/M)$ [CH'02]	$O(nm \operatorname{poly}(k))$ [BLRSSW'15]

N: length of uncompressed text
n: size of compressed text
k: number of mismatches/errors

M: length of uncompressed pattern
m: size of compressed pattern

Known Results

Problem	uncompressed text and pattern	SLP text and pattern $n = \Omega(\log N)$, $m = \Omega(\log M)$
Pattern Matching	$O(N+M)$ [KMP'77]	$\tilde{O}(n+m)$ [Jeż'15]
PM with k Mismatches	$\tilde{O}(N + k^2 \cdot N/M)$, $\tilde{O}(N + kN/\sqrt{M})$ [CFPSS'16] [GU'18]	$\tilde{O}(nk^4 + Mk)$ [BK**W**'19]
PM with k Errors	$O(N + k^4 \cdot N/M)$ [CH'02]	$O(nm \operatorname{poly}(k))$ [BLRSSW'15]

N: length of uncompressed text
n: size of compressed text
k: number of mismatches/errors

M: length of uncompressed pattern
m: size of compressed pattern

Main Results

Problem	uncompressed text and pattern	SLP text and pattern $n = \Omega(\log N)$, $m = \Omega(\log M)$
Pattern Matching	$O(N + M)$ [KMP'77]	$\tilde{O}(n + m)$ [Jeż'15]
PM with k Mismatches	$\tilde{O}(N + k^2 \cdot N/M)$, $\tilde{O}(N + kN/\sqrt{M})$ [CFPSS'16] [GU'18]	~~$\tilde{O}(nk^4 + Mk)$~~ $\tilde{O}(nk^2 + m)$
PM with k Errors	$O(N + k^4 \cdot N/M)$ [CH'02]	~~$O(nm\,\mathrm{poly}(k))$~~ $\tilde{O}(nk^4 + m)$

N: length of uncompressed text
n: size of compressed text
k: number of mismatches/errors
M: length of uncompressed pattern
m: size of compressed pattern

Main Results

Problem	uncompressed text and pattern	SLP text and pattern $n = \Omega(\log N)$, $m = \Omega(\log M)$
Pattern Matching	$O(N + M)$ [KMP'77]	$\tilde{O}(n + m)$ [Jeż'15]
PM with k Mismatches	$\tilde{O}(N + k^2 \cdot N/M)$, $\tilde{O}(N + kN/\sqrt{M})$ $\tilde{O}(N + k^2 \cdot N/M)$	$\cancel{\tilde{O}(nk^4 + Mk)}$ $\tilde{O}(nk^2 + m)$
PM with k Errors	$O(N + k^4 \cdot N/M)$ $O(N + k^4 \cdot N/M)$	$\cancel{O(nm\,\mathrm{poly}(k))}$ $\tilde{O}(nk^4 + m)$

N: length of uncompressed text
n: size of compressed text
k: number of mismatches/errors

M: length of uncompressed pattern
m: size of compressed pattern

Main Results

Problem	uncompressed text and pattern	SLP text and pattern $n = \Omega(\log N)$, $m = \Omega(\log M)$
Pattern Matching	$O(N + M)$ [KMP'77]	$\tilde{O}(n + m)$ [Jeż'15]
PM with k Mismatches	$\tilde{O}(N + k^2 \cdot N/M)$, $\tilde{O}(N + kN/\sqrt{M})$ $\tilde{O}(N + k^2 \cdot N/M)$	$\cancel{\tilde{O}(nk^4 + Mk)}$ $\tilde{O}(nk^2 + m)$
PM with k Errors	$O(N + k^4 \cdot N/M)$ $O(N + k^4 \cdot N/M)$	$\cancel{O(nm\,\mathrm{poly}(k))}$ $\tilde{O}(nk^4 + m)$

Improvements obtained via improved/new structural insight in solution structure.

Known Structural Results for PM with Mismatches

Theorem [BKW'19]

Given a pattern P of length M, a text T of length $N \leq {}^3/_2\, M$, and a threshold $k \leq M$, at least one of the following holds:

- The number of k-mismatch occurrences of P in T is at most $O(k^2)$.
- The pattern P is almost periodic (at HD $\leq 6k$ to a string Q with period $O(M/k)$).

Known Structural Results for PM with Mismatches

Theorem [BKW'19]

Given a pattern P of length M, a text T of length $N \le {}^{3}\!/_{2}\, M$, and a threshold $k \le M$, at least one of the following holds:

- The number of k-mismatch occurrences of P in T is at most $O(k^2)$.
- The pattern P is almost periodic (at HD $\le 6k$ to a string Q with period $O(M/k)$).

- Additional insights into the almost periodic case

Structural Results for PM with Mismatches

Main Structural Theorem (HD)

Given a pattern P of length M, a text T of length $N \le {}^3\!/_2\, M$, and a threshold $k \le M$, at least one of the following holds:

- The number of k-mismatch occurrences of P in T is at most $\cancel{O(k^2)}$ $O(k)$.
- The pattern P is almost periodic (at HD $\cancel{\le 6k}$ $< 2k$ to a string Q with period $O(M/k)$).

Structural Results for PM with Mismatches

Main Structural Theorem (HD)

Given a pattern P of length M, a text T of length $N \leq {}^3/_2\, M$, and a threshold $k \leq M$, at least one of the following holds:

- The number of k-mismatch occurrences of P in T is at most $\cancel{O(k^2)}$ $O(k)$.
- The pattern P is almost periodic (at HD $\cancel{\leq 6k}$ $< 2k$ to a string Q with period $O(M/k)$).

Are these bounds optimal?

Structural Results for PM with Mismatches

Main Structural Theorem (HD)

Given a pattern P of length M, a text T of length $N \le {}^3\!/_2\, M$, and a threshold $k \le M$, at least one of the following holds:

- The number of k-mismatch occurrences of P in T is at most $\cancel{O(k^2)}$ $O(k)$.
- The pattern P is almost periodic (at HD $\cancel{\le 6k}$ $< 2k$ to a string Q with period $O(M/k)$).

Are these bounds optimal?—Yes, see long version.

What about PM with errors?

Structural Results for PM with Errors

Main Structural Theorem (HD)

Given a pattern P of length M, a text T of length $N \le {}^3\!/_2\, M$, and a threshold $k \le M$, at least one of the following holds:

- The number of k-mismatch occurrences of P in T is at most $O(k)$.
- The pattern P is almost periodic (at HD $< 2k$ to a string Q with period $O(M/k)$).

Main Structural Theorem (ED)

Given a pattern P of length M, a text T of length $N \le {}^3\!/_2\, M$, and a threshold $k \le M$, at least one of the following holds:

- The number of (starting positions of) k-error occurrences of P in T is at most $O(k^2)$.
- The pattern P is almost periodic (at ED $< 2k$ to a string Q with period $O(M/k)$).

Structural Results for PM with Errors

Main Structural Theorem (HD)

Given a pattern P of length M, a text T of length $N \le {}^{3}\!/_{2}\, M$, and a threshold $k \le M$, at least one of the following holds:

- The number of k-mismatch occurrences of P in T is at most $O(k)$.
- The pattern P is almost periodic (at HD < $2k$ to a string Q with period $O(M/k)$).

Main Structural Theorem (ED)

Given a pattern P of length M, a text T of length $N \le {}^{3}\!/_{2}\, M$, and a threshold $k \le M$, at least one of the following holds:

- The starting positions of all k-error occurrences of P in T lie in $O(k)$ intervals of length $O(k)$ each.
- The pattern P is almost periodic (at ED < $2k$ to a string Q with period $O(M/k)$).

In this talk, Hamming distance only.

Proofs for edit distance work similarly.

Main Structural Theorem (HD)

Given a pattern P of length M, a text T of length $N \le {}^{3}\!/_{2}\, M$, and a threshold $k \le M$, at least one of the following holds:

- The number of k-mismatch occurrences of P in T is at most $O(k)$.
- The pattern P is almost periodic (at HD $< 2k$ to a string with period $O(M/k)$).

Main Structural Theorem (HD)

Given a pattern P of length M, a text T of length $N \le {}^3\!/_2\, M$, and a threshold $k \le M$, at least one of the following holds:

- The number of k-mismatch occurrences of P in T is at most $O(k)$.
- The pattern P is almost periodic (at HD $< 2k$ to a string with period $O(M/k)$).

Key Lemma (Analyze)

For each string P of length M, at least one of the following holds:

- P contains $2k$ disjoint breaks; each break has length $M/8k$ and period $> M/128k$.
- P contains disjoint repetitive regions R_i with total length $\ge 3/8 \cdot M$;
  each region has length $\ge M/8k$ and is almost periodic with HD exactly $8k/M \cdot |R_i|$.
- P is almost periodic (at HD $< 8k$ to a string with period $M/128k$).

Main Structural Theorem (HD)

Given a pattern P of length M, a text T of length $N \le {}^3/_2\, M$, and a threshold $k \le M$, at least one of the following holds:

- The number of k-mismatch occurrences of P in T is at most $O(k)$.
- The pattern P is almost periodic (at HD < $2k$ to a string with period $O(M/k)$).

Key Lemma (Analyze)

For each string P of length M, at least one of the following holds:

- P contains $2k$ disjoint breaks; each break has length $M/8k$ and period $> M/128k$.
- P contains disjoint repetitive regions R_i with total length $\ge 3/8 \cdot M$; each region has length $\ge M/8k$ and is almost periodic with HD exactly $8k/M \cdot |R_i|$.
- P is almost periodic (at HD < $8k$ to a string with period $M/128k$).

Long version:

- Key Lemma $\Longrightarrow$ Main Structural Theorem
- Proof idea for Key Lemma

How do we turn the structural insights into algorithms?

Obtaining Faster Algorithms

- All of our algorithms rely on a small set of essential operations:
  - `LCP`(S, T): Compute the length of the longest common prefix of S and T.
  - `LCP`^{R}(S, T): Compute the length of the longest common suffix of S and T.
  - `IPM`(P, T): Compute all exact matches of P in T.
  - `Length`(S): Compute the length $|S|$ of S.
  - `Access`(S, i): Retrieve the character $S[i]$.
  - `Extract`(S, ℓ, r): Extract the fragment (or substring) $S[\ell\,..\,r)$ from S.

The PILLAR Model

- All of our algorithms rely on a small set of essential operations:
  - LC**P**(S, T): Compute the length of the longest common prefix of S and T.
  - **L**CP$^R(S, T)$: Compute the length of the longest common suffix of S and T.
  - **I**PM(P, T): Compute all exact matches of P in T.
  - **L**ength(S): Compute the length $|S|$ of S.
  - **A**ccess(S, i): Retrieve the character $S[i]$.
  - Ext**R**act(S, ℓ, r): Extract the fragment (or substring) $S[\ell\,..\,r)$ from S.

The PILLAR Model

- All of our algorithms rely on a small set of essential operations:
  - $\texttt{LCP}(S, T)$: Compute the length of the longest common prefix of S and T.
  - $\texttt{LCP}^R(S, T)$: Compute the length of the longest common suffix of S and T.
  - $\texttt{IPM}(P, T)$: Compute all exact matches of P in T.
  - $\texttt{Length}(S)$: Compute the length $|S|$ of S.
  - $\texttt{Access}(S, i)$: Retrieve the character $S[i]$.
  - $\texttt{ExtRact}(S, \ell, r)$: Extract the fragment (or substring) $S[\ell\,..\,r)$ from S.

Theorem

Fast PILLAR Algorithm

+

Theorem

Fast PILLAR Implementation

=

Theorem

Fast Algorithm

The PILLAR Model

- All of our algorithms rely on a small set of essential operations:
  - LC**P**(S, T): Compute the length of the longest common prefix of S and T.
  - **L**CP^R(S, T): Compute the length of the longest common suffix of S and T.
  - **I**PM(P, T): Compute all exact matches of P in T.
  - **L**ength(S): Compute the length $|S|$ of S.
  - **A**ccess(S, i): Retrieve the character $S[i]$.
  - Ext**R**act(S, ℓ, r): Extract the fragment (or substring) $S[\ell\,..\,r)$ from S.

Theorem		**Theorem**		**Theorem**
Fast PILLAR Algorithm	+	Fast PILLAR Implementation	=	Fast Algorithm

This work

The PILLAR Model

- All of our algorithms rely on a small set of essential operations:
  - `LC`**`P`**`(`S`, `T`)`: Compute the length of the longest common prefix of S and T.
  - **`L`**`CP`^{R}`(`S`, `T`)`: Compute the length of the longest common suffix of S and T.
  - **`I`**`PM(`P`, `T`)`: Compute all exact matches of P in T.
  - **`L`**`ength(`S`)`: Compute the length $|S|$ of S.
  - **`A`**`ccess(`S`, `i`)`: Retrieve the character $S[i]$.
  - `Ext`**`R`**`act(`S`, `ℓ`, `r`)`: Extract the fragment (or substring) $S[\ell\,..\,r)$ from S.

Theorem		**Theorem**		**Theorem**
Fast PILLAR Algorithm	+	Fast PILLAR Implementation	=	Fast Algorithm
This work		This work: Combination of known results for 3 settings *(see long version)*		

Obtaining Fast Algorithms: The Fully-Compressed Setting

The `PILLAR` operations: `LCP`, $\texttt{LCP}^R$, `IPM`, `Length`, `Access`, `Extract`

SLPs: G_P of size m generating pattern P, G_T of size n generating text T.

Obtaining Fast Algorithms: The Fully-Compressed Setting

The `PILLAR` operations: `LCP`, $\texttt{LCP}^R$, `IPM`, `Length`, `Access`, `Extract`

SLPs: G_P of size m generating pattern P, G_T of size n generating text T.

Using Recompression [Jeż'15], we can implement each operation in $O(\log^3(|P| + |T|))$ time.

(After $O((n + m)\log(|P| + |T|))$ preprocessing.)

Obtaining Fast Algorithms: The Fully-Compressed Setting

The PILLAR operations: LCP, LCP^R, IPM, Length, Access, Extract

SLPs: G_P of size m generating pattern P, G_T of size n generating text T.

Using Recompression [Jeż'15], we can implement each operation in $O(\log^3(|P| + |T|))$ time.

(After $O((n + m)\log(|P| + |T|))$ preprocessing.)

Theorem (Algorithm for PM w/ Mism.)

For any positive threshold $k \leq |P|$, we can compute the number of all k-mismatch occ's of P in T in time $O(m \log(|P| + |T|) + nk^2 \log^3(|P| + |T|))$.

(Reporting of all occ's takes time linear in the number of occ's.)

Theorem (Algorithm for PM w/ Errors)

For any positive threshold $k \leq |P|$, we can compute the number of all k-error occ's of P in T in time $O(m \log(|P| + |T|) + nk^4 \log^3(|P| + |T|))$.

(Reporting of all occ's takes time linear in the number of occ's.)

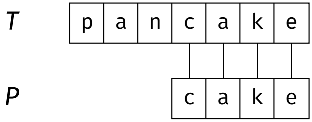
Thank You!

Full paper: `arxiv.org/abs/2004.08350`

Approximate Pattern Matching

Pattern Matching

Given a text T and a pattern P, identify the occurrences of P as a substring of T.

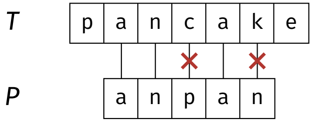


Finding cake

Approximate Pattern Matching

Pattern Matching with Mismatches

Given a text T, a pattern P, and an integer k, identify the length-$|P|$ substrings of T with **Hamming distance** at most k to P.



Finding anpan, $k = 2$

Approximate Pattern Matching

Pattern Matching with Mismatches

Given a text T, a pattern P, and an integer k, identify the length-$|P|$ substrings of T with **Hamming distance** at most k to P.

Thm. [Clifford et al.'16]

Pattern matching with k mismatches on a text of length N and a pattern of length M can be solved in time $\tilde{O}(N + k^2 \cdot N/M)$.

Approximate Pattern Matching

Pattern Matching with Mismatches

Given a text T, a pattern P, and an integer k, identify the length-$|P|$ substrings of T with **Hamming distance** at most k to P.

Thm. [Gawrychowski, Uznański'18]

Pattern matching with k mismatches on a text of length N and a pattern of length M can be solved in time $\tilde{O}(N + kN/\sqrt{M})$.

Approximate Pattern Matching

Pattern Matching with Mismatches

Given a text T, a pattern P, and an integer k, identify the length-$|P|$ substrings of T with **Hamming distance** at most k to P.

Thm. [Gawrychowski, Uznański'18]

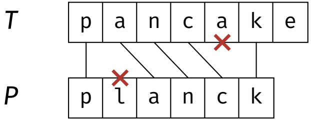
Pattern matching with k mismatches on a text of length N and a pattern of length M can be solved in time $\tilde{O}(N + kN/\sqrt{M})$.

Matching (conditional) lower bound (for combinatorial algorithms) [GU'18]

Approximate Pattern Matching

Pattern Matching with Errors

Given a text T, a pattern P, and an integer k, identify the (starting positions of) substrings of T that are at **edit distance** of at most k to P.



Finding `planck`, $k = 2$

Approximate Pattern Matching

Pattern Matching with Errors

Given a text T, a pattern P, and an integer k, identify the (starting positions of) substrings of T that are at **edit distance** of at most k to P.

Thm. [Cole, Hariharan'02]

Pattern matching with k errors on a text of length N and a pattern of length M can be solved in time $\tilde{O}(N + k^4 \cdot N/M)$.

Approximate Pattern Matching

Pattern Matching with Errors

Given a text T, a pattern P, and an integer k, identify the (starting positions of) substrings of T that are at **edit distance** of at most k to P.

Thm. [Cole, Hariharan'02]

Pattern matching with k errors on a text of length N and a pattern of length M can be solved in time $\tilde{O}(N + k^4 \cdot N/M)$.

$\Omega(N + k^2 \cdot N/M)$ conditional lower bound [Backurs, Indyk'18]

Approximate Pattern Matching

Pattern Matching with Errors

Given a text T, a pattern P, and an integer k, identify the (starting positions of) substrings of T that are at **edit distance** of at most k to P.

Thm. [Cole, Hariharan'02]

Pattern matching with k errors on a text of length N and a pattern of length M can be solved in time $\tilde{O}(N + k^4 \cdot N/M)$.

$\Omega(N + k^2 \cdot N/M)$ conditional lower bound [Backurs, Indyk'18]

Not in this talk: How to shrink the gap.

What if the text and the pattern are *huge*?

o	f	s	w	e	e	t	b	e	e	n	c	e	n	t	e	r	e	d	i	n	t	h	e	a	n	p	a	n	f	o	r	e	x	a	m

n	p	a	s	t	e	a	n	p	a	n	c	a	n	a	l	s	o	b	e	p	r	e	p	a	r	e	d	w	i	t	h	o	t	h	e

What if the text and the pattern are *huge*?

anpanisoneofthepopularjapanesesweetbunwithsweetbeaninthecentertodaytherearemanytypesofsweetbeencenteredintheanpanforexamplegomaanshiroanuguisuankurianandetcbuttheoriginalofitisthenormalankomadewithredbean

anpanisajapanesesweetrollmostcommonlyfilledwithredbeanpasteanpancanalsobepreparedwithotherfillingsincludingwhitebeansgreenbeanssesameandchestnut

What if the text and the pattern are given in a compressed representation?

anpanisoneofthepopularjapanesesweetbunwithsweetbeaninthecentertodaytherearemanytypesofsweetbeencenteredintheanpanforexamplegomaanshiroanuguisuankurianandetcbuttheoriginalofitisthenormalankomadewithredbean

anpanisajapanesesweetrollmostcommonlyfilledwithredbeanpasteanpancanalsobepreparedwithotherfillingsincludingwhitebeansgreenbeanssesameandchestnut

Grammar Compression

Grammar Compression

For a string T, a grammar compression of T is a context-free grammar G_T that generates $\{T\}$. The grammar G_T is wlog. a straight-line program or SLP.

Grammar Compression

Straight-Line Program (SLP)

An SLP G_T is a set of non-terminals $\{T_1, \dots, T_n\}$ and productions of the form $T_i \to a, a \in \Sigma$ or $T_i \to T_\ell T_r$, where $\ell, r < i$. The starting symbol is T_n.

Grammar Compression

Straight-Line Program (SLP)

An SLP G_T is a set of non-terminals $\{T_1, \dots, T_n\}$ and productions of the form $T_i \to a, a \in \Sigma$ or $T_i \to T_\ell T_r$, where $\ell, r < i$. The starting symbol is T_n.

$T_1 \to \text{a}; \quad T_2 \to \text{n}; \quad T_3 \to \text{p}$

$$\begin{array}{ccc} T_1 & T_2 & T_3 \\ | & | & | \\ \text{a} & \text{n} & \text{p} \end{array}$$

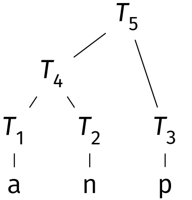
Grammar Compression

Straight-Line Program (SLP)

An SLP G_T is a set of non-terminals $\{T_1, \ldots, T_n\}$ and productions of the form $T_i \to a, a \in \Sigma$ or $T_i \to T_\ell T_r$, where $\ell, r < i$. The starting symbol is T_n.

$$T_1 \to \mathrm{a}; \quad T_2 \to \mathrm{n}; \quad T_3 \to \mathrm{p}$$

$$T_4 \to T_1 T_2; \qquad T_5 \to T_4 T_3$$



Grammar Compression

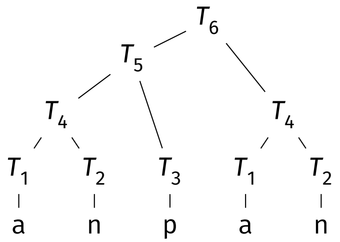
Straight-Line Program (SLP)

An SLP G_T is a set of non-terminals $\{T_1, \ldots, T_n\}$ and productions of the form $T_i \to a, a \in \Sigma$ or $T_i \to T_\ell T_r$, where $\ell, r < i$. The starting symbol is T_n.

$$T_1 \to \mathrm{a}; \quad T_2 \to \mathrm{n}; \quad T_3 \to \mathrm{p}$$

$$T_4 \to T_1 T_2; \qquad T_5 \to T_4 T_3$$

$$T_6 \to T_5 T_4$$

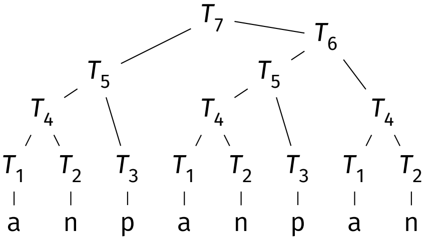


Grammar Compression

Straight-Line Program (SLP)

An SLP G_T is a set of non-terminals $\{T_1, \ldots, T_n\}$ and productions of the form $T_i \to a, a \in \Sigma$ or $T_i \to T_\ell T_r$, where $\ell, r < i$. The starting symbol is T_n.

$$T_1 \to \mathrm{a}; \quad T_2 \to \mathrm{n}; \quad T_3 \to \mathrm{p}$$
$$T_4 \to T_1 T_2; \qquad T_5 \to T_4 T_3$$
$$T_6 \to T_5 T_4; \qquad T_7 \to T_5 T_6$$



Structure of Exact Pattern Matching

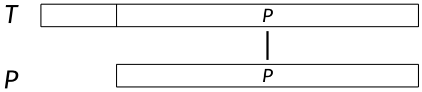
Fact (Folklore)

Let text T and pattern P, $|T| \leq {}^3/_2\,|P|$, be given such that there are ≥ 2 matches of P in T that together match T completely. Then, both P and T are periodic with some period x and every match of P in T starts at a position $1 + i \cdot |x|$.

Structure of Exact Pattern Matching

Fact (Folklore)

Let text T and pattern P, $|T| \leq {}^{3}\!/_{2}\, |P|$, be given such that there are ≥ 2 matches of P in T that together match T completely. Then, both P and T are periodic with some period x and every match of P in T starts at a position $1 + i \cdot |x|$.



Structure of Exact Pattern Matching

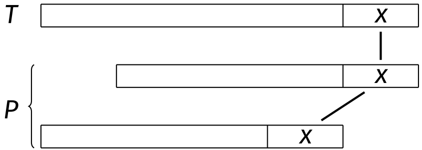
Fact (Folklore)

Let text T and pattern P, $|T| \leq {}^3/_2\,|P|$, be given such that there are ≥ 2 matches of P in T that together match T completely. Then, both P and T are periodic with some period x and every match of P in T starts at a position $1 + i \cdot |x|$.

Structure of Exact Pattern Matching

Fact (Folklore)

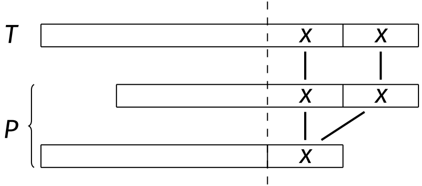
Let text T and pattern P, $|T| \le {}^3\!/_2\,|P|$, be given such that there are ≥ 2 matches of P in T that together match T completely. Then, both P and T are periodic with some period x and every match of P in T starts at a position $1 + i \cdot |x|$.



Structure of Exact Pattern Matching

Fact (Folklore)

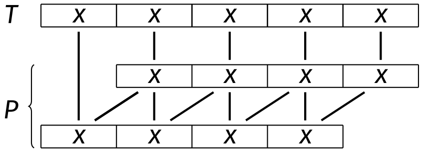
Let text T and pattern P, $|T| \leq {}^{3}\!/_{2}\,|P|$, be given such that there are ≥ 2 matches of P in T that together match T completely. Then, both P and T are periodic with some period x and every match of P in T starts at a position $1 + i \cdot |x|$.



Structure of Exact Pattern Matching

Fact (Folklore)

Let text T and pattern P, $|T| \leq {}^3/_2\,|P|$, be given such that there are ≥ 2 matches of P in T that together match T completely. Then, both P and T are periodic with some period x and every match of P in T starts at a position $1 + i \cdot |x|$.



Structural Results for PM with Mismatches

Main Structural Theorem (HD)

Given a pattern P of length M, a text T of length $N \le {}^3\!/_2\, M$, and a threshold $k \le M$, at least one of the following holds:

- The number of k-mismatch occurrences of P in T is at most $\cancel{O(k^2)}$ $O(k)$.
- The pattern P is almost periodic (at HD $\cancel{\le 6k}$ $< 2k$ to a string Q with period $O(M/k)$).

Structural Results for PM with Mismatches

Main Structural Theorem (HD)

Given a pattern P of length M, a text T of length $N \le {}^3\!/_2\, M$, and a threshold $k \le M$, at least one of the following holds:

- The number of k-mismatch occurrences of P in T is at most $\cancel{O(k^2)}$ $O(k)$.
- The pattern P is almost periodic (at HD $\cancel{\le 6k}$ $< 2k$ to a string Q with period $O(M/k)$).

Are these bounds optimal?

Structural Results for PM with Mismatches

Main Structural Theorem (HD)

Given a pattern P of length M, a text T of length $N \le {}^3/_2\, M$, and a threshold $k \le M$, at least one of the following holds:

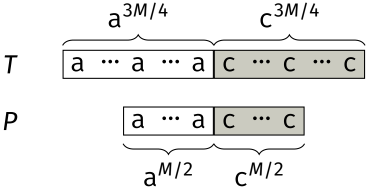
- The number of k-mismatch occurrences of P in T is at most $O(k)$.
- The pattern P is almost periodic (at HD < $2k$ to a string Q with period $O(M/k)$).

Structural Results for PM with Mismatches

Main Structural Theorem (HD)

Given a pattern P of length M, a text T of length $N \le {}^3\!/_2\, M$, and a threshold $k \le M$, at least one of the following holds:

- The number of k-mismatch occurrences of P in T is at most $O(k)$.
- The pattern P is almost periodic (at HD < $2k$ to a string Q with period $O(M/k)$).



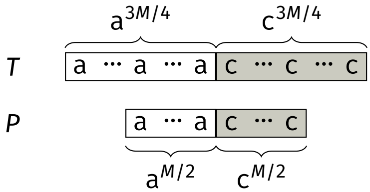
- Both P and T far from periodic, but there are $2k+1$ k-mismatch occurrences of P in T.

Structural Results for PM with Mismatches

Main Structural Theorem (HD)

Given a pattern P of length M, a text T of length $N \le {}^{3}\!/_{2}\, M$, and a threshold $k \le M$, at least one of the following holds:

- The number of k-mismatch occurrences of P in T is at most $O(k)$.
- The pattern P is almost periodic (at HD $< 2k$ to a string Q with period $O(M/k)$).



- Both P and T far from periodic, but there are $2k + 1$ k-mismatch occurrences of P in T.

Structural Results for PM with Mismatches

Main Structural Theorem (HD)

Given a pattern P of length M, a text T of length $N \le {}^{3}\!/_{2}\, M$, and a threshold $k \le M$, at least one of the following holds:

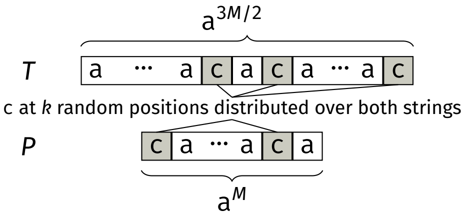
- The number of k-mismatch occurrences of P in T is at most $O(k)$.
- The pattern P is almost periodic (at HD $< 2k$ to a string Q with period $O(M/k)$).

Structural Results for PM with Mismatches

Main Structural Theorem (HD)

Given a pattern P of length M, a text T of length $N \le {}^{3}/_{2}\,M$, and a threshold $k \le M$, at least one of the following holds:

- The number of k-mismatch occurrences of P in T is at most $O(k)$.
- The pattern P is almost periodic (at HD $< 2k$ to a string Q with period $O(M/k)$).



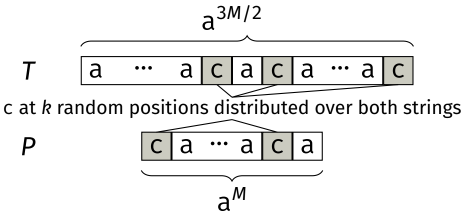
- Both P and T at HD up to k from periodic, and there are $M/2$ k-mismatch occurrences.

Structural Results for PM with Mismatches

Main Structural Theorem (HD)

Given a pattern P of length M, a text T of length $N \le {}^3/_2\, M$, and a threshold $k \le M$, at least one of the following holds:

- The number of k-mismatch occurrences of P in T is at most $O(k)$.
- The pattern P is almost periodic (at HD < $2k$ to a string Q with period $O(M/k)$).



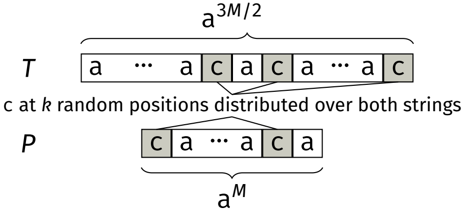
- Both P and T at HD up to k from periodic, and there are $M/2$ k-mismatch occurrences.

Structural Results for PM with Mismatches

Main Structural Theorem (HD)

Given a pattern P of length M, a text T of length $N \le {}^3\!/_2\, M$, and a threshold $k \le M$, at least one of the following holds:

- The number of k-mismatch occurrences of P in T is at most $O(k)$.
- The pattern P is almost periodic (at HD $< (1+\varepsilon)k$ to a string Q with period $O(M/k)$).



- Both P and T at HD up to k from periodic, and there are $M/2$ k-mismatch occurrences.

Main Structural Theorem (HD)

Given a pattern P of length M, a text T of length $N \le {}^3/_2\, M$, and a threshold $k \le M$, at least one of the following holds:

- The number of k-mismatch occurrences of P in T is at most $O(k)$.
- The pattern P is almost periodic (at HD $< 2k$ to a string with period $O(M/k)$).

Main Structural Theorem (HD)

Given a pattern P of length M, a text T of length $N \leq {}^3/_2\, M$, and a threshold $k \leq M$, at least one of the following holds:

- The number of k-mismatch occurrences of P in T is at most $O(k)$.
- The pattern P is almost periodic (at HD $< 2k$ to a string with period $O(M/k)$).

Key Lemma (Analyze)

For each string P of length M, at least one of the following holds:

- P contains $2k$ disjoint breaks; each break has length $M/8k$ and period $> M/128k$.
- P contains disjoint repetitive regions R_i with total length $\geq 3/8 \cdot M$; each region has length $\geq M/8k$ and is almost periodic with HD exactly $8k/M \cdot |R_i|$.
- P is almost periodic (at HD $< 8k$ to a string with period $M/128k$).

Main Structural Theorem (HD)

Given a pattern P of length M, a text T of length $N \leq {}^{3}\!/_{2}\, M$, and a threshold $k \leq M$, at least one of the following holds:

- The number of k-mismatch occurrences of P in T is at most $O(k)$.
- The pattern P is almost periodic (at HD < $2k$ to a string with period $O(M/k)$).

Key Lemma (Analyze)

For each string P of length M, at least one of the following holds:

- P contains $2k$ disjoint breaks; each break has length $M/8k$ and period $> M/128k$.
- P contains disjoint repetitive regions R_i with total length $\geq 3/8 \cdot M$;
  each region has length $\geq M/8k$ and is almost periodic with HD exactly $8k/M \cdot |R_i|$.
- P is almost periodic (at HD < $8k$ to a string with period $M/128k$).

Main Structural Theorem (HD)

Given a pattern P of length M, a text T of length $N \le {}^3/_2 M$, and a threshold $k \le M$, at least one of the following holds:

- The number of k-mismatch occurrences of P in T is at most $O(k)$.
- The pattern P is almost periodic (at HD < $2k$ to a string with period $O(M/k)$).

Key Lemma (Analyze)

For each string P of length M, at least one of the following holds:

- P contains $2k$ disjoint breaks; each break has length $M/8k$ and period > $M/128k$.
- P contains disjoint repetitive regions R_i with total length $\ge 3/8 \cdot M$;
  each region has length $\ge M/8k$ and is almost periodic with HD exactly $8k/M \cdot |R_i|$.
- P is almost periodic (at HD < $8k$ to a string with period $M/128k$).

Observation [BKW'19, refined]

If P has HD $\ge 2k$ and < $8k$ to a string w/ period $O(M/k)$, there are $O(k)$ k-mism. occ's of P in T.

Main Structural Theorem (HD)

Given a pattern P of length M, a text T of length $N \le {}^{3}/_{2}\, M$, and a threshold $k \le M$, at least one of the following holds:

- The number of k-mismatch occurrences of P in T is at most $O(k)$.
- The pattern P is almost periodic (at HD $< 2k$ to a string with period $O(M/k)$).

Key Lemma (Analyze)

For each string P of length M, at least one of the following holds:

- P contains $2k$ disjoint breaks; each break has length $M/8k$ and period $> M/128k$.
- P contains disjoint repetitive regions R_i with total length $\ge 3/8 \cdot M$;
  each region has length $\ge M/8k$ and is almost periodic with HD exactly $8k/M \cdot |R_i|$.
- P is almost periodic (at HD $< 8k$ to a string with period $M/128k$).

Observation [BKW'19, refined]

If P has HD $\ge 2k$ and $< 8k$ to a string w/ period $O(M/k)$, there are $O(k)$ k-mism. occ's of P in T.

Main Structural Theorem (HD)

Given a pattern P of length M, a text T of length $N \le {}^3/_2\, M$, and a threshold $k \le M$, at least one of the following holds:

- The number of k-mismatch occurrences of P in T is at most $O(k)$.
- The pattern P is almost periodic (at HD $< 2k$ to a string with period $O(M/k)$).

Key Lemma (Analyze)

For each string P of length M, at least one of the following holds:

- P contains $2k$ disjoint breaks; each break has length $M/8k$ and period $> M/128k$.
- P contains disjoint repetitive regions R_i with total length $\ge 3/8 \cdot M$; each region has length $\ge M/8k$ and is almost periodic with HD exactly $8k/M \cdot |R_i|$.
- P is almost periodic (at HD $< 8k$ to a string with period $M/128k$).

Main Structural Theorem (HD)

Given a pattern P of length M, a text T of length $N \le {}^3/_2 M$, and a threshold $k \le M$, at least one of the following holds:

- The number of k-mismatch occurrences of P in T is at most $O(k)$.
- The pattern P is almost periodic (at HD < $2k$ to a string with period $O(M/k)$).

Key Lemma (Analyze)

For each string P of length M, at least one of the following holds:

- P contains $2k$ disjoint breaks; each break has length $M/8k$ and period $> M/128k$.
- P contains disjoint repetitive regions R_i with total length $\ge 3/8 \cdot M$; each region has length $\ge M/8k$ and is almost periodic with HD exactly $8k/M \cdot |R_i|$.
- P is almost periodic (at HD < $8k$ to a string with period $M/128k$).

Observation [BKW'19, refined]

If P has HD $\ge 2k$ and $< 8k$ to a string w/ period $O(M/k)$, there are $O(k)$ k-mism. occ's of P in T.

Main Structural Theorem (HD)

Given a pattern P of length M, a text T of length $N \le {}^3/_2\, M$, and a threshold $k \le M$, at least one of the following holds:

- The number of k-mismatch occurrences of P in T is at most $O(k)$.
- The pattern P is almost periodic (at HD < $2k$ to a string with period $O(M/k)$).

Key Lemma (Analyze)

For each string P of length M, at least one of the following holds:

- P contains $2k$ disjoint breaks; each break has length $M/8k$ and period $> M/128k$.
- P contains disjoint repetitive regions R_i with total length $\ge 3/8 \cdot M$; each region has length $\ge M/8k$ and is almost periodic with HD exactly $8k/M \cdot |R_i|$.
- P is almost periodic (at HD < $8k$ to a string with period $M/128k$).

Main Structural Theorem (HD)

Given a pattern P of length M, a text T of length $N \le {}^3\!/_2 M$, and a threshold $k \le M$, at least one of the following holds:

- The number of k-mismatch occurrences of P in T is at most $O(k)$.
- The pattern P is almost periodic (at HD < $2k$ to a string with period $O(M/k)$).

Key Lemma (Analyze)

For each string P of length M, at least one of the following holds:

- P contains $2k$ disjoint breaks; each break has length $M/8k$ and period $> M/128k$.
- P contains disjoint repetitive regions R_i with total length $\ge 3/8 \cdot M$;
  each region has length $\ge M/8k$ and is almost periodic with HD exactly $8k/M \cdot |R_i|$.
- P is almost periodic (at HD < $8k$ to a string with period $M/128k$).

Observation [BKW'19, refined]

If P contains $\ge 2k$ disjoint breaks, there are $O(k)$ k-mismatch occ's of P in T.

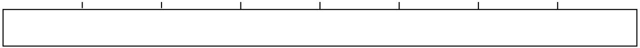
Analyzing the Pattern, Proof Idea

Key Lemma (Analyze)

For each string P of length M, at least one of the following holds:

- P contains $2k$ disjoint breaks; each break has length $M/8k$ and period $> M/128k$.
- P contains disjoint repetitive regions R_i with total length $\geq 3/8 \cdot M$;
  each region has length $\geq M/8k$ and is almost periodic with HD exactly $8k/M \cdot |R_i|$.
- P is almost periodic (at HD $< 8k$ to a string with period $M/128k$).

P

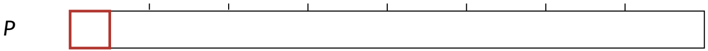


Analyzing the Pattern, Proof Idea

Key Lemma (Analyze)

For each string P of length M, at least one of the following holds:

- P contains $2k$ disjoint breaks; each break has length $M/8k$ and period $> M/128k$.
- P contains disjoint repetitive regions R_i with total length $\geq 3/8 \cdot M$;
  each region has length $\geq M/8k$ and is almost periodic with HD exactly $8k/M \cdot |R_i|$.
- P is almost periodic (at HD $< 8k$ to a string with period $M/128k$).



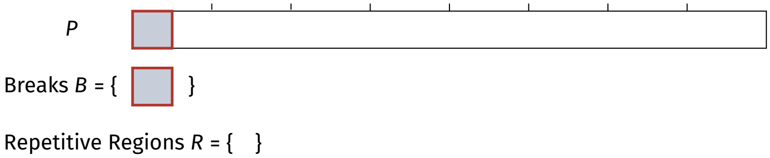
- Process P from left to right, $M/8k$ new characters at a time.

Analyzing the Pattern, Proof Idea

Key Lemma (Analyze)

For each string P of length M, at least one of the following holds:

- P contains $2k$ disjoint breaks; each break has length $M/8k$ and period $> M/128k$.
- P contains disjoint repetitive regions R_i with total length $\geq 3/8 \cdot M$; each region has length $\geq M/8k$ and is almost periodic with HD exactly $8k/M \cdot |R_i|$.
- P is almost periodic (at HD $< 8k$ to a string with period $M/128k$).



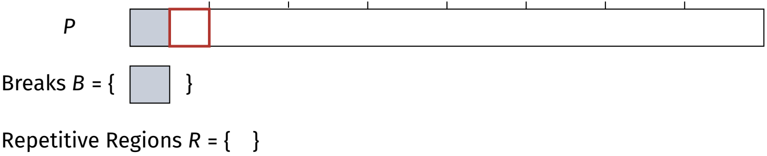
- If a fragment has a period $> M/128k$, add it to the found breaks.

Analyzing the Pattern, Proof Idea

Key Lemma (Analyze)

For each string P of length M, at least one of the following holds:

- P contains $2k$ disjoint breaks; each break has length $M/8k$ and period $> M/128k$.
- P contains disjoint repetitive regions R_i with total length $\geq 3/8 \cdot M$;
  each region has length $\geq M/8k$ and is almost periodic with HD exactly $8k/M \cdot |R_i|$.
- P is almost periodic (at HD $< 8k$ to a string with period $M/128k$).



- Otherwise, find the shortest prefix (longer than $M/8k$) that is a repetitive region.

Analyzing the Pattern, Proof Idea

Key Lemma (Analyze)

For each string P of length M, at least one of the following holds:

- P contains $2k$ disjoint breaks; each break has length $M/8k$ and period $> M/128k$.
- P contains disjoint repetitive regions R_i with total length $\geq 3/8 \cdot M$;
  each region has length $\geq M/8k$ and is almost periodic with HD exactly $8k/M \cdot |R_i|$.
- P is almost periodic (at HD $< 8k$ to a string with period $M/128k$).

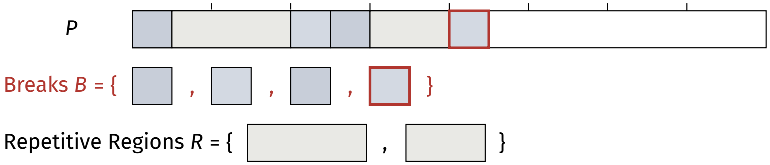
- Otherwise, find the shortest prefix (longer than $M/8k$) that is a repetitive region.

Analyzing the Pattern, Proof Idea

Key Lemma (Analyze)

For each string P of length M, at least one of the following holds:

- P contains $2k$ disjoint breaks; each break has length $M/8k$ and period $> M/128k$.
- P contains disjoint repetitive regions R_i with total length $\geq 3/8 \cdot M$; each region has length $\geq M/8k$ and is almost periodic with HD exactly $8k/M \cdot |R_i|$.
- P is almost periodic (at HD $< 8k$ to a string with period $M/128k$).



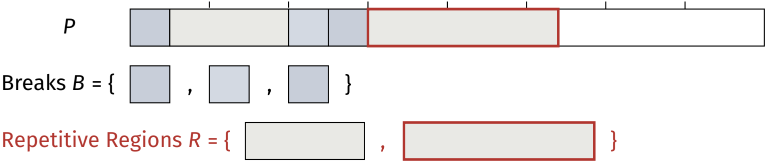
- If we found $2k$ breaks, return the breaks.

Analyzing the Pattern, Proof Idea

Key Lemma (Analyze)

For each string P of length M, at least one of the following holds:

- P contains $2k$ disjoint breaks; each break has length $M/8k$ and period $> M/128k$.
- P contains disjoint repetitive regions R_i with total length $\geq 3/8 \cdot M$; each region has length $\geq M/8k$ and is almost periodic with HD exactly $8k/M \cdot |R_i|$.
- P is almost periodic (at HD $< 8k$ to a string with period $M/128k$).



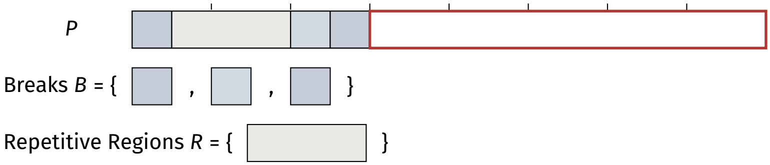
- If the total length of the repetitive regions is $> 3/8 \cdot M$, return the repetitive regions.

Analyzing the Pattern, Proof Idea

Key Lemma (Analyze)

For each string P of length M, at least one of the following holds:

- P contains $2k$ disjoint breaks; each break has length $M/8k$ and period $> M/128k$.
- P contains disjoint repetitive regions R_i with total length $\geq 3/8 \cdot M$;
  each region has length $\geq M/8k$ and is almost periodic with HD exactly $8k/M \cdot |R_i|$.
- P is almost periodic (at HD $< 8k$ to a string with period $M/128k$).



- If we reach the end of P, try to find a single repetitive region starting from the end.

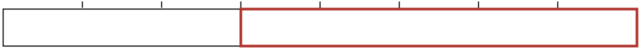
Analyzing the Pattern, Proof Idea

Key Lemma (Analyze)

For each string P of length M, at least one of the following holds:

- P contains $2k$ disjoint breaks; each break has length $M/8k$ and period $> M/128k$.
- P contains disjoint repetitive regions R_i with total length $\geq 3/8 \cdot M$;
  each region has length $\geq M/8k$ and is almost periodic with HD exactly $8k/M \cdot |R_i|$.
- P is almost periodic (at HD $< 8k$ to a string with period $M/128k$).

P



Breaks $B = \{ \quad \}$

Repetitive Regions $R = \{ \quad \}$

- If we reach the end of P, try to find a single repetitive region starting from the end.

Analyzing the Pattern, Proof Idea

Key Lemma (Analyze)

For each string P of length M, at least one of the following holds:

- P contains $2k$ disjoint breaks; each break has length $M/8k$ and period $> M/128k$.
- P contains disjoint repetitive regions R_i with total length $\geq 3/8 \cdot M$; each region has length $\geq M/8k$ and is almost periodic with HD exactly $8k/M \cdot |R_i|$.
- P is almost periodic (at HD $< 8k$ to a string with period $M/128k$).

Breaks $B = \{ \quad \}$

Repetitive Regions $R = \{$  $\}$

- If we found a repetitive region, return it.

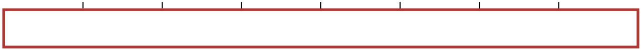
Analyzing the Pattern, Proof Idea

Key Lemma (Analyze)

For each string P of length M, at least one of the following holds:

- P contains $2k$ disjoint breaks; each break has length $M/8k$ and period $> M/128k$.
- P contains disjoint repetitive regions R_i with total length $\geq 3/8 \cdot M$;
  each region has length $\geq M/8k$ and is almost periodic with HD exactly $8k/M \cdot |R_i|$.
- P is almost periodic (at HD $< 8k$ to a string with period $M/128k$).

P



Breaks $B = \{\quad\}$

Repetitive Regions $R = \{\quad\}$

- If we again don't obtain a repetitive region, P is almost periodic.

Analyzing the Pattern, Proof Idea

Key Lemma (Analyze) ✓

For each string P of length M, at least one of the following holds:

- P contains $2k$ disjoint breaks; each break has length $M/8k$ and period $> M/128k$.
- P contains disjoint repetitive regions R_i with total length $\geq 3/8 \cdot M$;
  each region has length $\geq M/8k$ and is almost periodic with HD exactly $8k/M \cdot |R_i|$.
- P is almost periodic (at HD $< 8k$ to a string with period $M/128k$).

Obtaining Fast Algorithms: Fast PILLAR Algorithms

The PILLAR operations: LCP, LCP^R, IPM, Length, Access, Extract

Theorem (PILLAR Alg. for PM w/ Mism.)

Given a pattern P of length m, a text T of length n, and a positive threshold $k \le m$, we can compute (a representation of) all k-mismatch occurrences of P in T using $O(n/m \cdot k^2 \log\log k)$ time plus $O(n/m \cdot k^2)$ PILLAR operations.

Obtaining Fast Algorithms: Fast PILLAR Algorithms

The PILLAR operations: LCP, LCP^R, IPM, Length, Access, Extract

Theorem (PILLAR Alg. for PM w/ Mism.)

Given a pattern P of length m, a text T of length n, and a positive threshold $k \le m$, we can compute (a representation of) all k-mismatch occurrences of P in T using $O(n/m \cdot k^2 \log\log k)$ time plus $O(n/m \cdot k^2)$ PILLAR operations.

Theorem (PILLAR Alg. for PM w/ Errors)

Given a pattern P of length m, a text T of length n, and a positive threshold $k \le m$, we can compute (a representation of) all k-error occurrences of P in T using $O(n/m \cdot k^4)$ PILLAR operations.

Obtaining Fast Algorithms: The Standard Setting

The PILLAR operations: LCP, LCP^R, IPM, Length, Access, Extract

Uncompressed strings: pattern P of length M, text T of length N.

Obtaining Fast Algorithms: The Standard Setting

The `PILLAR` operations: `LCP`, $\texttt{LCP}^R$, `IPM`, `Length`, `Access`, `Extract`

Uncompressed strings: pattern P of length M, text T of length N.

We can implement each operation in $O(1)$ time. (After $O(N + M)$ preprocessing.)

Obtaining Fast Algorithms: The Standard Setting

The PILLAR operations: LCP, LCP^R, IPM, Length, Access, Extract

Uncompressed strings: pattern P of length M, text T of length N.

We can implement each operation in $O(1)$ time. (After $O(N+M)$ preprocessing.)

Theorem (Algorithm for PM w/ Mism.)

For any positive threshold $k \le M$,
we can compute all k-mismatch occ's of P in T in time $O(N + N/M \cdot k^2 \log\log k)$.

Theorem (Algorithm for PM w/ Errors)

For any positive threshold $k \le M$,
we can compute (starting positions of) all k-error occ's of P in T in time $O(N + N/M \cdot k^4)$.

Obtaining Fast Algorithms: The Fully-Compressed Setting

The PILLAR operations: LCP, LCP^R, IPM, Length, Access, Extract

SLPs: G_P of size m generating pattern P, G_T of size n generating text T.

Obtaining Fast Algorithms: The Fully-Compressed Setting

The PILLAR operations: LCP, LCP^R, IPM, Length, Access, Extract

SLPs: G_P of size m generating pattern P, G_T of size n generating text T.

Using Recompression [Jeż'15], we can implement each operation in $O(\log^3(|P| + |T|))$ time.

(After $O((n + m) \log(|P| + |T|))$ preprocessing.)

Obtaining Fast Algorithms: The Fully-Compressed Setting

The `PILLAR` operations: `LCP`, $\texttt{LCP}^R$, `IPM`, `Length`, `Access`, `Extract`

SLPs: G_P of size m generating pattern P, G_T of size n generating text T.

Using Recompression [Jeż'15], we can implement each operation in $O(\log^3(|P| + |T|))$ time.

(After $O((n+m)\log(|P| + |T|))$ preprocessing.)

Theorem (Algorithm for PM w/ Mism.)

For any positive threshold $k \leq |P|$, we can compute the number of all k-mismatch occ's of P in T in time $O(m\log(|P| + |T|) + nk^2\log^3(|P| + |T|))$.

(Reporting of all occ's takes time linear in the number of occ's.)

Theorem (Algorithm for PM w/ Errors)

For any positive threshold $k \leq |P|$, we can compute the number of all k-error occ's of P in T in time $O(m\log(|P| + |T|) + nk^4\log^3(|P| + |T|))$.

(Reporting of all occ's takes time linear in the number of occ's.)

Obtaining Fast Algorithms: The Dynamic Setting

The `PILLAR` operations: `LCP`, $\mathtt{LCP}^R$, `IPM`, `Length`, `Access`, `Extract`

Dynamic maintenance of a collection of (non-empty persistent) strings $\mathcal{X}$ of total length N; supporting `makestring`, `concat`, `split`.

Obtaining Fast Algorithms: The Dynamic Setting

The PILLAR operations: LCP, LCP^R, IPM, Length, Access, Extract

Dynamic maintenance of a collection of (non-empty persistent) strings X of total length N; supporting makestring, concat, split.

Using Optimal Dynamic Strings [Gawrychowski et al.'18], we can implement each PILLAR operation in (w.h.p) $O(\log^2 N)$ time.

Obtaining Fast Algorithms: The Dynamic Setting

The PILLAR operations: LCP, LCP^R, IPM, Length, Access, Extract

Dynamic maintenance of a collection of (non-empty persistent) strings X of total length N; supporting makestring, concat, split.

Using Optimal Dynamic Strings [Gawrychowski et al.'18], we can implement each PILLAR operation in (w.h.p) $O(\log^2 N)$ time.

Theorem (Algorithm for PM w/ Mism.)

For any two strings $P, T \in X$ and any threshold k, we support the additional operation "Find all k-mismatch occ's of P in T" in (w.h.p) $O(|T|/|P| \cdot k^2 \log^2 N)$ time.

Theorem (Algorithm for PM w/ Errors)

For any two strings $P, T \in X$ and any threshold k, we support the additional operation "Find all k-error occ's of P in T" in (w.h.p) $O(|T|/|P| \cdot k^4 \log^2 N)$ time.

Navigation

Start

End