Few Matches or Almost Periodicity: Faster Pattern Matching with Mismatches in Compressed Texts

Karl Bringmann, Marvin Künnemann, and Philip Wellnitz

Max Planck Institute for Informatics, Saarland Informatics Campus (SIC), Saarbrücken, Germany

April 25, 2020
Pattern Matching with Mismatches

Pattern Matching
Given a text $t$ and a pattern $p$, is $p$ a substring of $t$?

Finding CAKE
Pattern Matching with Mismatches

Given a text $t$, a pattern $p$, and an integer $k$, does $t$ have a length-$|p|$ substring with Hamming-distance at most $k$ to $p$?

Finding ANPAN, $k = 2$
Pattern Matching with Mismatches

**Thm. [Gawrychowski, Uznanski’18]**

Pattern matching with $k$ mismatches on a text of length $n$ and a pattern of length $m$ can be solved in time $\tilde{O}((m + k\sqrt{m}) \cdot n/m)$. 
Pattern Matching with Mismatches

Given a text $t$, a pattern $p$, and an integer $k$, does $t$ have a length-$|p|$ substring with Hamming-distance at most $k$ to $p$?

Thm. [Gawrychowski,Uznanski’18]
Pattern matching with $k$ mismatches on a text of length $n$ and a pattern of length $m$ can be solved in time $\tilde{O}((m + k\sqrt{m}) \cdot n/m)$.

Matching (conditional) lower bound [GU’18]
What if the text is much larger than the pattern?

ANPAN

ANPAN

sesweet roll most commonly filled with red bean paste. Anpan can also be prepared with other fillings.

Karl Bringmann, Marvin Künnemann, and Philip Wellnitz

Faster Pattern Matching with Mismatches in Compressed Texts
What if the text is much larger than the pattern?

ANPAN is a Japanese sweet roll most commonly filled with red bean paste. ANPAN can also be prepared with other fillings including white beans, green beans, sesame, and chestnut.

ANPAN
What if the text is much larger than the pattern and given in a compressed representation?

ANPAN is a Japanese sweet roll. Most commonly filled with red bean paste, anpan can also be prepared with other fillings including white beans, green beans, sesame, and chestnut.
Grammar Compression

**Straight-Line Program (SLP)**

A Straight-Line Program or SLP $\mathcal{T}$ is a context-free grammar that generates exactly one string $\text{eval}(\mathcal{T})$. 
Grammar Compression

**Straight-Line Program (SLP)**

An SLP $\mathcal{T}$ is a set of non-terminals $\{T_1, \ldots, T_n\}$ and productions of the form $T_i \rightarrow \sigma$ or $T_i \rightarrow T_\ell T_r$, where $\ell, r < i$.

We write $\text{eval}(\mathcal{T}) = \text{eval}(T_n)$ for the generated string.
Grammar Compression

**Straight-Line Program (SLP)**

An SLP $\mathcal{T}$ is a set of non-terminals $\{T_1, \ldots, T_n\}$ and productions of the form $T_i \rightarrow \sigma$ or $T_i \rightarrow T_\ell T_r$, where $\ell, r < i$. We write $\text{eval}(\mathcal{T}) = \text{eval}(T_n)$ for the generated string.

$$T_1 \rightarrow A; \ T_2 \rightarrow N; \ T_3 \rightarrow P$$

Karl Bringmann, Marvin Künnemann, and Philip Wellnitz

Faster Pattern Matching with Mismatches in Compressed Texts
Grammar Compression

**Straight-Line Program (SLP)**

An SLP $\mathcal{T}$ is a set of non-terminals $\{T_1, \ldots, T_n\}$ and productions of the form $T_i \rightarrow \sigma$ or $T_i \rightarrow T_\ell T_r$, where $\ell, r < i$.

We write $\text{eval}(\mathcal{T}) = \text{eval}(T_n)$ for the generated string.

$$
\begin{align*}
T_1 & \rightarrow A; \quad T_2 \rightarrow N; \quad T_3 \rightarrow P \\
T_4 & \rightarrow T_1 T_2; \quad T_5 \rightarrow T_4 T_3
\end{align*}
$$
Grammar Compression

**Straight-Line Program (SLP)**

An SLP $\mathcal{T}$ is a set of non-terminals $\{T_1, \ldots, T_n\}$ and productions of the form $T_i \to \sigma$ or $T_i \to T_\ell T_r$, where $\ell, r < i$.

We write $\text{eval}(\mathcal{T}) = \text{eval}(T_n)$ for the generated string.

\[
\begin{align*}
T_1 & \to A; \\
T_2 & \to N; \\
T_3 & \to P \\
T_4 & \to T_1 T_2; \\
T_5 & \to T_4 T_3 \\
T_6 & \to T_5 T_4
\end{align*}
\]
Grammar Compression

Straight-Line Program (SLP)

An SLP $\mathcal{T}$ is a set of non-terminals $\{T_1, \ldots, T_n\}$ and productions of the form $T_i \rightarrow \sigma$ or $T_i \rightarrow T_\ell T_r$, where $\ell, r < i$.

We write $\text{eval}(\mathcal{T}) = \text{eval}(T_n)$ for the generated string.

$$
\begin{align*}
T_1 & \rightarrow A; \quad T_2 \rightarrow N; \quad T_3 \rightarrow P \\
T_4 & \rightarrow T_1 T_2; \quad T_5 \rightarrow T_4 T_3 \\
T_6 & \rightarrow T_5 T_4; \quad T_7 \rightarrow T_6 T_4
\end{align*}
$$
### Known Results

<table>
<thead>
<tr>
<th>Problem</th>
<th>uncompressed</th>
<th>LZW/LZ78 text</th>
<th>SLP text</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$n = \Omega(\sqrt{N})$</td>
<td>$n = \Omega(\log N)$</td>
<td></td>
</tr>
<tr>
<td><strong>Pattern</strong></td>
<td>$O(N + m)$ [KMP’77]</td>
<td>$O(n + m)$ ※ [G’12]</td>
<td>$\tilde{O}(n + m)$ ※ [J’15]</td>
</tr>
<tr>
<td><strong>Matching</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>PM with $k$</strong></td>
<td>$\tilde{O}\left(\frac{N}{m}(m + k\sqrt{m})\right)$ [GU’18]</td>
<td>$O(n\sqrt{mk^2})$ [GS’13]</td>
<td>$\tilde{O}(nm\text{poly}(k))$ [T’14,BLRS’15]</td>
</tr>
<tr>
<td><strong>Mismatches</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

$N$: length of uncompressed text  
$m$: length of pattern  
$n$: length of compressed text  
$k$: number of mismatches  
※: allows compressed pattern
# Known Results

<table>
<thead>
<tr>
<th>Problem</th>
<th>uncompressed</th>
<th>LZW/LZ78 text</th>
<th>SLP text</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$n = \Omega(\sqrt{N})$</td>
<td>$n = \Omega(\log N)$</td>
<td></td>
</tr>
<tr>
<td>Pattern Matching</td>
<td>$O(N + m)$</td>
<td>$O(n + m)$ ※</td>
<td>$\tilde{O}(n + m)$ ※</td>
</tr>
<tr>
<td></td>
<td>[KMP’77]</td>
<td>[G’12]</td>
<td>[J’15]</td>
</tr>
<tr>
<td>PM with $k$</td>
<td>$\tilde{O}(\frac{N}{m}(m + k\sqrt{m}))$</td>
<td>$O(n\sqrt{mk^2})$</td>
<td>$\tilde{O}(nm\text{poly}(k))$</td>
</tr>
<tr>
<td>Mismatches</td>
<td>[GU’18]</td>
<td>[GS’13]</td>
<td>[T’14,BLRS’15]</td>
</tr>
</tbody>
</table>

$N$: length of uncompressed text  
$m$: length of pattern  
$n$: length of compressed text  
$k$: number of mismatches  
※: allows compressed pattern
### Known Results

<table>
<thead>
<tr>
<th>Problem</th>
<th>uncompressed</th>
<th>LZW/LZ78 text $n = \Omega(\sqrt{N})$</th>
<th>SLP text $n = \Omega(\log N)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pattern Matching</td>
<td>$O(N + m)$</td>
<td>$O(n + m)$ ※</td>
<td>$\tilde{O}(n + m)$ ※</td>
</tr>
<tr>
<td></td>
<td>[KMP’77]</td>
<td></td>
<td>[G’12]</td>
</tr>
<tr>
<td>PM with $k$</td>
<td>$\tilde{O}(\frac{N}{m}(m + k\sqrt{m}))$</td>
<td>$O(n\sqrt{mk^2})$</td>
<td>$\tilde{O}(nm\text{poly}(k))$</td>
</tr>
<tr>
<td>Mismatches</td>
<td>[GU’18]</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

$N$: length of uncompressed text  
$n$: length of compressed text  
$m$: length of pattern  
$k$: number of mismatches  
※: allows compressed pattern

Karl Bringmann, Marvin Künnemann, and Philip Wellnitz
Faster Pattern Matching with Mismatches in Compressed Texts
## Known Results

<table>
<thead>
<tr>
<th>Problem</th>
<th>uncompressed</th>
<th>LZW/LZ78 text $n = \Omega(\sqrt{N})$</th>
<th>SLP text $n = \Omega(\log N)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pattern Matching</td>
<td>$O(N + m)$</td>
<td>$O(n + m)$ ※</td>
<td>$\tilde{O}(n + m)$ ※</td>
</tr>
<tr>
<td></td>
<td>[KMP’77]</td>
<td>[G’12]</td>
<td>[J’15]</td>
</tr>
<tr>
<td>PM with $k$ Mismatches</td>
<td>$\tilde{O}\left(\frac{N}{m}(m + k\sqrt{m})\right)$</td>
<td>$O(n\sqrt{mk^2})$</td>
<td>$\tilde{O}(nm \text{poly}(k))$</td>
</tr>
<tr>
<td></td>
<td>[GU’18]</td>
<td></td>
<td>$\tilde{O}(nk^4 + mk)$</td>
</tr>
</tbody>
</table>

$N$: length of uncompressed text  
$n$: length of compressed text  
$m$: length of pattern  
$k$: number of mismatches  
※: allows compressed pattern

Improvement obtained via new structural insight in solution structure
Solution Structure of Pattern Matching

**Fact (Folklore)**

Let text $t$ and pattern $p$, $|t| \leq \frac{3}{2} |p|$, be given such that there are $\geq 2$ 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 $1 + i \cdot |x|$.  

Karl Bringmann, Marvin Künnemann, and Philip Wellnitz  
Faster Pattern Matching with Mismatches in Compressed Texts
Solution Structure of Pattern Matching

Fact (Folklore)

Let text $t$ and pattern $p$, $|t| \leq \frac{3}{2} |p|$, be given such that there are $\geq 2$ 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 $1 + i \cdot |x|$.

\[
\begin{array}{c}
t \quad p \\
p \\
p 
\end{array}
\]
Solution Structure of Pattern Matching

Fact (Folklore)

Let text \( t \) and pattern \( p \), \( |t| \leq \frac{3}{2}|p| \), be given such that there are \( \geq 2 \) 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 \( 1 + i \cdot |x| \).
Solution Structure of Pattern Matching

**Fact (Folklore)**

Let text $t$ and pattern $p$, $|t| \leq \frac{3}{2}|p|$, be given such that there are $\geq 2$ 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 $1 + i \cdot |x|$. 
Solution Structure of Pattern Matching

Fact (Folklore)

Let text $t$ and pattern $p$, $|t| \leq \frac{3}{2}|p|$, be given such that there are $\geq 2$ 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 $1 + i \cdot |x|$. 

$$
\begin{array}{c}
 t \\
 p \end{array}
$$

$X X$

$X X$

$X$

Karl Bringmann, Marvin Künnemann, and Philip Wellnitz

Faster Pattern Matching with Mismatches in Compressed Texts
Solution Structure of Pattern Matching

Fact (Folklore)

Let text $t$ and pattern $p$, $|t| \leq \frac{3}{2}|p|$, be given such that there are $\geq 2$ 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 $1 + i \cdot |x|$. 
What is the solution structure of Pattern Matching with Mismatches?
Solution Structure of Pattern Matching with Mismatches

If there are at least 2 $k$-matches of $p$ in $t$, then $p$ and $t$ are periodic and every $k$-match of $p$ starts at a position $1 + i|x|$?
Solution Structure of Pattern Matching with Mismatches

If there are at least two $k$-matches of $p$ in $t$, then $p$ and $t$ are periodic and every $k$-match of $p$ starts at a position $1 + |x|$. 

![Diagram showing periodicity and matching structure]
Solution Structure of Pattern Matching with Mismatches

If there are at least two $k$-matches of $p$ in $t$, then $p$ and $t$ are periodic and every $k$-match of $p$ starts at a position $1 + i|x|$?

- $p$ and $t$ not periodic, but $2k$ $k$-matches of $p$ in $t$
Solution Structure of Pattern Matching with Mismatches

If there are at least two \( \Omega(\text{poly}(k)) \) \( k \)-matches of \( p \) in \( t \), then \( p \) and \( t \) are periodic and every \( k \)-match of \( p \) starts at a position \( 1 + i|x|? \)

```
Insight 1
Periodicity only if number of \( k \)-matches of \( p \) in \( t \) is \( \Omega(\text{poly}(k)) \)
```
Solution Structure of Pattern Matching with Mismatches

If there are at least $\Omega(poly(k))$ $k$-matches of $p$ in $t$, then $p$ and $t$ are periodic and every $k$-match of $p$ starts at a position $1 + i|x|$?
Solution Structure of Pattern Matching with Mismatches

If there are at least $\Omega(poly(k))$ $k$-matches of $p$ in $t$, then $p$ and $t$ are periodic and every $k$-match of $p$ starts at a position $1 + i|x|$?

---

Karl Bringmann, Marvin Künnemann, and Philip Wellnitz

Faster Pattern Matching with Mismatches in Compressed Texts
Solution Structure of Pattern Matching with Mismatches

If there are at least $\Omega(poly(k))$ $k$-matches of $p$ in $t$, then $p$ and $t$ are periodic and every $k$-match of $p$ starts at a position $1 + i|x|$?

- $O(m)$ $k$-matches of $p$ in $t$, but $p$ and $t$ not perfectly periodic

\[ t \begin{array}{c}
\text{A} \text{A} \cdots \text{A} \text{A} \text{B} \text{B} \text{A} \text{A} \cdots \text{A} \text{B}
\end{array}\]

\[ p \begin{array}{c}
\text{B} \text{A} \cdots \text{A} \text{B} \text{A}
\end{array}\]
Solution Structure of Pattern Matching with Mismatches

If there are at least $\Omega(\text{poly}(k))$ $k$-matches of $p$ in $t$, then $p$ and $t$ are periodic up to $O(k)$ mismatches and every $k$-match of $p$ starts at a position $1 + i|x|$.

**Insight 2**

Periodicity only up to $O(k)$ mismatches
Solution Structure of Pattern Matching with Mismatches

If there are at least $\Omega(poly(k))$ $k$-matches of $p$ in $t$, then $p$ and $t$ are periodic up to $O(k)$ mismatches and every $k$-match of $p$ starts at a position $1 + i|\chi|$?
Main Result

**Theorem (Structural Insight)**

For pattern $p$ and text $t$, $|t| \leq 2|p|$, at least one of the following holds:

- The number of $k$-matches of $p$ in $t$ is at most $O(k^2)$, or

- $t'$: shortest substring of $t$ such that any $k$-match of $p$ in $t$ is also a $k$-match in $t'$
  Both $t'$ and $p$ have HD $O(k)$ to the same periodic string $x$ and all
  $k$-matches of $p$ in $t'$ start at a position $1 + i \cdot |x|$. 
Main Result

**Theorem (Structural Insight)**

For pattern $p$ and text $t$, $|t| \leq 2|p|$, at least one of the following holds:

- The number of $k$-matches of $p$ in $t$ is at most $O(k^2)$, or
- $t'$: shortest substring of $t$ such that any $k$-match of $p$ in $t$ is also a $k$-match in $t'$
  
  Both $t'$ and $p$ have HD $O(k)$ to the same periodic string $x$ and all $k$-matches of $p$ in $t'$ start at a position $1 + i \cdot |x|$. 
Main Result, Proof Overview

Theorem (Structural Insight)

For pattern \( p \) and text \( t \), \(|t| \leq 2|p|\), at least one of the following holds:

- The number of \( k \)-matches of \( p \) in \( t \) is at most \( 1000k^2 \), or
- Both \( t' \) and \( p \) have HD < \( 20k \) to a periodic \( x \); all \( k \)-matches start at position \( 1 + i \cdot |x| \).

\( t \)

\( p \)
Main Result, Proof Overview

**Theorem (Structural Insight)**

For pattern $p$ and text $t$, $|t| \leq 2|p|$, at least one of the following holds:

- The number of $k$-matches of $p$ in $t$ is at most $1000k^2$, or
- Both $t'$ and $p$ have HD $< 20k$ to a periodic $x$; all $k$-matches start at position $1 + i \cdot |x|$.

Consider $t'$: shortest substring of $t$ that contains all $k$-matches.
Main Result, Proof Overview

**Theorem (Structural Insight)**

For pattern $p$ and text $t$, $|t| \leq 2|p|$, at least one of the following holds:

- The number of $k$-matches of $p$ in $t$ is at most $1000k^2$, or
- Both $t'$ and $p$ have HD $< 20k$ to a periodic $x$; all $k$-matches start at position $1 + i \cdot |x|$.

- Split $p$ into $16k$ parts $p_i$ of equal length
Main Result, Proof Overview

Theorem (Structural Insight)
For pattern $p$ and text $t$, $|t| \leq 2|p|$, at least one of the following holds:
- The number of $k$-matches of $p$ in $t$ is at most $1000k^2$, or
- Both $t'$ and $p$ have HD $< 20k$ to a periodic $x$; all $k$-matches start at position $1 + i \cdot |x|$.

- Fix a $p_i$
Main Result, Proof Overview

**Theorem (Structural Insight)**

For pattern $p$ and text $t$, $|t| \leq 2|p|$, at least one of the following holds:

- The number of $k$-matches of $p$ in $t$ is at most $1000k^2$, or
- Both $t'$ and $p$ have HD $< 20k$ to a periodic $x$; all $k$-matches start at position $1 + i \cdot |x|$.

Consider prefix $x_i$ of $p_i$ that is also a period of $p_i$. 
Main Result, Proof Overview

Theorem (Structural Insight)

For pattern $p$ and text $t$, $|t| \leq 2|p|$, at least one of the following holds:

- The number of $k$-matches of $p$ in $t$ is at most $1000k^2$, or
- Both $t'$ and $p$ have HD $< 20k$ to a periodic $x$; all $k$-matches start at position $1 + i \cdot |x|$.

- Find first $3k$ mismatches between $p$ and $x_i^*$ before and after $p_i$
Main Result, Proof Overview

**Theorem (Structural Insight)**

For pattern $p$ and text $t$, $|t| \leq 2|p|$, at least one of the following holds:

- The number of $k$-matches of $p$ in $t$ is at most $1000k^2$, or
- Both $t'$ and $p$ have HD $< 20k$ to a periodic $x$; all $k$-matches start at position $1 + i \cdot |x|$. 

- Find first $3k$ mismatches between $p$ and $x_i^*$ before and after $p_i$
Main Result, Proof Overview

**Theorem (Structural Insight)**

For pattern $p$ and text $t$, $|t| \leq 2|p|$, at least one of the following holds:

- The number of $k$-matches of $p$ in $t$ is at most $1000k^2$, or
- Both $t'$ and $p$ have HD $< 20k$ to a periodic $x$; all $k$-matches start at position $1 + i \cdot |x|$.

![Diagram](https://via.placeholder.com/150)

- $t'$
- $x_i^*$
- $p$

$< 6k$ mism.
Main Result, Proof Overview

**Theorem (Structural Insight)**

For pattern $p$ and text $t$, $|t| \leq 2|p|$, at least one of the following holds:
- The number of $k$-matches of $p$ in $t$ is at most $1000k^2$, or
- Both $t'$ and $p$ have HD < $20k$ to a periodic $x$; all $k$-matches start at position $1 + i \cdot |x|$.

$$< 2 \cdot (6 + 1)k = 14k \text{ mism.}$$

$$< 6k \text{ mism.}$$
Main Result, Proof Overview

**Theorem (Structural Insight)**

For pattern $p$ and text $t$, $|t| \leq 2|p|$, at least one of the following holds:
- The number of $k$-matches of $p$ in $t$ is at most $1000k^2$, or
- Both $t'$ and $p$ have HD $< 20k$ to a periodic $x$; all $k$-matches start at position $1 + i \cdot |x|$.
Main Result, Proof Overview

**Theorem (Structural Insight)**

For pattern $p$ and text $t$, $|t| \leq 2|p|$, at least one of the following holds:

- The number of $k$-matches of $p$ in $t$ is at most $1000k^2$, or
- Both $t'$ and $p$ have HD $< 20k$ to a periodic $x$; all $k$-matches start at position $1 + i \cdot |x|$.
Main Result, Proof Overview

**Theorem (Structural Insight)**

For pattern $p$ and text $t$, $|t| \leq 2|p|$, at least one of the following holds:

- The number of $k$-matches of $p$ in $t$ is at most $1000k^2$, or
- Both $t'$ and $p$ have HD < $20k$ to a periodic $x$; all $k$-matches start at position $1 + i \cdot |x|$.

---

**Insight**

Any $k$-match of $p$ in $t'$ must match at least one $p_i$'s exactly.
Main Result, Proof Overview

**Theorem (Structural Insight)**

For pattern $p$ and text $t$, $|t| \leq 2|p|$, at least one of the following holds:

- The number of $k$-matches of $p$ in $t$ is at most $1000k^2$, or
- Both $t'$ and $p$ have $\text{HD} < 20k$ to a periodic $x$; all $k$-matches start at position $1 + i \cdot |x|$.

- Fix a $p_i$
Main Result, Proof Overview

**Theorem (Structural Insight)**

For pattern $p$ and text $t$, $|t| \leq 2|p|$, at least one of the following holds:

- The number of $k$-matches of $p$ in $t$ is at most $1000k^2$, or
- Both $t'$ and $p$ have $\text{HD} < 20k$ to a periodic $x$; all $k$-matches start at position $1 + i \cdot |x|$.

- Fix a $p_i$; count $k$-matches where $p_i$ is matched exactly.
Main Result, Proof Overview

**Theorem (Structural Insight)**

For pattern $p$ and text $t$, $|t| \leq 2|p|$, at least one of the following holds:
- The number of $k$-matches of $p$ in $t$ is at most $1000k^2$, or
- Both $t'$ and $p$ have HD $< 20k$ to a periodic $x$; all $k$-matches start at position $1 + i \cdot |x|$.

Consider occurrences of $x_i$ in $t'$
Main Result, Proof Overview

Theorem (Structural Insight)

For pattern $p$ and text $t$, $|t| \leq 2|p|$, at least one of the following holds:

- The number of $k$-matches of $p$ in $t$ is at most $1000k^2$, or
- Both $t'$ and $p$ have HD $< 20k$ to a periodic $x$; all $k$-matches start at position $1 + i \cdot |x|$.

Problem

Up to $O(m)$ exact matches of $x_i$ in $t'$.
Main Result, Proof Overview

**Theorem (Structural Insight)**

For pattern $p$ and text $t$, $|t| \leq 2|p|$, at least one of the following holds:

- The number of $k$-matches of $p$ in $t$ is at most $1000k^2$, or
- Both $t'$ and $p$ have HD $< 20k$ to a periodic $x$; all $k$-matches start at position $1 + i \cdot |x|$.

- Consider **power stretches** of $x_i$ in $t'$ of length $\geq |p_i|$.
Main Result, Proof Overview

**Theorem (Structural Insight)**
For pattern $p$ and text $t$, $|t| \leq 2|p|$, at least one of the following holds:
- The number of $k$-matches of $p$ in $t$ is at most $1000k^2$, or
- Both $t'$ and $p$ have HD $< 20k$ to a periodic $x$; all $k$-matches start at position $1 + i \cdot |x|$.

- Consider *power stretches* of $x_i$ in $t'$ of length $\geq |p_i|$
- $\leadsto$ at most $150k$ different power stretches
Main Result, Proof Overview

**Theorem (Structural Insight)**
For pattern \( p \) and text \( t \), \( |t| \leq 2|p| \), at least one of the following holds:
- The number of \( k \)-matches of \( p \) in \( t \) is at most \( 1000k^2 \), or
- Both \( t' \) and \( p \) have HD < 20\( k \) to a periodic \( x \); all \( k \)-matches start at position \( 1 + i \cdot |x| \).

- Fix a power stretch \( t_j \) of \( x_i \) in \( t' \).
Main Result, Proof Overview

**Theorem (Structural Insight)**

For pattern \( p \) and text \( t \), \(|t| \leq 2|p|\), at least one of the following holds:
- The number of \( k \)-matches of \( p \) in \( t \) is at most \( 1000k^2 \), or
- Both \( t' \) and \( p \) have HD < 20k to a periodic \( x \); all \( k \)-matches start at position \( 1 + i \cdot |x| \).

Fix a power stretch \( t_j \) of \( x_i \) in \( t' \).
Main Result, Proof Overview

**Theorem (Structural Insight)**

For pattern $p$ and text $t$, $|t| \leq 2|p|$, at least one of the following holds:
- The number of $k$-matches of $p$ in $t$ is at most $1000k^2$, or
- Both $t'$ and $p$ have HD $< 20k$ to a periodic $x$; all $k$-matches start at position $1 + i \cdot |x|$.

**Insight**

Must align at least one mismatch.
Main Result, Proof Overview

**Theorem (Structural Insight)**

For pattern $p$ and text $t$, $|t| \leq 2|p|$, at least one of the following holds:

- **The number of $k$-matches of $p$ in $t$ is at most** $1000k^2$, or
- Both $t'$ and $p$ have HD $< 20k$ to a periodic $x$; all $k$-matches start at position $1 + i \cdot |x|$.

**Insight**

At most $O(k^4)$ matches: $O(k)$ parts in $p$, $O(k)$ stretches, $O(k^2)$ matches per combination.
Main Result

**Theorem (Structural Insight) ✓**

For pattern $p$ and text $t$, $|t| \leq 2|p|$, at least one of the following holds:

- The number of $k$-matches of $p$ in $t$ is at most $O(k^2)$, or
- $t'$: shortest substring of $t$ such that any $k$-match of $p$ in $t$ is also a $k$-match in $t'$
  
  Both $t'$ and $p$ have Hamming distance $O(k)$ to the same periodic string $x$ and all $k$-matches of $p$ in $t'$ start at a position $1 + i \cdot |x|$. 
Faster Algorithm

**Theorem (Algorithm)**

Pattern matching with $k$ mismatches on a text $t$ given by an SLP of size $n$ and a pattern $p$ of length $m$ can be solved in time $O(n k^3 (k \log k + \log m) + k m)$. 

Karl Bringmann, Marvin Künnemann, and Philip Wellnitz

Faster Pattern Matching with Mismatches in Compressed Texts
**Theorem (Algorithm)**

Pattern matching with $k$ mismatches on a text $t$ given by an SLP of size $n$ and a pattern $p$ of length $m$ can be solved in time $O(n k^3 (k \log k + \log m) + k m)$.

**Pattern-Compressed String [GS’13]**

Let $p$ be a string of length $m$. We call a string $f = v_1 \ldots v_q, \sum_{i=1}^{q} |v_i| \leq 2m$ a $p$-pattern-compressed string (pc-string) if every $v_i$ is a substring of $p$. We call the $v_i$’s factors of $f$. 
Faster Algorithm for Pattern-Compressed Strings

Pattern-Compressed String [GS’13]

Let $p$ be a string of length $m$. We call a string $f = v_1 \ldots v_q$, $\sum_{i=1}^{q} |v_i| \leq 2m$ a $p$-pattern-compressed string (pc-string) if every $v_i$ is a substring of $p$. We call the $v_i$’s factors of $f$. 
Faster Algorithm for Pattern-Compressed Strings

Pattern-Compressed String [GS’13]

Let $p$ be a string of length $m$. We call a string $f = v_1 \ldots v_q$, $\sum_{i=1}^{q} |v_i| \leq 2m$ a $p$-pattern-compressed string (pc-string) if every $v_i$ is a substring of $p$. We call the $v_i$’s factors of $f$.

![Diagram of SLP instance](image)
Faster Algorithm for Pattern-Compressed Strings

Pattern-Compressed String [GS’13]

Let $p$ be a string of length $m$. We call a string $f = v_1 \ldots v_q, \sum_{i=1}^q |v_i| \leq 2m$ a $p$-pattern-compressed string (pc-string) if every $v_i$ is a substring of $p$. We call the $v_i$’s factors of $f$.

**SLP Instance $I$**

```
\[
\begin{array}{c}
\text{SLP } & T_1 & T_2 & T_3 \\
\text{Instance } I & A & B \\
\end{array}
\]

**SLP** $\mathcal{T} = T_1, \ldots, T_n$

$k, p$ of length $m$

**PC-String, inst. $J_1$**

$k, p, f_1$ with $O(k)$ factors

**PC-String, inst. $J_2$**

$k, p, f_2$ with $O(k)$ factors

\[O(n k^3 (k \log k + \log m) + k m)\]

**PC-String, inst. $J_n$**

$k, p, f_n$ with $O(k)$ factors

\[O(k^3 (k \log k + \log m))\]

Karl Bringmann, Marvin Künnemann, and Philip Wellnitz

Faster Pattern Matching with Mismatches in Compressed Texts
Faster Algorithm for Pattern-Compressed Strings

Pattern-Compressed String [GS’13]

Let $p$ be a string of length $m$. We call a string $f = v_1 \ldots v_q, \sum_{i=1}^{q} |v_i| \leq 2m$ a $p$-pattern-compressed string (pc-string) if every $v_i$ is a substring of $p$. We call the $v_i$’s factors of $f$.

SLP Instance $I$

\[
\begin{array}{c}
\text{SLP} \\
\text{Instance } I \\
T_1, T_2, T_3 \\
A, B \\
\text{SLP } \mathcal{T} = T_1, \ldots, T_n \\
k, p \text{ of length } m
\end{array}
\]

$O(nk^3(k \log k + \log m) + km)$ algorithm

PC-String, inst. $J_1$

$k, p, f_1$ with $O(k)$ factors

PC-String, inst. $J_2$

$k, p, f_2$ with $O(k)$ factors

...$

PC-String, inst. J_n$

$k, p, f_n$ with $O(k)$ factors

$O(k^3(k \log k + \log m))$ algorithm

Karl Bringmann, Marvin Künnemann, and Philip Wellnitz

Faster Pattern Matching with Mismatches in Compressed Texts
Faster Algorithm for Pattern-Compressed Strings

**Theorem (Algorithm for pc-strings)**

Pattern matching with $k$ mismatches on a pattern $p$ of length $m$ and a $p$-pc-string $f$ of size $O(k)$ representing at most $2m$ characters, can be solved in time $O(k^3(k \log k + \log m))$.

(With $O(km)$ preprocessing on $p$.)

- Implementation of structural insight
- Need e.g. tools for finding first $O(k)$ mismatches to a periodic string or finding all power stretches of a given string in a pc-string
Faster Algorithm for Pattern-Compressed Strings

**Theorem (Algorithm for pc-strings)**

Pattern matching with $k$ mismatches on a pattern $p$ of length $m$ and a $p$-pc-string $f$ of size $O(k)$ representing at most $2m$ characters, can be solved in time $O(k^3(k \log k + \log m))$.

(With $O(km)$ preprocessing on $p$.)

- Implementation of structural insight
- Need e.g. tools for finding first $O(k)$ mismatches to a periodic string or finding all power stretches of a given string in a pc-string
Faster Algorithm for Pattern-Compressed Strings

**Theorem (Algorithm for pc-strings)**

Pattern matching with \( k \) mismatches on a pattern \( p \) of length \( m \) and a \( p \)-pc-string \( f \) of size \( O(k) \) representing at most \( 2m \) characters, can be solved in time \( O(k^3(k \log k + \log m)) \).

(With \( O(km) \) preprocessing on \( p \).)

- Implementation of structural insight
- Need e.g. tools for finding first \( O(k) \) mismatches to a periodic string or finding all power stretches of a given string in a pc-string
Faster Algorithm

**Theorem (Algorithm)**

Pattern matching with $k$ mismatches on a text $t$ given by an SLP of size $n$ and a pattern $p$ of length $m$ can be solved in time $O(n k^3 (k \log k + \log m) + km)$. 

Karl Bringmann, Marvin Künnemann, and Philip Wellnitz

Faster Pattern Matching with Mismatches in Compressed Texts
Open Problems
Open Problems

- Improve insight to $O(k)$ mismatches in the aperiodic case

---

**Theorem (Structural Insight') [KW'19+]**

For pattern $p$ and text $t$, $|t| \leq 2|p|$, it holds at least one of:

- The number of $k$-matches of $p$ in $t$ is at most $O(k)$, or
- $t'$: shortest substring of $t$ such that any $k$-match of $p$ in $t$ is also a $k$-match in $t'$. Both $t'$ and $p$ have Hamming distance $O(k)$ to the same periodic string $x$ and all $k$-matches of $p$ in $t'$ start at a position $1 + i \cdot |x|$.
Open Problems

- Improve insight to $O(k)$ mismatches in the aperiodic case
- Improve dependence on $k$ in the algorithm

**Theorem (Algorithm)**

Pattern matching with $k$ mismatches on a text $t$ given by an SLP of size $n$ and a pattern $p$ of length $m$ can be solved in time $O(n k^3 (k \log k + \log m) + k m)$. 
Open Problems

- Improve insight to $O(k)$ mismatches in the aperiodic case
- Improve dependence on $k$ in the algorithm
- Fully compressed setting ($p$ also given as an SLP)
- Pattern Matching with Errors (Edit distance instead of Hamming distance)
Solution Structure of Pattern Matching with Mismatches

\[
\begin{align*}
  &A^{2m} \\
  t &= \text{A \cdots A A A \cdots A A A \cdots A A} \\
  &A^m \\
  p &= \text{A \cdots A A \cdots A A \cdots A A} 
\end{align*}
\]
Solution Structure of Pattern Matching with Mismatches

$A^{2m/3-1}$ \quad $A^{2m/3-1}$ \quad $A^{2m/3}$

$\begin{array}{c}
    t \\
    A \cdot A \cdot B \cdot A \cdot A \cdot B \cdot A \cdot A \\
\end{array}$

B at $2m/3$, $4m/3$ in $t$ and the middle $k + 1$ positions in $p$

$\begin{array}{c}
    p \\
    A \cdot A \cdot B \cdot B \cdot B \cdot A \cdot A \\
\end{array}$

$A^{(m-k-1)/2}$ \quad $B^{k+1}$ \quad $A^{(m-k-1)/2}$

Karl Bringmann, Marvin Künnemann, and Philip Wellnitz
Faster Pattern Matching with Mismatches in Compressed Texts
Solution Structure of Pattern Matching with Mismatches

- All matches start at the union of two intervals.
Solution Structure of Pattern Matching with Mismatches

Insight 3
Arithmetic progression only approximates all matches
Main Result

**Theorem (Structural Insight)**

Given strings $p$ of length $m$ and $t$ of length at most $2m$, at least one of the following holds:

- The number of $k$-matches of $p$ in $t$ is at most $O(k^2)$.
- $t'$: shortest substring of $t$ such that any $k$-match of $p$ in $t$ is also a $k$-match in $t'$
  There is a substring $x$ of $p$, with $|x| = O(m/k)$, such that
  $\delta_H(p, x^*[1, m]) \leq O(k)$ and $\delta_H(t', x^*[1, |t'|]) \leq O(k)$.

Moreover, any $k$-match of $p$ in $t'$ starts at a position of the form $1 + i \cdot |x|$ with $0 \leq i \leq (|t'| - |p|)/|x|$ (but not every starting position $1 + i \cdot |x|$ necessarily yields a $k$-match).
Main Result

**Theorem (Structural Insight)**

Given strings \( p \) of length \( m \) and \( t \) of length at most \( 2m \), at least one of the following holds:

- The number of \( k \)-matches of \( p \) in \( t \) is at most \( O(k^2) \).
- \( t' \): shortest substring of \( t \) such that any \( k \)-match of \( p \) in \( t \) is also a \( k \)-match in \( t' \)
  
  There is a substring \( x \) of \( p \), with \( |x| = O(m/k) \), such that 
  
  \[ \delta_H(p, x^*[1, m]) \leq O(k) \quad \text{and} \quad \delta_H(t', x^*[1, |t'|]) \leq O(k). \]

  Moreover, any \( k \)-match of \( p \) in \( t' \) starts at a position of the form \( 1 + i \cdot |x| \) 
  
  with \( 0 \leq i \leq (|t'| - |p|)/|x| \) (but not every starting position \( 1 + i \cdot |x| \) 
  
  necessarily yields a \( k \)-match).
Main Result

**Theorem (Structural Insight)**

Given strings $p$ of length $m$ and $t$ of length at most $2m$, at least one of the following holds:

- The number of $k$-matches of $p$ in $t$ is at most $O(k^2)$.

- $t'$: shortest substring of $t$ such that any $k$-match of $p$ in $t$ is also a $k$-match in $t'$
  
  There is a substring $x$ of $p$, with $|x| = O(m/k)$, such that $\delta_H(p, x[1, m]) \leq O(k)$ and $\delta_H(t', x[1, |t'|]) \leq O(k)$.

  Moreover, any $k$-match of $p$ in $t'$ starts at a position of the form $1 + i \cdot |x|$ with $0 \leq i \leq (|t'| - |p|)/|x|$ (but not every starting position $1 + i \cdot |x|$ necessarily yields a $k$-match).
Main Result

**Theorem (Structural Insight)**

Given strings $p$ of length $m$ and $t$ of length at most $2m$, at least one of the following holds:

- The number of $k$-matches of $p$ in $t$ is at most $O(k^2)$.
- $t'$: shortest substring of $t$ such that any $k$-match of $p$ in $t$ is also a $k$-match in $t'$

There is a substring $x$ of $p$, with $|x| = O(m/k)$, such that

$$\delta_H(p, x^*[1, m]) \leq O(k)$$

and

$$\delta_H(t', x^*[1, |t'|]) \leq O(k).$$

Moreover, any $k$-match of $p$ in $t'$ starts at a position of the form $1 + i \cdot |x|$ with $0 \leq i \leq (|t'| - |p|)/|x|$ (but not every starting position $1 + i \cdot |x|$ necessarily yields a $k$-match).
<table>
<thead>
<tr>
<th>Theorem (Structural Insight)</th>
</tr>
</thead>
<tbody>
<tr>
<td>For pattern $p$ and text $t$, $</td>
</tr>
<tr>
<td>- The number of $k$-matches of $p$ in $t$ is at most $O(k^2)$, and</td>
</tr>
<tr>
<td>- Both $t$ and $p$ have HD $O(k)$ to the same periodic string.</td>
</tr>
</tbody>
</table>
Main Result

**Theorem (Structural Insight)**

For pattern $p$ and text $t$, $|t| \leq 2|p|$, it holds at least one of:
- The number of $k$-matches of $p$ in $t$ is at most $O(k^2)$, and
- Both $t$ and $p$ have HD $O(k)$ to the same periodic string.

Finding $ANPAN$, $k = 2$

non-periodic case

Finding $PANPAN$, $k = 2$

periodic case
Main Result

**Theorem (Structural Insight)**

For pattern $p$ and text $t$, $|t| \leq 2|p|$, it holds at least one of:
- The number of $k$-matches of $p$ in $t$ is at most $O(k^2)$, and
- Both $t$ and $p$ have HD $O(k)$ to the same periodic string.

Finding ANPAN, $k = 2$

non-periodic case

Finding PANPAN, $k = 2$

periodic case

Karl Bringmann, Marvin Künnemann, and Philip Wellnitz

Faster Pattern Matching with Mismatches in Compressed Texts
Main Result

**Theorem (Structural Insight)**

For pattern $p$ and text $t$, $|t| \leq 2|p|$, it holds at least one of:

- The number of $k$-matches of $p$ in $t$ is at most $O(k^2)$, and
- Both $t$ and $p$ have HD $O(k)$ to the same periodic string.

Finding $\text{ANPAN}$, $k = 2$  
non-periodic case

Finding $\text{PANPAN}$, $k = 2$  
periodic case
Main Result

Theorem (Structural Insight)
For pattern $p$ and text $t$, $|t| \leq 2|p|$, it holds at least one of:
- The number of $k$-matches of $p$ in $t$ is at most $O(k^2)$, and
- Both $t$ and $p$ have HD $O(k)$ to the same periodic string.

Finding ANPAN, $k = 2$
non-periodic case

Finding PANPAN, $k = 2$
periodic case

Karl Bringmann, Marvin Künnemann, and Philip Wellnitz
Faster Pattern Matching with Mismatches in Compressed Texts
Main Result, Proof Overview

**Theorem (Structural Insight)**

For pattern $p$ and text $t$, $|t| \leq 2|p|$, it holds at least one of:

- The number of $k$-matches of $p$ in $t$ is at most $O(k^2)$, and
- Both $t$ and $p$ have HD $O(k)$ to the same periodic string.
Main Result, Proof Overview

**Theorem (Structural Insight)**

Fix a pattern $p$ of length $m$ and a text $t$ of length at most $2m$. If the number of $k$-matches of $p$ in $t$ is at least $1000k^2$, then both $t$ and $p$ have a HD $< 20k$ to the same periodic string.
Main Result, Proof Overview

**Theorem (Structural Insight)**

Fix a pattern $p$ of length $m$ and a text $t$ of length at most $2m$.
If the number of $k$-matches of $p$ in $t$ is at least $1000k^2$, then both $t$ and $p$ have a HD $< 20k$ to the same periodic string.

Main Steps:

- At least $1000k^2$ $k$-matches of $p$ in $t$ and $p$ has a HD $< 6k$ to a specific periodic string $x \in x(p)$.
  $$\implies t \text{ has a Hamming Distance } < 20k \text{ to } x$$

- $p$ has HD $\geq 6k$ to any specific periodic string $x \in x(p)$.
  $$\implies \text{Less than } 1000k^2 \text{ } k\text{-matches of } p \text{ in } t$$
Main Result, Proof Overview

**Theorem (Structural Insight)**

Fix a pattern $p$ of length $m$ and a text $t$ of length at most $2m$. If the number of $k$-matches of $p$ in $t$ is at least $1000k^2$, then both $t$ and $p$ have a HD $< 20k$ to the same periodic string.

Main Steps:

- At least $1000k^2$ $k$-matches of $p$ in $t$ and $p$ has a HD $< 6k$ to a specific periodic string $x \in x(p)$
  \[ \implies t \text{ has a Hamming Distance } < 20k \text{ to } x \]

- $p$ has HD $\geq 6k$ to any specific periodic string $x \in x(p)$
  \[ \implies \text{Less than } 1000k^2 \text{ } k\text{-matches of } p \text{ in } t \]
Main Result, Proof Overview

Lemma (Step 1)
Fix a pattern $p$ of length $m$ and a text $t$ of length at most $2m$. If the number of $k$-matches of $p$ in $t$ is at least $1000k^2$, and $p$ has HD $< 6k$ to a periodic string $x \in x(p)$, then $t$ has HD $< 20k$ to $x$. 
Main Result, Proof Overview

**Lemma (Step 1)**
Fix a pattern $p$ of length $m$ and a text $t$ of length at most $2m$. If the number of $k$-matches of $p$ in $t$ is at least $1000k^2$, and $p$ has HD < $6k$ to a periodic string $x \in x(p)$, then $t$ has HD < $20k$ to $x$.

$t$

$p$
Main Result, Proof Overview

**Lemma (Step 1)**

Fix a pattern $p$ of length $m$ and a text $t$ of length at most $2m$. If the number of $k$-matches of $p$ in $t$ is at least $1000k^2$, and $p$ has $\text{HD} < 6k$ to a periodic string $x \in x(p)$, then $t$ has $\text{HD} < 20k$ to $x$.

- Split $p$ into $16k$ parts $p_i$ of equal length
Main Result, Proof Overview

Lemma (Step 1)
Fix a pattern $p$ of length $m$ and a text $t$ of length at most $2m$.
If the number of $k$-matches of $p$ in $t$ is at least $1000k^2$, and $p$ has HD $< 6k$ to a periodic string $x \in x(p)$, then $t$ has HD $< 20k$ to $x$.

- Split $p$ into $16k$ parts $p_i$ of equal length
Main Result, Proof Overview

**Lemma (Step 1)**
Fix a pattern $p$ of length $m$ and a text $t$ of length at most $2m$. If the number of $k$-matches of $p$ in $t$ is at least $1000k^2$, and $p$ has HD $< 6k$ to a periodic string $x \in x(p)$, then $t$ has HD $< 20k$ to $x$.

$\blacksquare$ Fix a $p_i$
Main Result, Proof Overview

**Lemma (Step 1)**
Fix a pattern $p$ of length $m$ and a text $t$ of length at most $2m$.
If the number of $k$-matches of $p$ in $t$ is at least $1000k^2$, and $p$ has $HD < 6k$ to a periodic string $x \in x(p)$, then $t$ has $HD < 20k$ to $x$.

- Consider prefix $x_i$ of $p_i$ that is also a period of $p_i$
**Lemma (Step 1)**

Fix a pattern $p$ of length $m$ and a text $t$ of length at most $2m$.
If the number of $k$-matches of $p$ in $t$ is at least $1000k^2$, and $p$ has $HD < 6k$ to a periodic string $x \in x(p)$, then $t$ has $HD < 20k$ to $x$.

- Find first $3k$ mismatches between $p$ and $x_i^*$ before and after $p_i$
Main Result, Proof Overview

**Lemma (Step 1)**

Fix a pattern $p$ of length $m$ and a text $t$ of length at most $2m$.
If the number of $k$-matches of $p$ in $t$ is at least $1000k^2$, and $p$ has HD $< 6k$ to a periodic string $x \in x(p)$, then $t$ has HD $< 20k$ to $x$.

- Find first $3k$ mismatches between $p$ and $x_i^*$ before and after $p_i$.
Main Result, Proof Overview

Lemma (Step 1)

Fix a pattern $p$ of length $m$ and a text $t$ of length at most $2m$. If the number of $k$-matches of $p$ in $t$ is at least $1000k^2$, and $p$ has HD $< 6k$ to some $x^*_i$, $1 \leq i \leq 16k$, then $t$ has HD $< 20k$ to $x^*_i$. 

\[
\begin{array}{c}
\times \\
x^*_i \\
p \\
\times \times \times \times \times \times \times \times \times \\
\times \times \times \\
\times \\
\times \times \times \times \times \times \\
x \times x_i \\
\end{array}
\]

Karl Bringmann, Marvin Künnemann, and Philip Wellnitz
Faster Pattern Matching with Mismatches in Compressed Texts
Main Result, Proof Overview

Lemma (Step 1)
Fix a pattern $p$ of length $m$ and a text $t$ of length at most $2m$. If the number of $k$-matches of $p$ in $t$ is at least $1000k^2$, and $p$ has HD < $6k$ to some $x_i^*$, $1 \leq i \leq 16k$, then $t$ has HD < $20k$ to $x_i^*$.

Claim (Proof omitted)
If there are at least $2 + 16k$ $k$-matches of $p$ in $t$, all starting positions of $k$-matches differ by (integer) multiples of $|x_i|$. 
Main Result, Proof Overview

**Lemma (Step 1)**

Fix a pattern $p$ of length $m$ and a text $t$ of length at most $2m$. If the number of $k$-matches of $p$ in $t$ is at least $1000k^2$, and $p$ has $\text{HD} < 6k$ to some $x^*_i$, $1 \leq i \leq 16k$, then all starting positions of $k$-matches differ by multiples of $|x^*_i|$ and $t$ has $\text{HD} < 20k$ to $x^*_i$. 
Lemma (Step 1)

Fix a pattern $p$ of length $m$ and a text $t$ of length at most $2m$.
If the number of $k$-matches of $p$ in $t$ is at least $1000k^2$, and $p$ has HD $< 6k$ to some $x_i^*, 1 \leq i \leq 16k$, then all starting positions of $k$-matches differ by multiples of $|x_i|$ and $t$ has HD $< 20k$ to $x_i^*$. 

$t$

$p$

Karl Bringmann, Marvin Künnemann, and Philip Wellnitz
Faster Pattern Matching with Mismatches in Compressed Texts
**Lemma (Step 1)**

Fix a pattern $p$ of length $m$ and a text $t$ of length at most $2m$. If the number of $k$-matches of $p$ in $t$ is at least $1000k^2$, and $p$ has HD $< 6k$ to some $x_i^*$, $1 \leq i \leq 16k$, then all starting positions of $k$-matches differ by multiples of $|x_i|$ and $t$ has HD $< 20k$ to $x_i^*$. 
Main Result, Proof Overview

**Lemma (Step 1)**

Fix a pattern $p$ of length $m$ and a text $t$ of length at most $2m$. If the number of $k$-matches of $p$ in $t$ is at least $1000k^2$, and $p$ has HD $< 6k$ to some $x^*_i$, $1 \leq i \leq 16k$, then all starting positions of $k$-matches differ by multiples of $|x_i|$ and $t$ has HD $< 20k$ to $x^*_i$. 

![Diagram of pattern and text with mismatches highlighted.]
Main Result, Proof Overview

**Lemma (Step 1)**

Fix a pattern $p$ of length $m$ and a text $t$ of length at most $2m$. If the number of $k$-matches of $p$ in $t$ is at least $1000k^2$, and $p$ has HD $< 6k$ to some $x_i^*$, $1 \leq i \leq 16k$, then all starting positions of $k$-matches differ by multiples of $|x_i|$ and $t$ has HD $< 20k$ to $x_i^*$.

![Diagram of the lemma](image-url)
Main Result, Proof Overview

**Lemma (Step 1)**

Fix a pattern $p$ of length $m$ and a text $t$ of length at most $2m$. If the number of $k$-matches of $p$ in $t$ is at least $1000k^2$, and $p$ has HD < $6k$ to some $x_i^*$, $1 \leq i \leq 16k$, then all starting positions of $k$-matches differ by multiples of $|x_i|$ and $t$ has HD < $20k$ to $x_i^*$.

**Lemma (Step 2)**

Fix a pattern $p$ of length $m$ and a text $t$ of length at most $2m$. If the pattern $p$ has a HD $\geq 6k$ to all strings $x_i^*$, $1 \leq i \leq 16k$, then there are less than $1000k^2$ $k$-matches of $p$ in $t$. 
Main Result, Proof Overview

Lemma (Step 2)

Fix a pattern $p$ of length $m$ and a text $t$ of length at most $2m$. If the pattern $p$ has a HD $\geq 6k$ to all strings $x_i^*$, $1 \leq i \leq 16k$, then there are less than $1000k^2$ $k$-matches of $p$ in $t$. 

Karl Bringmann, Marvin Künnemann, and Philip Wellnitz
Faster Pattern Matching with Mismatches in Compressed Texts
Main Result, Proof Overview

**Lemma (Step 2)**

Fix a pattern $p$ of length $m$ and a text $t$ of length at most $2m$. If the pattern $p$ has a HD $\geq 6k$ to all strings $x_i^*$, $1 \leq i \leq 16k$, then there are less than $1000k^2$ $k$-matches of $p$ in $t$.

- Recall: Split $p$ into $16k$ parts $p_i$ of equal length
Main Result, Proof Overview

**Lemma (Step 2)**

Fix a pattern $p$ of length $m$ and a text $t$ of length at most $2m$.
If the pattern $p$ has a HD $\geq 6k$ to all strings $x_i^*, 1 \leq i \leq 16k$, then there are less than $1000k^2$ $k$-matches of $p$ in $t$.

```
p
  p1   p2   ...   pi   ...   p16k
```

**Insight**

Any $k$-match of $p$ in $t$ must match at least $15k$ $p_i$'s exactly.
Main Result, Proof Overview

**Lemma (Step 2)**

Fix a pattern $p$ of length $m$ and a text $t$ of length at most $2m$. If the pattern $p$ has a HD $\geq 6k$ to all strings $x^*_i, 1 \leq i \leq 16k$, then there are less than $1000k^2$ $k$-matches of $p$ in $t$.

- Fix a $p_i$
Main Result, Proof Overview

**Lemma (Step 2)**

Fix a pattern $p$ of length $m$ and a text $t$ of length at most $2m$. If the pattern $p$ has a HD $\geq 6k$ to all strings $x_i^*$, $1 \leq i \leq 16k$, then there are less than $1000k^2$ $k$-matches of $p$ in $t$.

- Fix a $p_i$; count $k$-matches where $p_i$ is matched exactly
Main Result, Proof Overview

**Lemma (Step 2)**

Fix a pattern $p$ of length $m$ and a text $t$ of length at most $2m$. If the pattern $p$ has a HD $\geq 6k$ to all strings $x_i^*$, $1 \leq i \leq 16k$, then there are less than $1000k^2$ $k$-matches of $p$ in $t$.

- Search for $x_i$ in $t$
Main Result, Proof Overview

Lemma (Step 2)
Fix a pattern $p$ of length $m$ and a text $t$ of length at most $2m$.
If the pattern $p$ has a HD $\geq 6k$ to all strings $x_i^*$, $1 \leq i \leq 16k$, then there are less than $1000k^2$ $k$-matches of $p$ in $t$.

Problem
Up to $O(m)$ exact matches of $x_i$ in $t$. 
Main Result, Proof Overview

Lemma (Step 2)

Fix a pattern $p$ of length $m$ and a text $t$ of length at most $2m$. If the pattern $p$ has a HD $\geq 6k$ to all strings $x_i^*, 1 \leq i \leq 16k$, then there are less than $1000k^2$ $k$-matches of $p$ in $t$.

- Search for power stretches of $x_i$ in $t$ of length $\geq |p_i|$
Main Result, Proof Overview

Lemma (Step 2)
Fix a pattern $p$ of length $m$ and a text $t$ of length at most $2m$. If the pattern $p$ has a HD $\geq 6k$ to all strings $x_i^*$, $1 \leq i \leq 16k$, then there are less than $1000k^2$ $k$-matches of $p$ in $t$.

Insight
Only $\leq 150k$ different power stretches of $x_i$ in $t$. 
Main Result, Proof Overview

**Lemma (Step 2)**

Fix a pattern $p$ of length $m$ and a text $t$ of length at most $2m$. If the pattern $p$ has a HD $\geq 6k$ to all strings $x_i^*$, $1 \leq i \leq 16k$, then there are less than $1000k^2 k$-matches of $p$ in $t$.

- Fix a power stretch $t_j$ of $x_i$ in $t$. 
Main Result, Proof Overview

**Lemma (Step 2)**

Fix a pattern $p$ of length $m$ and a text $t$ of length at most $2m$. If the pattern $p$ has a HD $\geq 6k$ to all strings $x_i^*$, $1 \leq i \leq 16k$, then there are less than $1000k^2$ $k$-matches of $p$ in $t$.

- Fix a power stretch $t_j$ of $x_i$ in $t$. 
Main Result, Proof Overview

**Lemma (Step 2)**

Fix a pattern $p$ of length $m$ and a text $t$ of length at most $2m$.
If the pattern $p$ has a HD $\geq 6k$ to all strings $x_i^*$, $1 \leq i \leq 16k$, then there are less than $1000k^2$ $k$-matches of $p$ in $t$. 

\[ t \]
\[ p \]

Karl Bringmann, Marvin Künnemann, and Philip Wellnitz

Faster Pattern Matching with Mismatches in Compressed Texts
Main Result, Proof Overview

**Lemma (Step 2)**

Fix a pattern $p$ of length $m$ and a text $t$ of length at most $2m$.

If the pattern $p$ has a HD $\geq 6k$ to all strings $x_i^*$, $1 \leq i \leq 16k$, then there are less than $1000k^2$ $k$-matches of $p$ in $t$.
Main Result, Proof Overview

\[ t \]

\[ x_i^* \]

\[ p \]

\[ \geq 2k \text{ mism.} \]

\[ \geq 3k \text{ mism.} \]
Main Result, Proof Overview

\[ t \]

\[ x_i^* \]

\[ p \]

Insight

Must align at least \( k \) mismatches.
Main Result, Proof Overview

\[ t \quad x_i \quad x_i \quad x_i \quad x_i \quad \]
\[ x_i^* \quad x_i \quad x_i \quad x_i \quad x_i \quad x_i \quad x_i \quad x_i \quad x_i \quad x_i \]
\[ p \quad x_i \quad x_i \quad \]
\[ p_i \quad \geq 3k \text{ mism.} \]
\[ t_j \quad \geq 2k \text{ mism.} \]

Insight
At most \( O(k^4) \) matches: \( O(k) \) parts in \( p \), \( O(k) \) streches, \( O(k^2) \) matches per combination.
Main Result, Proof Overview

**Lemma (Step 1)** ✓

Fix a pattern $p$ of length $m$ and a text $t$ of length at most $2m$. If the number of $k$-matches of $p$ in $t$ is at least $1000k^2$, and $p$ has HD $< 6k$ to some $x_i^*$, $1 \leq i \leq 16k$, then all starting positions of $k$-matches differ by multiples of $|x_i|$ and $t$ has HD $< 20k$ to $x_i^*$.

**Lemma (Step 2)** ✓

Fix a pattern $p$ of length $m$ and a text $t$ of length at most $2m$. If the pattern $p$ has a HD $\geq 6k$ to all strings $x_i^*$, $1 \leq i \leq 16k$, then there are less than $1000k^2$ $k$-matches of $p$ in $t$. 
Main Result, Proof Overview

**Theorem (Structural Insight)**

Given strings $p$ of length $m$ and $t$ of length at most $2m$, at least one of the following holds:

- The number of $k$-matches of $p$ in $t$ is at most $O(k^2)$.

- $t'$: shortest substring of $t$ such that any $k$-match of $p$ in $t$ is also a $k$-match in $t'$

  There is a substring $x$ of $p$, with $|x| = O(m/k)$, such that

  $\delta_H(p, x^*[1, m]) \leq O(k)$ and $\delta_H(t', x^*[1, |t'|]) \leq O(k)$.

Moreover, any $k$-match of $p$ in $t'$ starts at a position of the form $1 + i \cdot |x|$ with $0 \leq i \leq (|t'| - |p|)/|x|$ (but not every starting position $1 + i \cdot |x|$ necessarily yields a $k$-match).
Faster Algorithm for Pattern-Compressed Strings

Pattern-Compressed String [GS’13]

Let $p$ be a string of length $m$. We call a string $f = v_1 \ldots v_q, \sum_{i=1}^{q} |v_i| \leq 2m$ a $p$-pattern-compressed string (pc-string) if every $v_i$ is a substring of $p$. We call the $v_i$’s factors of $f$. 
Faster Algorithm for Pattern-Compressed Strings

Pattern-Compressed String [GS’13]

Let $p$ be a string of length $m$. We call a string $f = v_1 \ldots v_q$, $\sum_{i=1}^{q} |v_i| \leq 2m$ a $p$-pattern-compressed string (pc-string) if every $v_i$ is a substring of $p$. We call the $v_i$’s factors of $f$.

SLP Instance $I$

PC-String, inst. $J_1$
$k, p, f_1$ with $O(k)$ factors

PC-String, inst. $J_2$
$k, p, f_2$ with $O(k)$ factors

PC-String, inst. $J_n$
$k, p, f_n$ with $O(k)$ factors

$O(n (\log m + T(m, k)) + km)$ algorithm

$T(m, k)$ algorithm

Karl Bringmann, Marvin Künnemann, and Philip Wellnitz
Faster Pattern Matching with Mismatches in Compressed Texts
Faster Algorithm for Pattern-Compressed Strings

Pattern-Compressed String [GS’13]

Let $p$ be a string of length $m$. We call a string $f = v_1 \ldots v_q, \sum_{i=1}^{q} |v_i| \leq 2m$ a $p$-pattern-compressed string (pc-string) if every $v_i$ is a substring of $p$. We call the $v_i$’s factors of $f$.

<table>
<thead>
<tr>
<th>SLP Instance $I$</th>
<th>PC-String, inst. $J_1$</th>
<th>$k, p, f_1$ with $O(k)$ factors</th>
</tr>
</thead>
<tbody>
<tr>
<td>SLP $T = T_1, \ldots, T_n$ $k, p$ of length $m$</td>
<td>PC-String, inst. $J_2$ $k, p, f_2$ with $O(k)$ factors</td>
<td></td>
</tr>
<tr>
<td>$O(n k^3 (k \log k + \log m) + k m)$ algorithm</td>
<td>$O(k^3(k \log k + \log m))$ algorithm</td>
<td></td>
</tr>
</tbody>
</table>

Karl Bringmann, Marvin Künnemann, and Philip Wellnitz
Faster Pattern Matching with Mismatches in Compressed Texts
Faster Algorithm for Pattern-Compressed Strings

Pattern-Compressed String [GS’13]

Let $p$ be a string of length $m$. We call a string $f = v_1 \ldots v_q$, $\sum_{i=1}^{q} |v_i| \leq 2m$ a $p$-pattern-compressed string (pc-string) if every $v_i$ is a substring of $p$. We call the $v_i$’s factors of $f$.

SLP Instance $I$

$T = T_1, \ldots, T_n$

$k, p$ of length $m$

$O(n k^3 (k \log k + \log m) + k m)$

algorithm

PC-String, inst. $J_1$ $k, p, f_1$ with $O(k)$ factors

PC-String, inst. $J_2$ $k, p, f_2$ with $O(k)$ factors

PC-String, inst. $J_n$ $k, p, f_n$ with $O(k)$ factors

$O(k^3(k \log k + \log m))$

algorithm

Karl Bringmann, Marvin Künnemann, and Philip Wellnitz

Faster Pattern Matching with Mismatches in Compressed Texts
Faster Algorithm for Pattern-Compressed Strings

**Theorem (Algorithm for pc-strings)**

Pattern matching with $k$ mismatches on a pattern $p$ of length $m$ and a $p$-pc-string $f$ of size $O(k)$ representing at most $2m$ characters, can be solved in time $O(k^3(k \log k + \log m))$. (With $O(km)$ preprocessing on $p$.)

- Implementation of structural insight
- Need e.g. tools for finding first $O(k)$ mismatches to a periodic string or finding all power stretches of a given string in a pc-string
Theorem (Algorithm for pc-strings)

Pattern matching with $k$ mismatches on a pattern $p$ of length $m$ and a $p$-pc-string $f$ of size $O(k)$ representing at most $2m$ characters, can be solved in time $O(k^3(k \log k + \log m))$.
(With $O(km)$ preprocessing on $p$.)

- Implementation of structural insight
  - Need e.g. tools for finding first $O(k)$ mismatches to a periodic string or finding all power stretches of a given string in a pc-string
Faster Algorithm for Pattern-Compressed Strings

**Theorem (Algorithm for pc-strings)**

Pattern matching with \( k \) mismatches on a pattern \( p \) of length \( m \) and a \( p \)-pc-string \( f \) of size \( O(k) \) representing at most \( 2m \) characters, can be solved in time \( O(k^3(k \log k + \log m)) \).
(With \( O(km) \) preprocessing on \( p \).)

- Implementation of structural insight
- Need e.g. tools for finding first \( O(k) \) mismatches to a periodic string or finding all power stretches of a given string in a pc-string