A Polynomial Time Match Test for Large Classes of Extended Regular Expressions

Daniel Reidenbach, **Markus L. Schmid**,
Loughborough University UK

CIAA 2010, Winnipeg, Canada

Outline

Extended regular expressions (REGEX)

- A mechanism to define formal languages.

Extended regular expressions (REGEX)

- A mechanism to define formal languages.
- Enhancement of “standard” regular expressions.

Extended regular expressions (REGEX)

- A mechanism to define formal languages.
- Enhancement of "standard" regular expressions.
- Important elements:

Extended regular expressions (REGEX)

- A mechanism to define formal languages.
- Enhancement of “standard” regular expressions.
- Important elements:
  - Regular operations,

Extended regular expressions (REGEX)

- A mechanism to define formal languages.
- Enhancement of “standard” regular expressions.
- Important elements:
    - Regular operations,
    - back references.

Extended regular expressions (REGEX)

- A mechanism to define formal languages.
- Enhancement of “standard” regular expressions.
- Important elements:
  - Regular operations,
  - back references.
- Practical application.

Match Test

- Is a word w in a REGEX Language L?

Match Test

- Is a word w in a REGEX Language L?
- Complexity?

Match Test

- Is a word w in a REGEX Language L?
- Complexity?
- $\Rightarrow$ back references are crucial.

Match Test

- Is a word w in a REGEX Language L?
- Complexity?
- $\Rightarrow$ back references are crucial.
- Our model: pattern languages.

Pattern languages 1/3

- Σ: Finite alphabet of terminal symbols.
  (e. g. $\Sigma = \{a, b, c, d\}$)
- X: Infinite alphabet of variables.
  ($X = \{x_1, x_2, x_3, \ldots\}$)
- A string $\alpha \in (\Sigma \cup X)^+$ is called a *pattern*.

Pattern languages 2/3

- *Morphism*: Mapping $\sigma : \Gamma_1^* \to \Gamma_2^*$ with $\sigma(x \cdot y) = \sigma(x) \cdot \sigma(y)$.
- *Substitution*: Morphism $\sigma : (\Sigma \cup X)^* \to \Sigma^*$ with $\sigma(a) = a$ for all $a \in \Sigma$.
- $L_\Sigma(\alpha)$: $\{\sigma(\alpha) \mid \sigma$ is a substitution$\}$.

Pattern languages 3/3

Example

- Pattern: $\alpha = x_1 \cdot x_2 \cdot x_1 \cdot x_2 \cdot x_1$.

Pattern languages 3/3

Example

- Pattern: $\alpha = x_1 \cdot x_2 \cdot x_1 \cdot x_2 \cdot x_1$.
- $L_\Sigma(\alpha) = \{w \mid w = u \cdot v \cdot u \cdot v \cdot u \text{ where } u, v \in \Sigma^*\}$.

Pattern languages 3/3

Example

- Pattern: $\alpha = x_1 \cdot x_2 \cdot x_1 \cdot x_2 \cdot x_1$.
- $L_\Sigma(\alpha) = \{w \mid w = u \cdot v \cdot u \cdot v \cdot u \text{ where } u, v \in \Sigma^*\}$.
- Example word: *acabcbaacabcbaac*.

Pattern languages 3/3

Example

- Pattern: $\alpha = x_1 \cdot x_2 \cdot x_1 \cdot x_2 \cdot x_1$.
- $L_\Sigma(\alpha) = \{w \mid w = u \cdot v \cdot u \cdot v \cdot u$ where $u, v \in \Sigma^*\}$.
- Example word: $ac\,abcba\,ac\,abcba\,ac$.

Pattern languages 3/3

Example

- Pattern: $\alpha = x_1 \cdot x_2 \cdot x_1 \cdot x_2 \cdot x_1$.
- $L_\Sigma(\alpha) = \{w \mid w = u \cdot v \cdot u \cdot v \cdot u \text{ where } u, v \in \Sigma^*\}$.
- Example word: *ac abcba ac abcba ac*.

Pattern languages 3/3

Example

- Pattern: $\alpha = x_1 \cdot x_2 \cdot x_1 \cdot x_2 \cdot x_1$.
- $L_\Sigma(\alpha) = \{w \mid w = u \cdot v \cdot u \cdot v \cdot u$ where $u, v \in \Sigma^*\}$.
- Example word: *acabcbaacabcbaac*.

We will concentrate on patterns in X^+.

Notation

- $\mathrm{var}(\alpha)$: Set of variables occurring in α.
  E. g. $\mathrm{var}(x_1 ab x_2 ba x_1 x_2 c x_3) = \{x_1, x_2, x_3\}$.

- $|\alpha|_{x_i}$: Number of occurrences of variable x_i in α.

Match test for pattern languages

- Example pattern $\alpha = x_1 \cdot x_2 \cdot x_3 \cdot x_2 \cdot x_3 \cdot x_1 \cdot x_4 \cdot x_3 \cdot x_5 \cdot x_5 \cdot x_4$,
- Question: $w \in L_{\Sigma}(\alpha)$, $w \in \Sigma^*$?

Match test for pattern languages

- Example pattern $\alpha = x_1 \cdot x_2 \cdot x_3 \cdot x_2 \cdot x_3 \cdot x_1 \cdot x_4 \cdot x_3 \cdot x_5 \cdot x_5 \cdot x_4$,
- Question: $w \in L_\Sigma(\alpha)$, $w \in \Sigma^*$?

w :

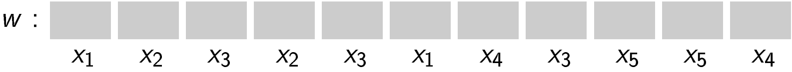
Match test for pattern languages

- Example pattern $\alpha = x_1 \cdot x_2 \cdot x_3 \cdot x_2 \cdot x_3 \cdot x_1 \cdot x_4 \cdot x_3 \cdot x_5 \cdot x_5 \cdot x_4$,
- Question: $w \in L_\Sigma(\alpha)$, $w \in \Sigma^*$?

w :

Match test for pattern languages

- Example pattern $\alpha = x_1 \cdot x_2 \cdot x_3 \cdot x_2 \cdot x_3 \cdot x_1 \cdot x_4 \cdot x_3 \cdot x_5 \cdot x_5 \cdot x_4$,
- Question: $w \in L_\Sigma(\alpha)$, $w \in \Sigma^*$?



Match test for pattern languages

- Example pattern $\alpha = x_1 \cdot x_2 \cdot x_3 \cdot x_2 \cdot x_3 \cdot x_1 \cdot x_4 \cdot x_3 \cdot x_5 \cdot x_5 \cdot x_4$,
- Question: $w \in L_\Sigma(\alpha)$, $w \in \Sigma^*$?

Match test for pattern languages

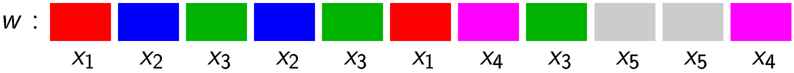
- Example pattern $\alpha = x_1 \cdot x_2 \cdot x_3 \cdot x_2 \cdot x_3 \cdot x_1 \cdot x_4 \cdot x_3 \cdot x_5 \cdot x_5 \cdot x_4$,
- Question: $w \in L_\Sigma(\alpha)$, $w \in \Sigma^*$?

Match test for pattern languages

- Example pattern $\alpha = x_1 \cdot x_2 \cdot x_3 \cdot x_2 \cdot x_3 \cdot x_1 \cdot x_4 \cdot x_3 \cdot x_5 \cdot x_5 \cdot x_4$,
- Question: $w \in L_\Sigma(\alpha)$, $w \in \Sigma^*$?

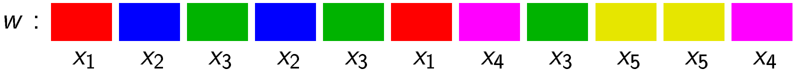
Match test for pattern languages

- Example pattern $\alpha = x_1 \cdot x_2 \cdot x_3 \cdot x_2 \cdot x_3 \cdot x_1 \cdot x_4 \cdot x_3 \cdot x_5 \cdot x_5 \cdot x_4$,
- Question: $w \in L_\Sigma(\alpha)$, $w \in \Sigma^*$?



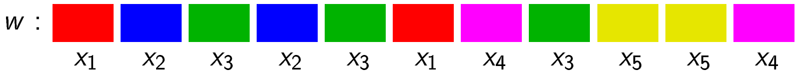
Match test for pattern languages

- Example pattern $\alpha = x_1 \cdot x_2 \cdot x_3 \cdot x_2 \cdot x_3 \cdot x_1 \cdot x_4 \cdot x_3 \cdot x_5 \cdot x_5 \cdot x_4$,
- Question: $w \in L_\Sigma(\alpha)$, $w \in \Sigma^*$?



Match test for pattern languages

- Example pattern $\alpha = x_1 \cdot x_2 \cdot x_3 \cdot x_2 \cdot x_3 \cdot x_1 \cdot x_4 \cdot x_3 \cdot x_5 \cdot x_5 \cdot x_4$,
- Question: $w \in L_\Sigma(\alpha)$, $w \in \Sigma^*$?



Number of factorisations: $O(|w|^{|\mathrm{var}(\alpha)|})$.

Match test for pattern languages

- Example pattern $\alpha = x_1 \cdot x_2 \cdot x_3 \cdot x_2 \cdot x_3 \cdot x_1 \cdot x_4 \cdot x_3 \cdot x_5 \cdot x_5 \cdot x_4$,
- Question: $w \in L_\Sigma(\alpha)$, $w \in \Sigma^*$?

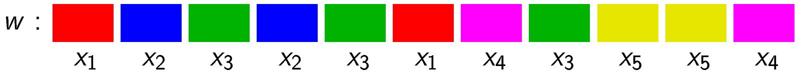
Number of factorisations: $O(|w|^{|\mathsf{var}(\alpha)|})$.

Theorem

The match test for pattern languages is NP-complete (Angluin, 80).

Match test for pattern languages

- Example pattern $\alpha = x_1 \cdot x_2 \cdot x_3 \cdot x_2 \cdot x_3 \cdot x_1 \cdot x_4 \cdot x_3 \cdot x_5 \cdot x_5 \cdot x_4$,
- Question: $w \in L_\Sigma(\alpha)$, $w \in \Sigma^*$?



Number of factorisations: $O(|w|^{|\mathrm{var}(\alpha)|})$.

Theorem

The match test for pattern languages is NP-complete (Angluin, 80).

Question: Are there classes of pattern languages with a polynomial match test?

Simple restrictions

Match test is polynomial if input (α, w) is restricted to

1. at most k different variables (for a constant k),
2. only one occurrence per variable,
3. $|\Sigma| = 1$.

A more useful restriction

We aim for a restriction with

- arbitrarily many variables,
- arbitrarily many occurrences of each variable,
- unrestricted cardinality of Σ,
- polynomial match test.

A more useful restriction

We aim for a restriction with

- arbitrarily many variables,
- arbitrarily many occurrences of each variable,
- unrestricted cardinality of Σ,
- polynomial match test.

Main idea: Establish a reasonable automata model to recognize pattern languages.

The Janus automaton

- Two two-way input heads.
- A constant number of k counters.
- A finite state control.

Counters

Each counter consists of

- a counter value,
- a counter bound.

Counters

Each counter consists of

- a counter value,
- a counter bound.

The counter value

- can be incremented or left unchanged,
- starts again at 1 if counter bound is reached or counter is reset.

Counters

Each counter consists of

- a counter value,
- a counter bound.

The counter value

- can be incremented or left unchanged,
- starts again at 1 if counter bound is reached or counter is reset.

The counter bound

- changes if a counter is reset.

Transitions

A transition, depending

- on the current state,
- on the currently scanned input symbols,
- on whether the counters have reached their counter bounds or not,

determines (completely deterministically)

- the next state,
- the input head movements,
- the counter instructions (increment, left unchanged, reset),

Transitions

A transition, depending

- on the current state,
- on the currently scanned input symbols,
- on whether the counters have reached their counter bounds or not,

determines (completely deterministically)

- the next state,
- the input head movements,
- the counter instructions (increment, left unchanged, reset),

RESET: A new counter bound is nondeterministically guessed between 0 and $|w|$. Counter value is set to 1.

Notation

- The automata model is denoted by $JFA(k)$,
- $L(M)$ is the language accepted by a $JFA(k)$ M,
- $\mathfrak{L}_k := \{L(M) \mid \text{ M is a } JFA(k)\}$.

Back to pattern languages

Proposition

Let α be a pattern with $|\mathrm{var}(\alpha)| = m$. Then $L_{\Sigma}(\alpha) \in \mathfrak{L}_m$.

Back to pattern languages

Proposition

Let α be a pattern with $|\text{var}(\alpha)| = m$. Then $L_\Sigma(\alpha) \in \mathfrak{L}_m$.

Example:

Is $w \in L_\Sigma(\alpha)$ where $\alpha = x_1 \cdot x_2 \cdot x_3 \cdot x_2 \cdot x_3 \cdot x_1 \cdot x_4 \cdot x_3 \cdot x_5 \cdot x_5 \cdot x_4$?

Back to pattern languages

Proposition

Let α be a pattern with $|\text{var}(\alpha)| = m$. Then $L_\Sigma(\alpha) \in \mathfrak{L}_m$.

Example:

Is $w \in L_\Sigma(\alpha)$ where $\alpha = x_1 \cdot x_2 \cdot x_3 \cdot x_2 \cdot x_3 \cdot x_1 \cdot x_4 \cdot x_3 \cdot x_5 \cdot x_5 \cdot x_4$?

Reset all counters and obtain counter bounds C_1, C_2, C_3, C_4, C_5.

Back to pattern languages

Proposition

Let α be a pattern with $|\text{var}(\alpha)| = m$. Then $L_\Sigma(\alpha) \in \mathfrak{L}_m$.

Example:

Is $w \in L_\Sigma(\alpha)$ where $\alpha = x_1 \cdot x_2 \cdot x_3 \cdot x_2 \cdot x_3 \cdot x_1 \cdot x_4 \cdot x_3 \cdot x_5 \cdot x_5 \cdot x_4$?

Reset all counters and obtain counter bounds C_1, C_2, C_3, C_4, C_5.

w :

Back to pattern languages

Proposition

Let α be a pattern with $|\mathrm{var}(\alpha)| = m$. Then $L_\Sigma(\alpha) \in \mathfrak{L}_m$.

Example:

Is $w \in L_\Sigma(\alpha)$ where $\alpha = x_1 \cdot x_2 \cdot x_3 \cdot x_2 \cdot x_3 \cdot x_1 \cdot x_4 \cdot x_3 \cdot x_5 \cdot x_5 \cdot x_4$?

Reset all counters and obtain counter bounds C_1, C_2, C_3, C_4, C_5.

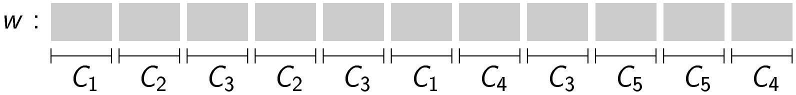
w :

Back to pattern languages

Proposition

Let α be a pattern with $|\mathrm{var}(\alpha)| = m$. Then $L_\Sigma(\alpha) \in \mathfrak{L}_m$.

Example:
Is $w \in L_\Sigma(\alpha)$ where $\alpha = x_1 \cdot x_2 \cdot x_3 \cdot x_2 \cdot x_3 \cdot x_1 \cdot x_4 \cdot x_3 \cdot x_5 \cdot x_5 \cdot x_4$?
Reset all counters and obtain counter bounds C_1, C_2, C_3, C_4, C_5.



Back to pattern languages

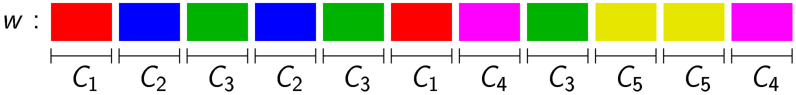
Proposition

Let α be a pattern with $|\text{var}(\alpha)| = m$. Then $L_\Sigma(\alpha) \in \mathfrak{L}_m$.

Example:

Is $w \in L_\Sigma(\alpha)$ where $\alpha = x_1 \cdot x_2 \cdot x_3 \cdot x_2 \cdot x_3 \cdot x_1 \cdot x_4 \cdot x_3 \cdot x_5 \cdot x_5 \cdot x_4$?

Reset all counters and obtain counter bounds C_1, C_2, C_3, C_4, C_5.



Back to pattern languages

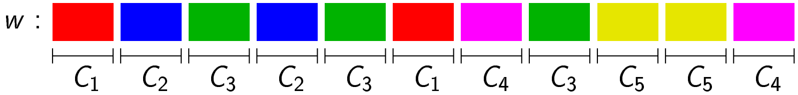
Proposition

Let α be a pattern with $|\mathrm{var}(\alpha)| = m$. Then $L_\Sigma(\alpha) \in \mathfrak{L}_m$.

Example:

Is $w \in L_\Sigma(\alpha)$ where $\alpha = x_1 \cdot x_2 \cdot x_3 \cdot x_2 \cdot x_3 \cdot x_1 \cdot x_4 \cdot x_3 \cdot x_5 \cdot x_5 \cdot x_4$?

Reset all counters and obtain counter bounds C_1, C_2, C_3, C_4, C_5.



Question

Is it possible to recognise $L_\Sigma(\alpha)$ by a $JFA(k)$ with $k < |\mathrm{var}(\alpha)|$?

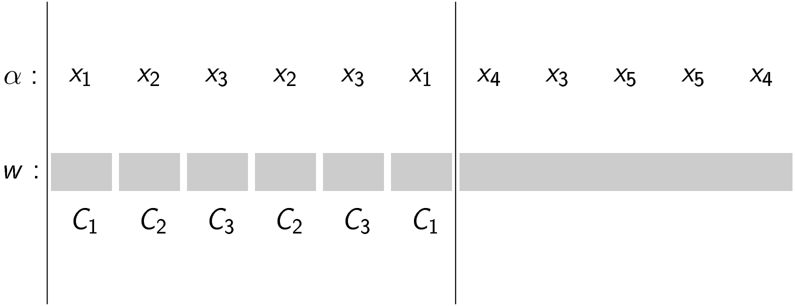
Example

α : x_1 x_2 x_3 x_2 x_3 x_1 x_4 x_3 x_5 x_5 x_4

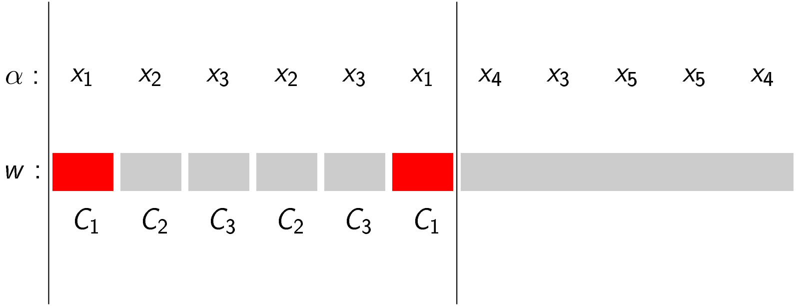
w :

Example

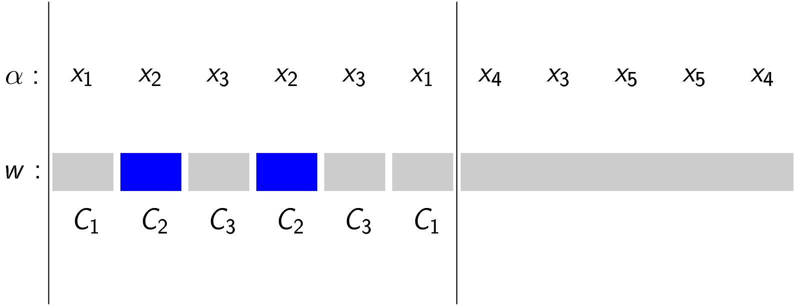
Example



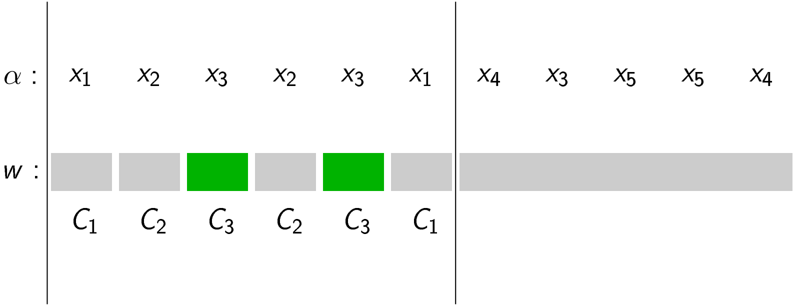
Example



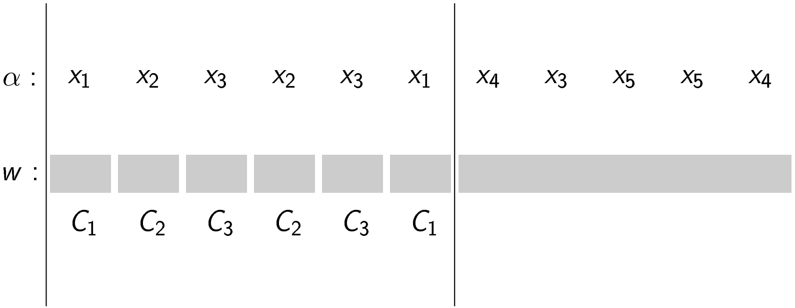
Example



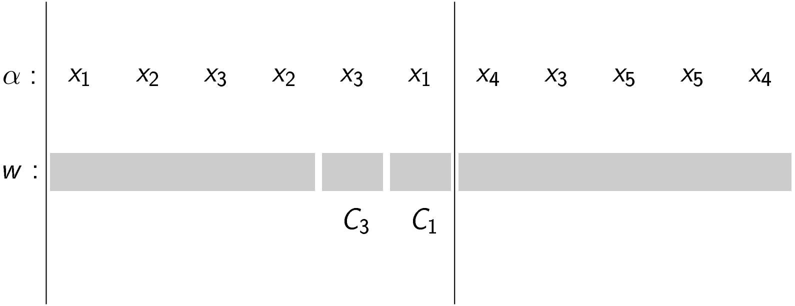
Example



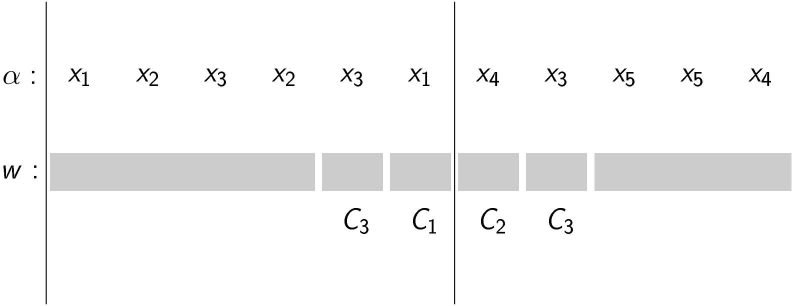
Example



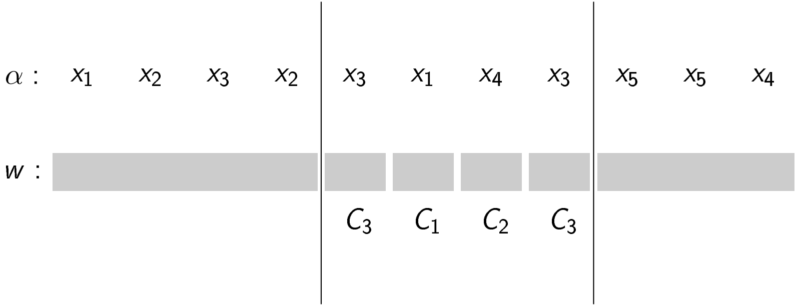
Example



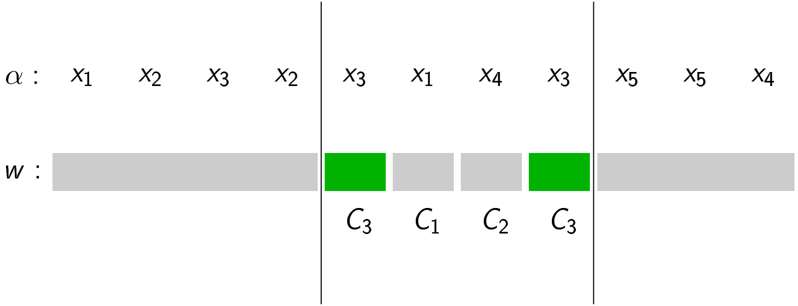
Example



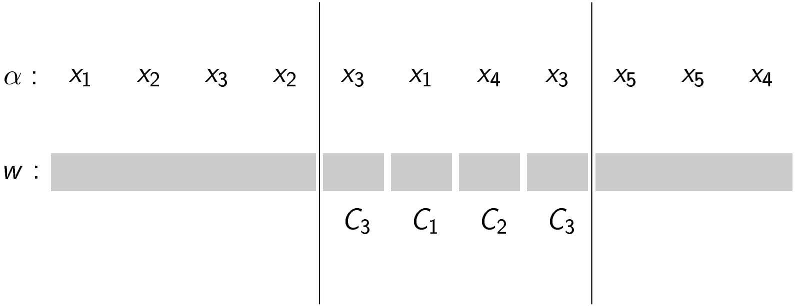
Example



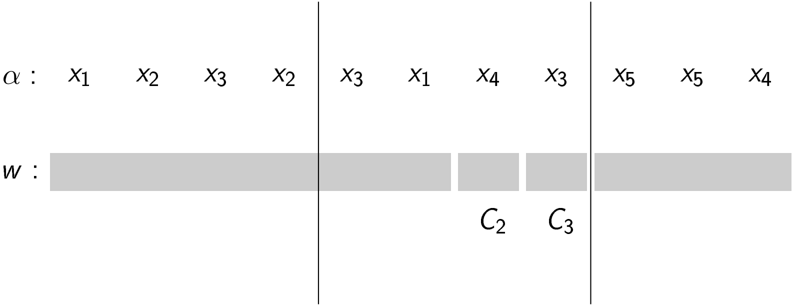
Example



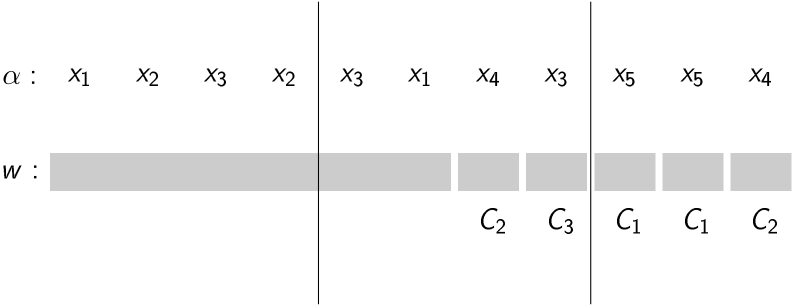
Example



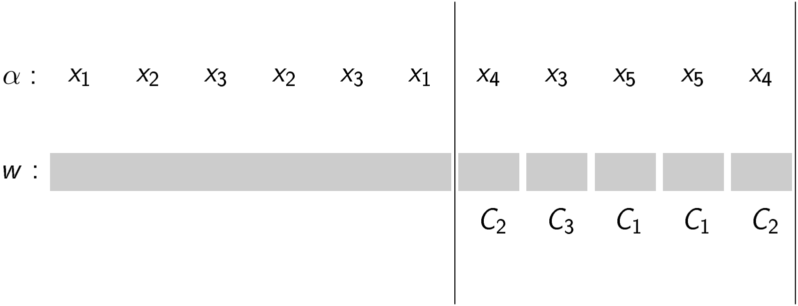
Example



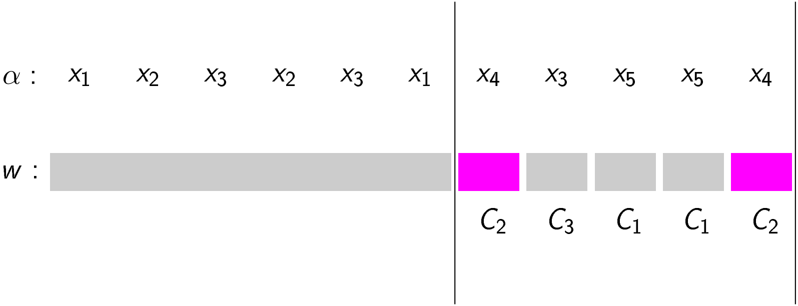
Example



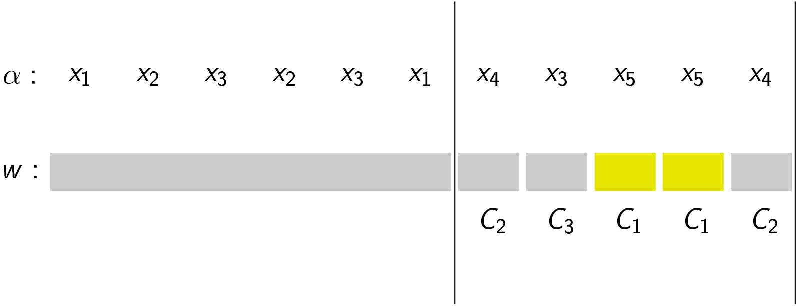
Example



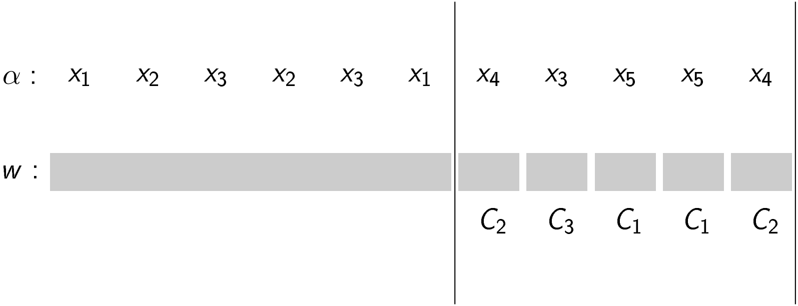
Example



Example



Example



Variable distance

The pattern α has a *variable distance of k* ($\text{vd}(\alpha) = k$) if and only if k is the smallest number such that, for each $x \in \text{var}(\alpha)$,

Variable distance

The pattern α has a *variable distance of k* ($\mathrm{vd}(\alpha) = k$) if and only if k is the smallest number such that, for each $x \in \mathrm{var}(\alpha)$,

- $\alpha = \beta \cdot x \cdot \gamma \cdot x \cdot \delta$,
- $|\gamma|_x = 0$

Variable distance

The pattern α has a *variable distance of k* ($\text{vd}(\alpha) = k$) if and only if k is the smallest number such that, for each $x \in \text{var}(\alpha)$,

- $\alpha = \beta \cdot x \cdot \gamma \cdot x \cdot \delta$,
- $|\gamma|_x = 0$

implies $|\text{var}(\gamma)| \leq k$.

Example

$$\mathrm{vd}(x_1 \cdot x_2 \cdot x_3 \cdot x_2 \cdot x_3 \cdot x_1 \cdot x_4 \cdot x_3 \cdot x_5 \cdot x_5 \cdot x_4) = ?$$

Example

$$\text{vd}(x_1 \cdot x_2 \cdot x_3 \cdot x_2 \cdot x_3 \cdot x_1 \cdot x_4 \cdot x_3 \cdot x_5 \cdot x_5 \cdot x_4) = ?$$

x_1 x_2 x_3 x_2 x_3 x_1 x_4 x_3 x_5 x_5 x_4

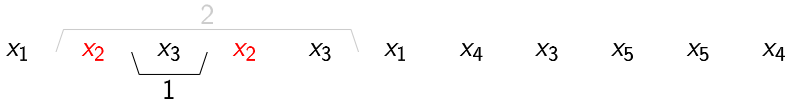
Example

$$\mathrm{vd}(x_1 \cdot x_2 \cdot x_3 \cdot x_2 \cdot x_3 \cdot x_1 \cdot x_4 \cdot x_3 \cdot x_5 \cdot x_5 \cdot x_4) = ?$$

$$\textcolor{red}{x_1} \; \overbrace{x_2 \;\; x_3 \;\; x_2 \;\; x_3}^{2} \; \textcolor{red}{x_1} \;\; x_4 \;\; x_3 \;\; x_5 \;\; x_5 \;\; x_4$$

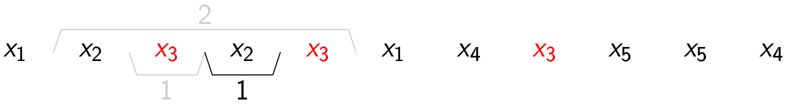
Example

$$\mathrm{vd}(x_1 \cdot x_2 \cdot x_3 \cdot x_2 \cdot x_3 \cdot x_1 \cdot x_4 \cdot x_3 \cdot x_5 \cdot x_5 \cdot x_4) = ?$$



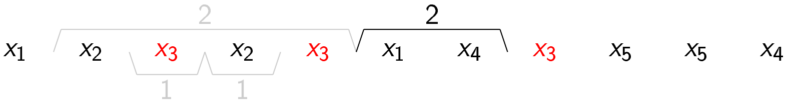
Example

$$\mathrm{vd}(x_1 \cdot x_2 \cdot x_3 \cdot x_2 \cdot x_3 \cdot x_1 \cdot x_4 \cdot x_3 \cdot x_5 \cdot x_5 \cdot x_4) = ?$$



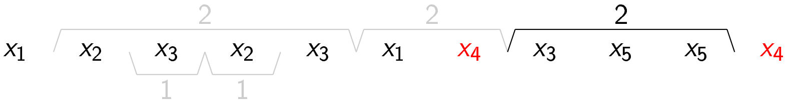
Example

$$\mathrm{vd}(x_1 \cdot x_2 \cdot x_3 \cdot x_2 \cdot x_3 \cdot x_1 \cdot x_4 \cdot x_3 \cdot x_5 \cdot x_5 \cdot x_4) = ?$$



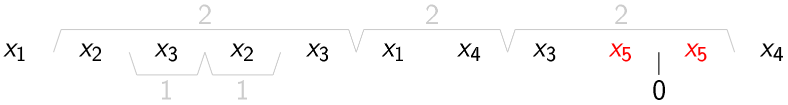
Example

$$\mathrm{vd}(x_1 \cdot x_2 \cdot x_3 \cdot x_2 \cdot x_3 \cdot x_1 \cdot x_4 \cdot x_3 \cdot x_5 \cdot x_5 \cdot x_4) = ?$$



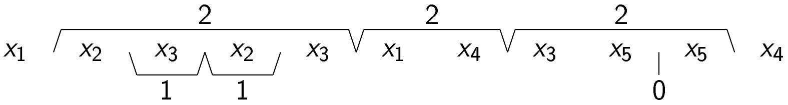
Example

$$\mathrm{vd}(x_1 \cdot x_2 \cdot x_3 \cdot x_2 \cdot x_3 \cdot x_1 \cdot x_4 \cdot x_3 \cdot x_5 \cdot x_5 \cdot x_4) = ?$$



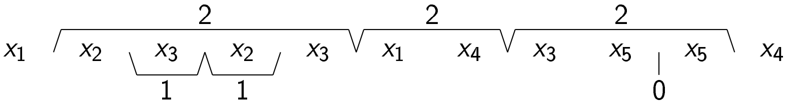
Example

$$\mathrm{vd}(x_1 \cdot x_2 \cdot x_3 \cdot x_2 \cdot x_3 \cdot x_1 \cdot x_4 \cdot x_3 \cdot x_5 \cdot x_5 \cdot x_4) = ?$$



Example

$$\text{vd}(x_1 \cdot x_2 \cdot x_3 \cdot x_2 \cdot x_3 \cdot x_1 \cdot x_4 \cdot x_3 \cdot x_5 \cdot x_5 \cdot x_4) = 2$$



Main Result

Theorem

Let α be a pattern with $\mathsf{vd}(\alpha) \leq k$. Then $L_\Sigma(\alpha) \in \mathfrak{L}_{k+1}$.

Main Result

Theorem

Let α be a pattern with $\mathsf{vd}(\alpha) \leq k$. Then $L_\Sigma(\alpha) \in \mathfrak{L}_{k+1}$.

Corollary

The match test for the class $\{L_\Sigma(\alpha) \mid \mathsf{vd}(\alpha) \leq k\}$ is solvable in time $O(|\alpha|^3 \, |w|^{(\mathsf{vd}(\alpha)+4)})$.

Test Results 1/3

Two algorithms:

Test Results 1/3

Two algorithms:

- JANUS: Java implementation of our JFA approach for the match test of pattern languages.

Test Results 1/3

Two algorithms:

- JANUS: Java implementation of our JFA approach for the match test of pattern languages.
- JREG: Java regex engine, implemented in the Java core packages.

Test Results 2/3

Test settings:

- $|\mathsf{var}(\alpha)| \in \{10, 15\}$,

Test Results 2/3

Test settings:

- $|\text{var}(\alpha)| \in \{10, 15\}$,
- $\text{vd}(\alpha) \in \{2, 3, 4\}$,

Test Results 2/3

Test settings:

- $|\text{var}(\alpha)| \in \{10, 15\}$,
- $\text{vd}(\alpha) \in \{2, 3, 4\}$,
- For each $x \in \text{var}(\alpha)$, $20 \leq |\alpha|_x \leq 25$,

Test Results 2/3

Test settings:

- $|\mathrm{var}(\alpha)| \in \{10, 15\}$,
- $\mathrm{vd}(\alpha) \in \{2, 3, 4\}$,
- For each $x \in \mathrm{var}(\alpha)$, $20 \leq |\alpha|_x \leq 25$,
- $w \in \Sigma^+$ with $|\Sigma| = 2$,

Test Results 2/3

Test settings:

- $|\mathrm{var}(\alpha)| \in \{10, 15\}$,
- $\mathrm{vd}(\alpha) \in \{2, 3, 4\}$,
- For each $x \in \mathrm{var}(\alpha)$, $20 \leq |\alpha|_x \leq 25$,
- $w \in \Sigma^+$ with $|\Sigma| = 2$,
- on average $|w| \approx 3\,|\alpha|$,

Test Results 2/3

Test settings:

- $|\text{var}(\alpha)| \in \{10, 15\}$,
- $\text{vd}(\alpha) \in \{2, 3, 4\}$,
- For each $x \in \text{var}(\alpha)$, $20 \leq |\alpha|_x \leq 25$,
- $w \in \Sigma^+$ with $|\Sigma| = 2$,
- on average $|w| \approx 3\,|\alpha|$,
- For each scenario, 250 instances with $w \in L_\Sigma(\alpha)$ and 250 instances with $w \notin L_\Sigma(\alpha)$.

Test Results 3/3

$\lvert var(\alpha) \rvert$	vd(α)	JANUS	JREG	JREG / JANUS
10	2	114 (0)	2653 (0)	23.2
10	3	5272 (0)	234903 (142)	44.56
10	4	36951 (0)	429527 (313)	11.62
15	2	440 (0)	10818 (0)	24.61
15	3	23506 (5)	371430 (232)	15.8
15	4	264239 (119)	526005 (403)	1.99

Questions

Thank you for your attention.

Questions?