Indexing Finite-State Automata Using Forward-Stable Partitions

Ruben Becker, Sung-Hwan Kim,
Nicola Prezza, **Carlo Tosoni**

DAIS, Ca' Foscari University, Venice

Monday 23rd September, 2024

The Burrows-Wheeler transform

The **Burrows-Wheeler transform** (**BWT**) is a famous reversible string transformation introduced by Burrows and Wheeler in 1994 [1]. In the following years, the BWT obtained significant interest from researchers, due to the fact that it proved to be an excellent transformation to **compress** and **index** strings.

1. *Burrows M., Wheeler D.: A block-sorting lossless data compression algorithm. SRS Research Report (1994)*

Wheeler NFAs

In 2017, Gagie et al. introduced **Wheeler nondeterministic finite automata (NFAs)** [2]. Wheeler NFAs represents a **generalisation of the original BWT** to automata.

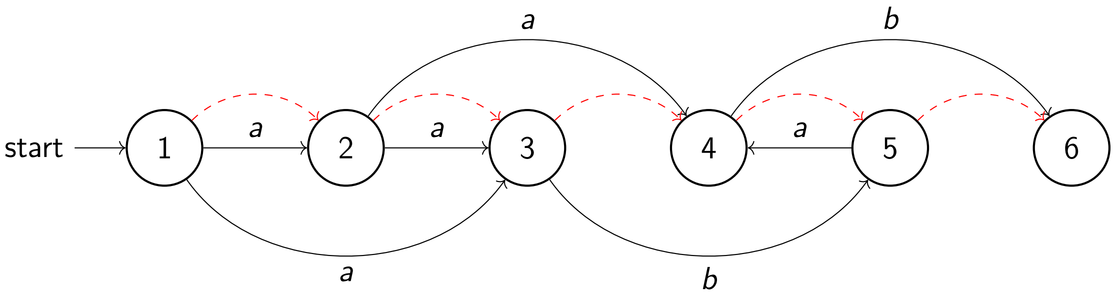
Similarly to the case of strings, **they can be compressed and indexed in almost optimal time and space**.

2. *Gagie et al.: Wheeler graphs: A framework for BWT-based data structures. Theor. Comput. Sci. (2017)*

Definition of Wheeler order

Let $\mathcal{A} = (Q, \delta, \Sigma, s)$ be an NFA. A total order $\leq$ of Q is a **Wheeler order** of $\mathcal{A}$ if for any pair $\mathbf{u} \in \delta(\mathbf{u}', \mathbf{a})$ and $\mathbf{v} \in \delta(\mathbf{v}', \mathbf{a}')$:

1. $\mathbf{a} < \mathbf{a}' \implies \mathbf{u} < \mathbf{v}$,
2. $(\mathbf{a} = \mathbf{a}') \wedge (\mathbf{u} < \mathbf{v}) \implies \mathbf{u}' \leq \mathbf{v}'$.



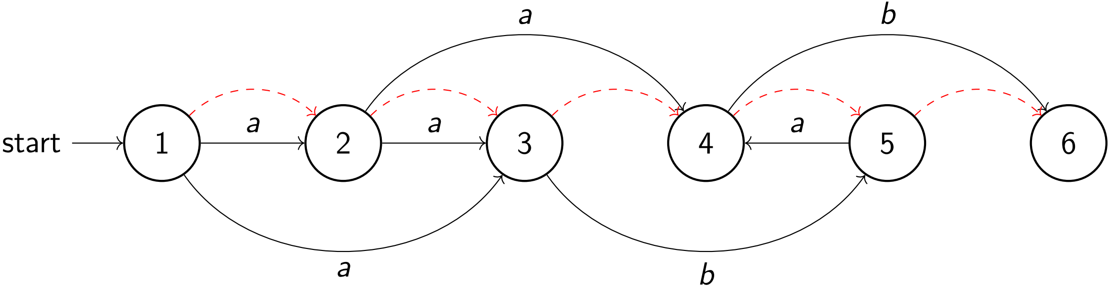
An NFA is **Wheeler** if it admits a **Wheeler order**. (NP-hard problem)

In other words...

A **Wheeler order** is a total order which sorts the states of an NFA based on the **strings reaching them**.

$$I_1 = \emptyset \qquad I_2 = \{a\} \qquad I_3 = \{a, aa\}$$

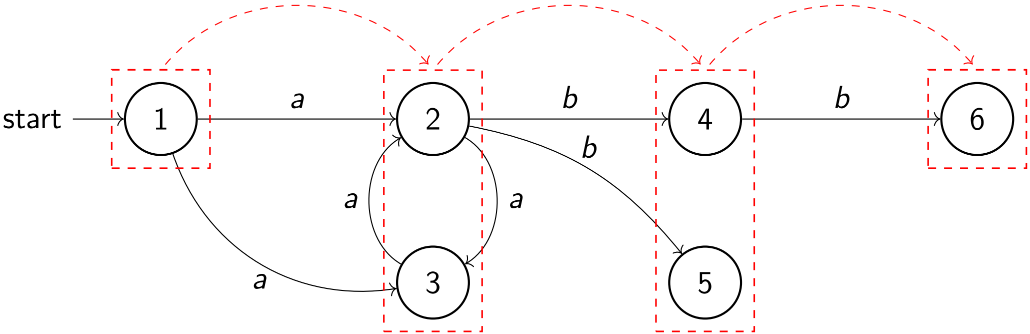
$$I_4 = \{aa, aba, aaba\} \qquad I_5 = \{ab, aab\} \qquad I_6 = \{aab, abab, aabab\}$$



We are comparing strings from **right to left**! (co-lexicographically) (NP-hard problem)

quasi-Wheeler NFAs

An NFA is **quasi-Wheeler** if it admits a **Wheeler preorder** [3]. The class of quasi-Wheeler NFAs is **strictly larger** than that of Wheeler NFAs



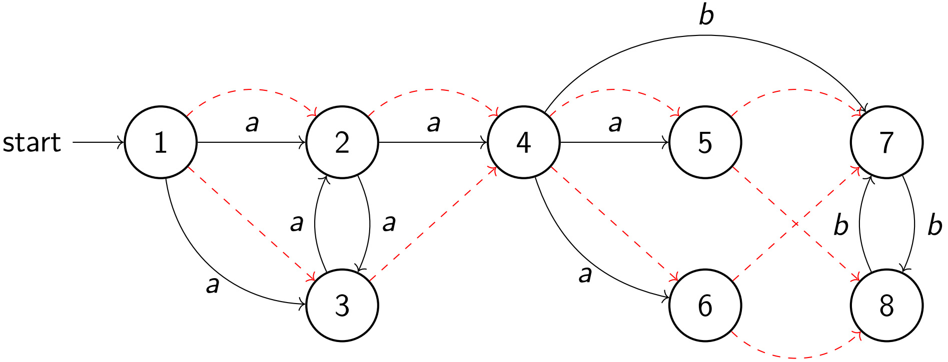
The NFA is not Wheleer, but it is quasi-Wheeler! (P problem)

3. *Alanko et al.: Wheeler languages. Information and Computation. (2021)*

Co-lex orders

Co-lex orders are a generalization of Wheeler orders to **arbitrary NFAs** [4].

Intuition: it's always possible to **partially sort** the states according to the strings reaching them

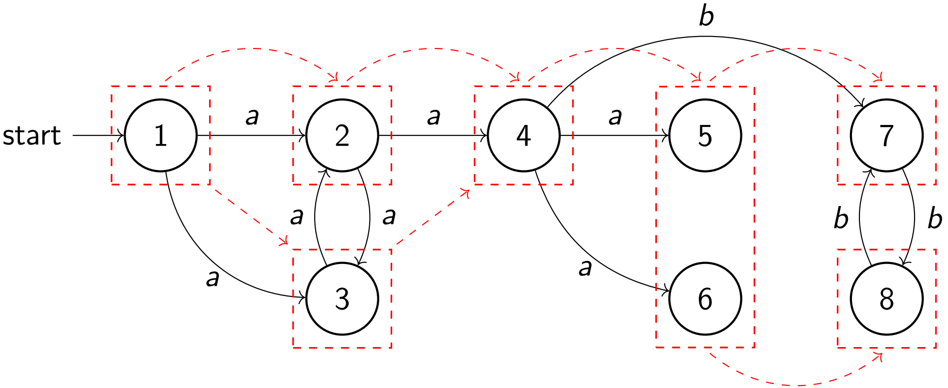


This NFA is neither Wheeler nor quasi-Wheeler. (NP-hard problem)

4. *Cotumaccio N., Prezza N.: On indexing and compressing finite automata. SODA. (2021)*

Maximum co-lex relation

The **maximum co-lex relation** is a partial order on **equivalence classes of states** [5]. For each NFA there exists an unique maximum co-lex relation.

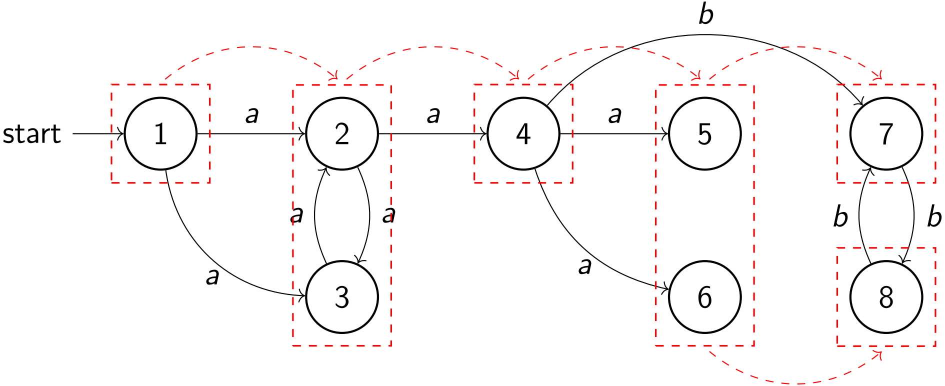


(P problem) States 5 and 6 can be merged into an unique state **without changing the language of the NFA**!

5. *Cotumaccio N.: Graphs can be succinctly indexed for pattern matching in* $O(|E|^2 + |V|^{5/2})$ *time. DCC. (2022)*

CFS order

The **coarsest forward-stable co-lex** (**CFS**) **order** is a partial order on equivalence classes that we introduced in this paper.



Compared to the maximum co-lex order, it manages to **merge more states** together!

Width of orders

- **the space/time efficency** of the index we generate **depends on the width** of the orders we have chosen [4].
- In fact, once we have chosen the order, **we can count the states reached by a pattern** of length **m** in:

$$O(m \cdot w^2 \cdot \log(w \cdot |\Sigma|))$$

4. *Cotumaccio N., Prezza N.: On indexing and compressing finite automata. SODA. (2021)*

Relation among orders

:D	**Existence**	**NP-hard**	**P**
Wheeler order		✓	
Wheeler preorder			✓
CFS order	✓		✓
Max co-lex rel.	✓		✓
Co-lex order*	✓	✓	

decreasing width

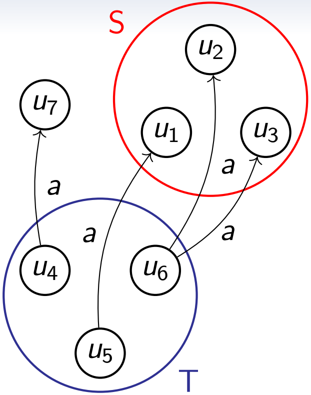
∗ It refers to co-lex orders of minimum width.

Forward-Stability

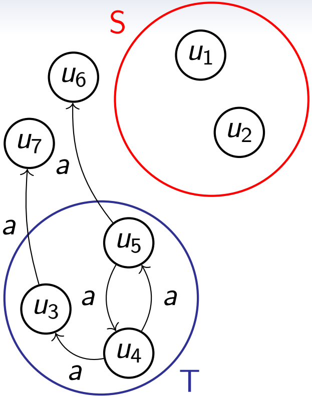
Before presenting our contribution we need to present the notion of **forward-stability**.

1. **Forward-stability between sets of states**.
2. **Forward-stable partitions**.

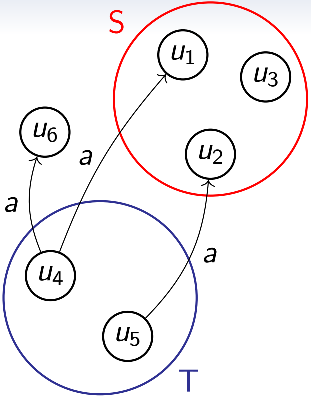
Forward-stability between set of states



All states of S reached by states of T $\Longrightarrow$
S is forward-stable wrt T!



All states of S not reached by states of T $\Longrightarrow$
S is forward-stable wrt T!



some states of S reached by T and some are not $\Longrightarrow$
They're not forward-stable!

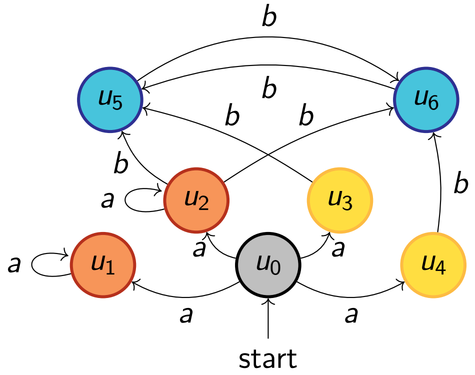
Coarsest forward-stable partition

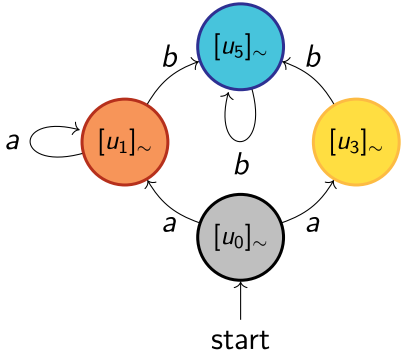
A partition $\mathcal{Q}$ is said to be **forward-stable** if for each parts $S, T \in \mathcal{Q}$ it holds that S **is forward stable wrt** T.

For each NFA there exists a **coarsest forward-stable partition** $Q/_{\sim_{FS}}$, that is, a forward-stable partition formed by the smallest number of parts.

We denote with $\mathcal{A}/_{\sim_{FS}}$ **the quotient automaton defined by the coarsest forward-stable partition**.

Example of a coarsest forward-stable partition





On the left, an automaton $\mathcal{A}$ whose coarsest forward-stable partition is $Q/_{\sim_{FS}} = \{\{u_0\}, \{u_1, u_2\}, \{u_3, u_4\}, \{u_5, u_6\}\}$. **On the right**, the quotient automaton $\mathcal{A}/_{\sim_{FS}}$.

Our contribution

The quotient automata $\mathcal{A}/_{\sim_{FS}}$ and $\mathcal{A}/_{\sim_R}$

$\mathcal{A}/_{\sim_R} = (Q/_{\sim_R}, \delta/_{\sim_R}, \Sigma, s/_{\sim_R})$ $\longrightarrow$ Quotient automaton defined by the **maximum co-lex relation** $\leq_R$ [5].

$\mathcal{A}/_{\sim_{FS}} = (Q/_{\sim_{FS}}, \delta/_{\sim_{FS}}, \Sigma, s/_{\sim_{FS}})$ $\longrightarrow$ Quotient automaton defined by the **coarsest forward-stable partition** [this work!].

Lemma 2

Consider the NFA $\mathcal{A}/_{\sim_R} = (Q/_{\sim_R}, \delta/_{\sim_R}, \Sigma, s/_{\sim_R})$. Then, $Q/_{\sim_R}$ is a **forward-stable partition** of $\mathcal{A}$.

5. *Cotumaccio N.: Graphs can be succinctly indexed for pattern matching in $O(|E|^2 + |V|^{5/2})$ time. DCC. (2022)*

The quotient automata $\mathcal{A}/_{\sim_{FS}}$ and $\mathcal{A}/_{\sim_R}$

$Q/_{\sim_R}$ is a forward-stable partition, however, $Q/_{\sim_{FS}}$ is the **coarsest** forward-stable partition, it follows that:

Theorem 1.3

Consider the partitions $Q/_{\sim_R}$ and $Q/_{\sim_{FS}}$, then the following statement holds:

$$|Q/_{\sim_{FS}}| \leq |Q/_{\sim_R}|$$

Maximum co-lex order of $\mathcal{A}/_{\sim_{FS}}$

A co-lex order of an NFA $\mathcal{A}$ is said to be the **maximum** co-lex order $\leq$ if it is equal to the **union of every co-lex order** of $\mathcal{A}$.

Lemma 3

Consider the NFA $\mathcal{A}/_{\sim_{FS}} = (Q/_{\sim_{FS}}, \delta/_{\sim_{FS}}, \Sigma, s/_{\sim_{FS}})$. Then $\mathcal{A}/_{\sim_{FS}}$ **admits a maximum co-lex order** $\leq$.

The CFS order of an NFA

Definition 10

Consider the NFA $\mathcal{A}/_{\sim_{FS}}$ and its maximum co-lex order $\leq$. Then the **CFS order** $\leq_{FS}$ is the partial order defined as follows:

$$\forall u, v \in Q, \; u \leq_{FS} v \iff [u]_{\sim_{FS}} \leq [v]_{\sim_{FS}}$$

Since the width of these orders is crucial for the time/space efficency of their corresponding indices, we have **compared the widths** of the **maximum co-lex relation** $\leq_R$ and of the **CFS order** $\leq$.

Relation between $\leq_{FS}$ and $\leq_R$

Lemma 6

Let $\leq_R$ and $\leq_{FS}$ be the **maximum co-lex relation** and the **CFS order** of an NFA $\mathcal{A}$, respectively. Then, $\leq_{FS}$ is a **superset** of $\leq_R$.

In other words, this means that;

$$\forall u, v \in Q : u \leq_R v \implies u \leq_{FS} v$$

Relation between $\leq_{FS}$ and $\leq_R$

This directly yields the following result;

Theorem 1.2

Let $\leq_R$ and $\leq_{FS}$ be the **maximum co-lex relation** and the **CFS order** of an NFA $\mathcal{A}$, respectively. Then;

The width of $\leq_{FS}$ *is* **smaller than or equal to** *the width of* $\leq_R$.

Relation between $\leq_{FS}$ and $\leq_R$

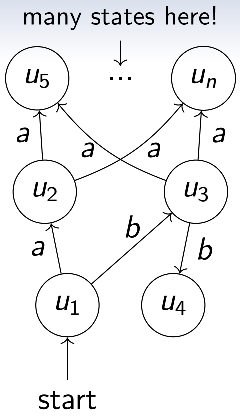
So we demonstrated that the width of $\leq_{FS}$ can be smaller than the width of $\leq_R$, **but how much smaller?**

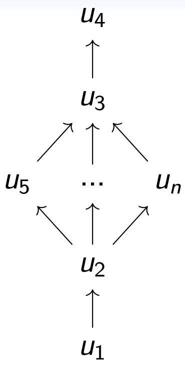
In the next theorem we demonstrated that in some cases it can be **asymptotically smaller**.

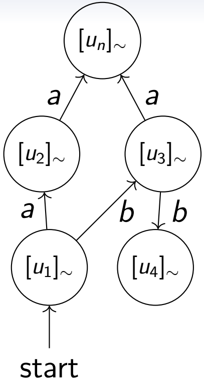
Theorem 2

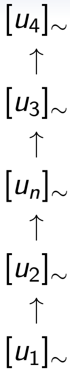
there exist NFAs $\mathcal{A}$ for which $\leq_R$ has **width $\boldsymbol{\Theta}(|\mathbf{Q}|)$** and $\leq_{FS}$ has **width 1**

Proof of Theorem 2









From left to right, **1.** an NFA $\mathcal{A}$ formed by n states, where $\mathcal{A} = \mathcal{A}/_{\sim_R}$. **2.** Hasse diagram of the maximum co-lex relation. **3.** The quotient automaton $\mathcal{A}/_{\sim_{FS}}$, formed by 5 states. **4.** The Hasse diagram of the CFS order.

Time complexity

Corollary 1

The CFS order of an NFA can be computed in $O(|\delta|^2)$ time.

In consideration of these facts, we can conclude that the **CFS order** beats the state-of-the-art competitor, i.e. the **maximum co-lex relation**, in every respect:

- Lower or same width.
- Lower or same number of states.
- Same time complexity.

Thank you for your attention ☺

Funded by ERC StG "REGINDEX: Compressed indexes for regular languages with applications to computational pan-genomics" grant nr 101039208. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Research Council Executive Agency. Neither the European Union nor the granting authority can be held responsible for them.