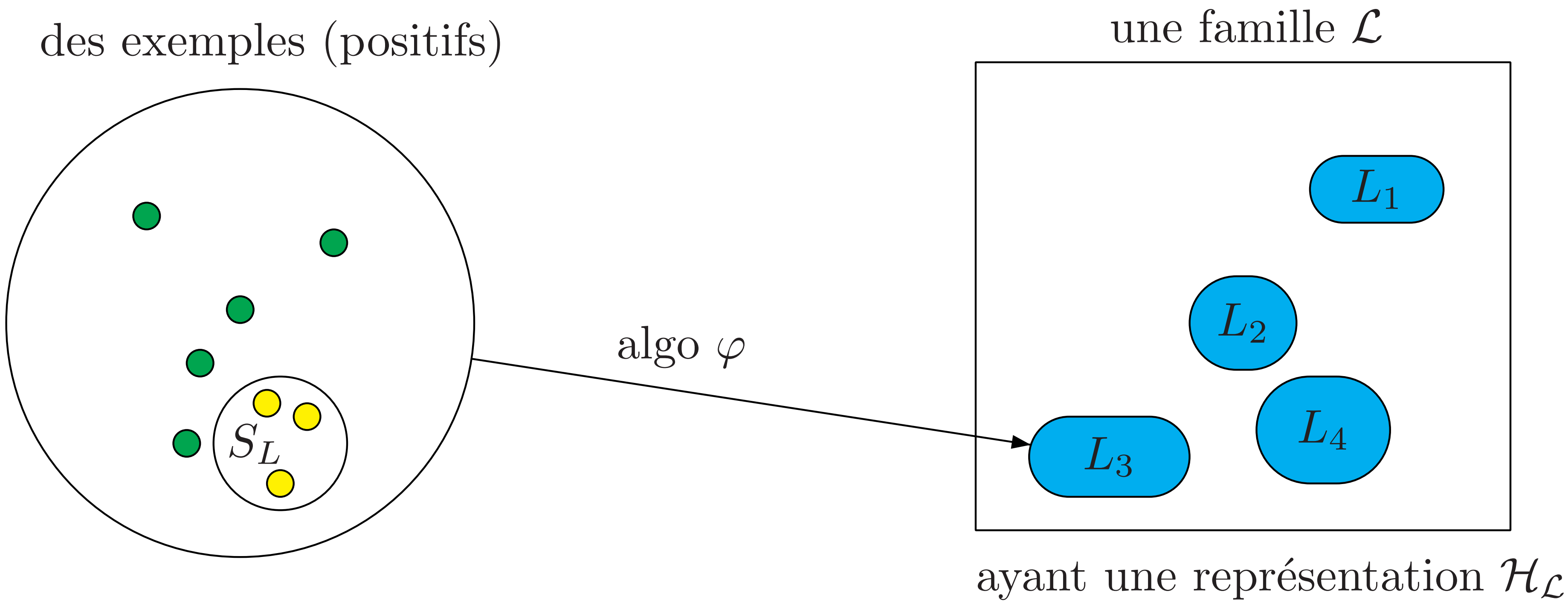
Identification des langages rationnels à résiduels k-disjoints

Alain Terlutte, Fabien Torre

Grappa, Mostrare, Lille

Identification à la limite avec échantillon caractéristique



$\mathcal{L}$ est *identifiable à la limite en temps et données polynômiaux*, à partir de présentations de type $\mathcal{S}_{\mathcal{L}}$, en utilisant la classe de représentations $\mathcal{H}_{\mathcal{L}}$, ssi il existe un algorithme φ tel que

- pour tout $L \in \mathcal{L}$, il existe un échantillon caractéristique S_L de taille polynomiale par rapport à H_L
- pour tout $S_L \subseteq S \in \mathcal{S}_{\mathcal{L}}$, on a $L_{\varphi(S)} = L$ et φ travaille en temps polynomial en $\|S\|$.

Les langages résiduels

Soit $L = \{w_1, w_2, ..., w_n, ...\}$ un langage.

Le *résiduel (droit)* de u dans L est défini par $u^{-1}L = \{v \in \Sigma^* \mid uv \in L\}$.

Exemple : $L = \{aa, ab, aab, aba, aabb, abab, ...\}$

$a^{-1}L = \{a, b, ab, ba, abb, bab, ...\}$ $\qquad L = \{aa, ab, aab, aba, aabb, abab, ...\}$

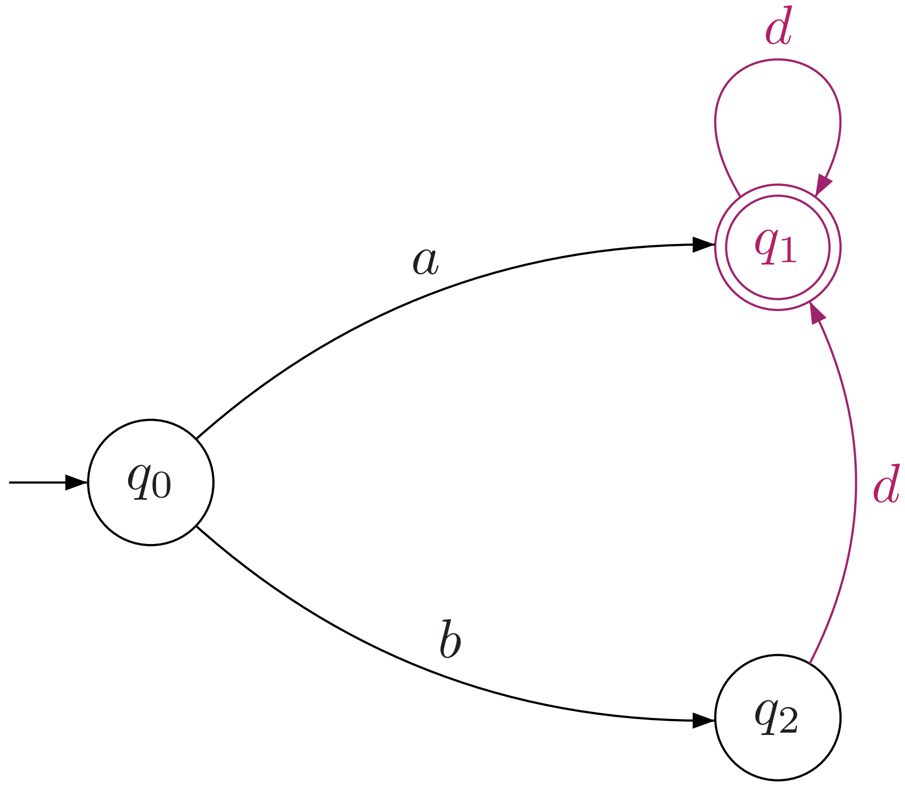
$(ab)^{-1}L = \{\varepsilon, a, ab, ...\}$ $\qquad L = \{aa, ab\varepsilon, aab, aba, aabb, abab, ...\}$

$\varepsilon^{-1}L = L$

$b^{-1}L = \emptyset$ s'il n'y a pas de mots qui commencent par b

- particulièrement utiles pour les langages rationnels.
- extensions aux langages linéaires, aux transductions, aux langages réguliers d'arbres.

AFD : Automate Fini Déterministe

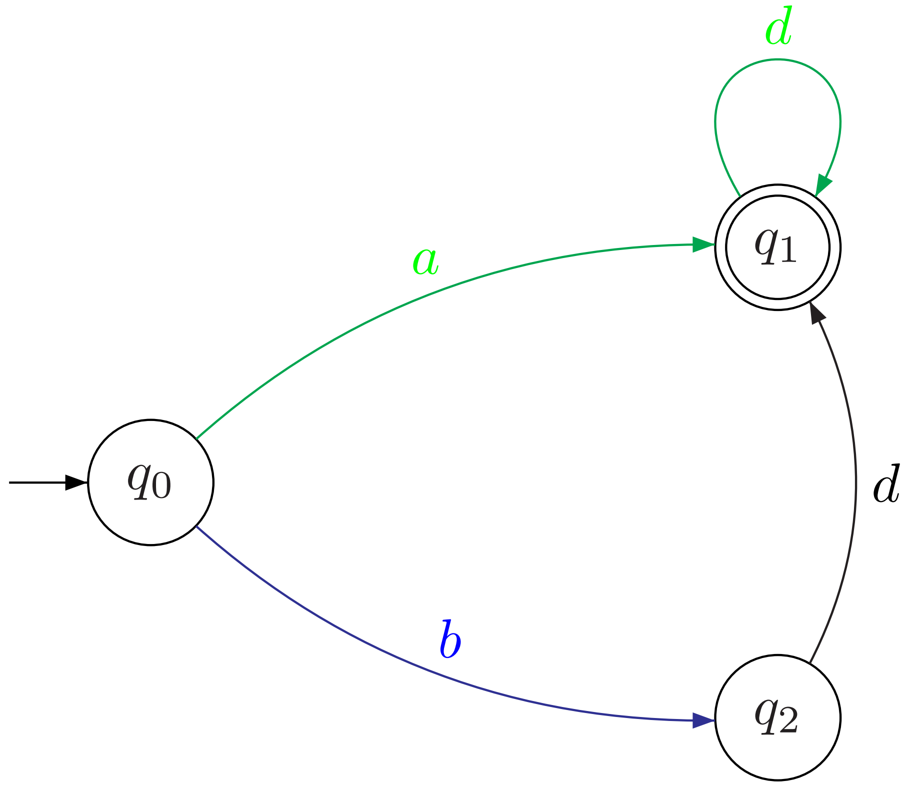


Un AFD reconnaissant $L = L_{q_0} = ad^* + bd^+$. C'est l'AFD minimal.

$L_{q_2} = d^+$

$L_{q_1} = d^*$

AFD et résiduels droits



Si l'automate est déterministe, un mot u atteint un seul état q et nous avons $u^{-1}L = L_q$.

$b^{-1}L = L_{q_2}$

$a^{-1}L = L_{q_1} = (ad)^{-1}L = (add)^{-1}L$

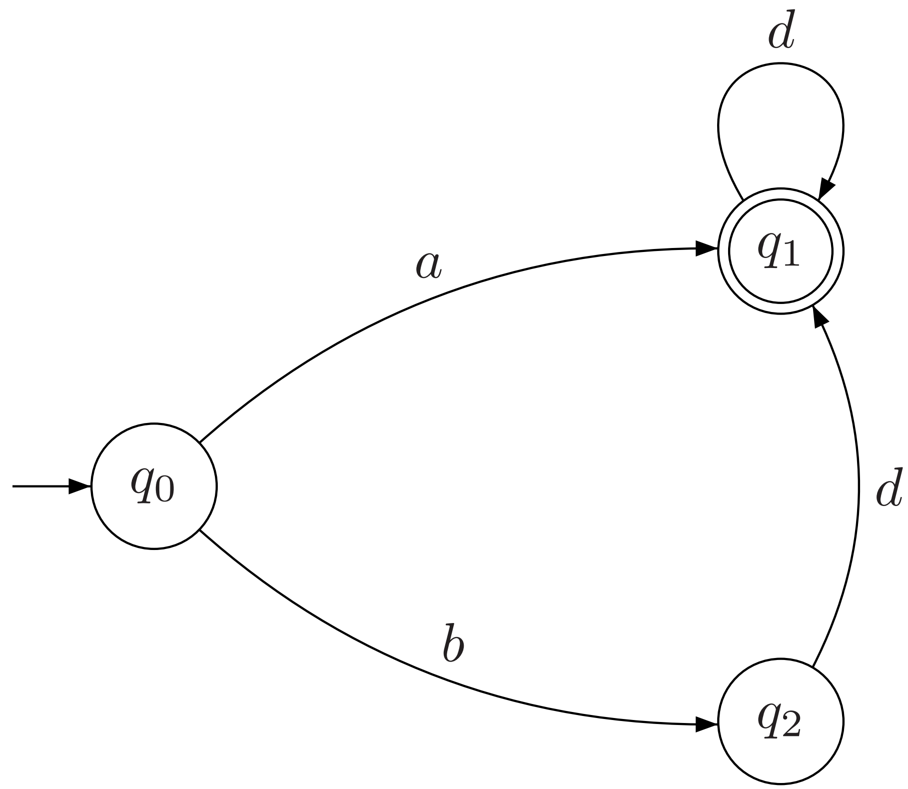
$\varepsilon^{-1}L = L = L_{q_0}$

$d^{-1}L = \emptyset$

AFD, résiduels, $F_{k,q}$

Chaque résiduel contient une partie finie qui peut le caractériser.

$F_{k,q} = \{v \in L_q \mid |v| < k\}$, les mots du résiduel L_q de longueur strictement inférieure à k

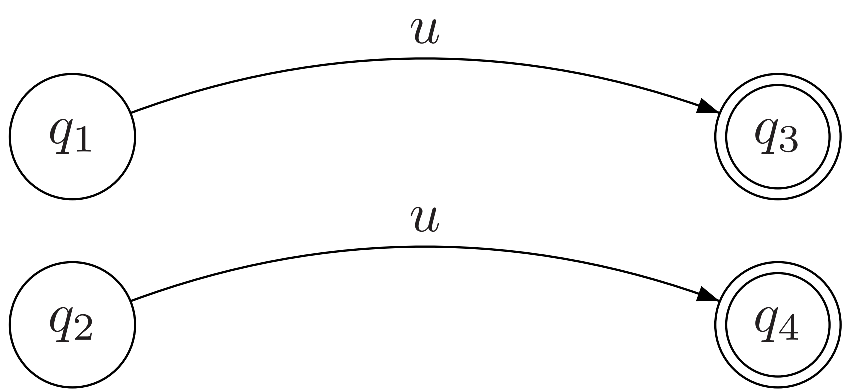


Pour chaque AFD, il existe une longueur k telle que les $F_{k,q}$ sont distincts.

$F_{1,q_0} = \emptyset$	$F_{1,q_1} = \{\varepsilon\}$	$F_{1,q_2} = \emptyset$
$F_{2,q_0} = \{a\}$	$F_{2,q_1} = \{\varepsilon, d\}$	$F_{2,q_2} = \{d\}$
$F_{3,q_0} = \{a, ad, bd\}$	$F_{3,q_1} = \{\varepsilon, d, dd\}$	$F_{3,q_2} = \{d, dd\}$

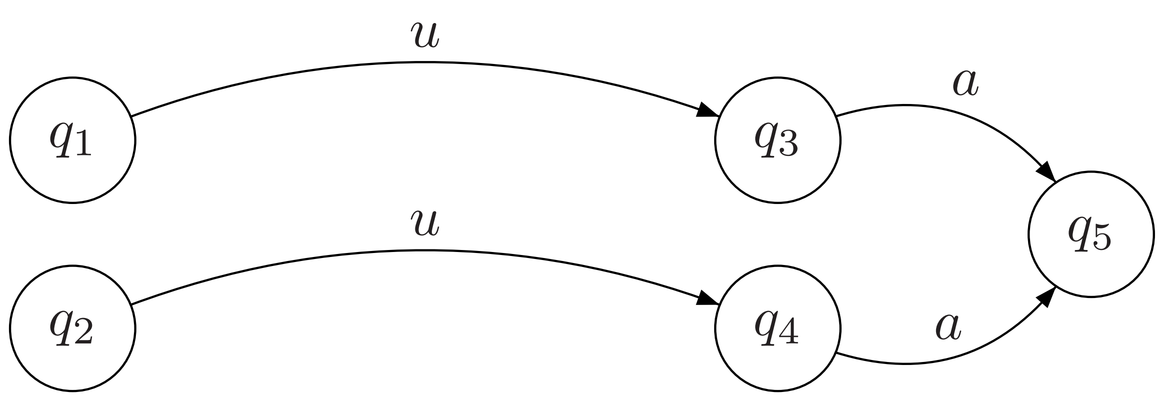
Automates k-réversibles

La condition de k-réversibilité



avec $|u| = k$

sont interdits



Cela signifie que $u^{-1}L_{q_1} \cap u^{-1}L_{q_2}$ doit être vide.

Ce qui donne

$$u^{-1}L_q \neq u^{-1}L_{q'} \quad \Longrightarrow \quad u^{-1}L_q \cap u^{-1}L_{q'} = \emptyset$$

Langages à résiduels k-disjoints

Nous allons utiliser les parties finies $F_{k,q}$ et la condition de k-réversibilité pour définir la hiérarchie des langages à résiduels k-disjoints.

La condition que doivent vérifier les états q et q' distincts est

$$L_q \neq L_{q'}$$

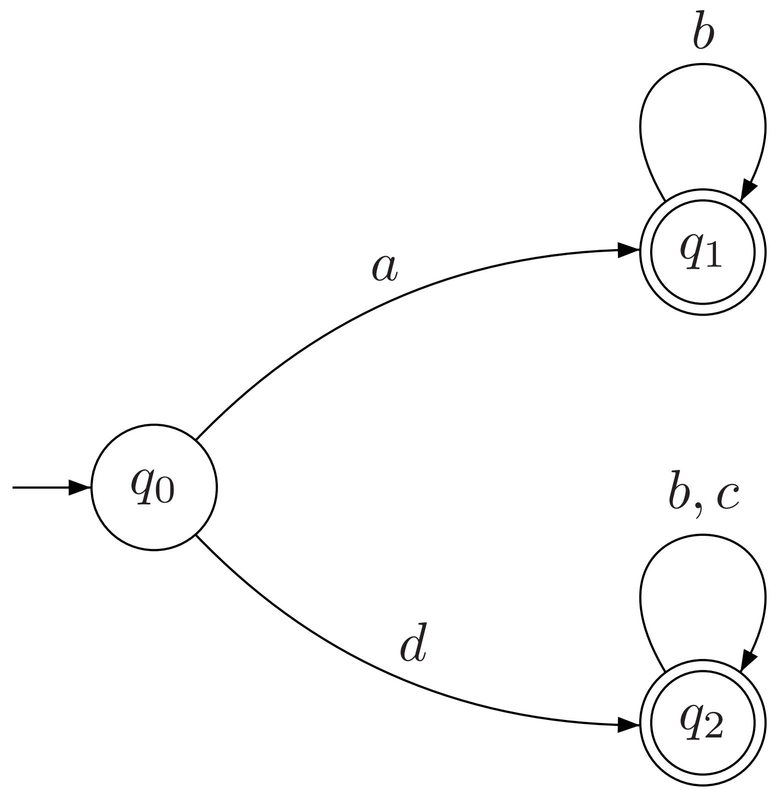
$$\Updownarrow$$

$$F_{k,q} \neq F_{k,q'}$$

$$\text{ou}$$

$$\exists u \in \Sigma^k, \left(u^{-1}L_q \neq \emptyset \text{ ou } u^{-1}L_{q'} \neq \emptyset\right) \text{ et } u^{-1}L_q \cap u^{-1}L_{q'} = \emptyset$$

Exemple



est 1-disjoint

$F_{1,q_0} = \emptyset$ $\quad$ $F_{1,q_1} = \{\varepsilon\}$ $\quad$ $F_{1,q_2} = \{\varepsilon\}$

Pour les états q_1 et q_2, les parties finies F_{1,q_1} et F_{1,q_2} sont identiques,
mais $c^{-1}L_{q_1}$ est vide alors que $c^{-1}L_{q_2}$ ne l'est pas ; $\quad$ $c^{-1}L_{q_1} \cap c^{-1}L_{q_2} = \emptyset$

Résultats

Chaque famille de langages à résiduels k-disjoints contient la famille de k-réversibles correspondante.

L'union de toutes les familles de langages à résiduels k-disjoints est égale à la famille des langages rationnels.

Identification des langages à résiduels k-disjoints

Soit $k \in \mathbb{N}$. Soit L un langage à résiduels k-disjoints.

On définit un échantillon caractéristique S_L pour L.

On adapte l'équivalence dans L en une équivalence dans S en supposant que $S_{q_\alpha} = \alpha^{-1}S$ est une estimation de $L_{q_\alpha} = \alpha^{-1}L$ et que S_{k,q_α} est une estimation de F_{k,q_α}.

$$\alpha \simeq_S \alpha'$$

$$\Updownarrow$$

$$S_{k,q_\alpha} = S_{k,q_{\alpha'}}$$

et

$$\forall u \in \Sigma^k, \left(u^{-1}S_{q_\alpha} = \emptyset \text{ et } u^{-1}S_{q_{\alpha'}} = \emptyset\right) \text{ ou } u^{-1}S_{q_\alpha} \cap u^{-1}S_{q_{\alpha'}} \neq \emptyset$$

Algorithme d'identification

```
construire l'ensemble Pref des préfixes de l'échantillon S
q0 = ε

tant que l'automate en construction n'est pas consistant
      et que Pref est non vide
  prendre le premier préfixe αx dans Pref
  si αx ≃_S α' avec α' ∈ Q
    alors
      ajouter une transition (α, x, α') à δ
    sinon
      ajouter q_αx à Q
      ajouter une transition (α, x, αx) à δ
  fin si

  si αx ∈ S alors ajouter q_αx à Q_F fin si
fin tant que

si A n'est pas k-disjoint ou pas consistant alors A = A_Σ* fin si
```

Résultat

Chaque famille de langages à résiduels k-disjoints peut être identifiée à la limite, en temps et données polynômiaux à partir d'exemples positifs, en les représentant par des AFD. L'algorithme travaille en temps $\mathcal{O}(|[S]|^4)$.

L'échantillon caractéristique a une taille en $\mathcal{O}(\|A_{Ldet}\|^3)$.

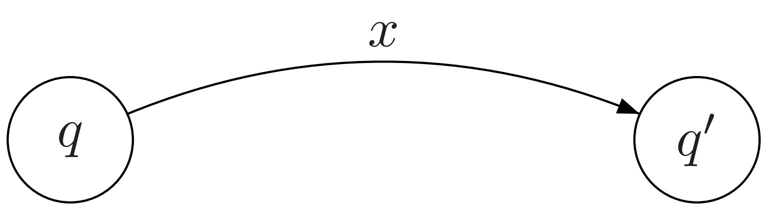
La hiérarchie de ces familles couvre tous les langages rationnels.

Détail de taille : les constantes du polynôme sont de l'ordre de $\|\Sigma\|^k$.

Conclusion

Nous avons obtenu un résultat théorique utilisant des sous-ensembles finis des résiduels.

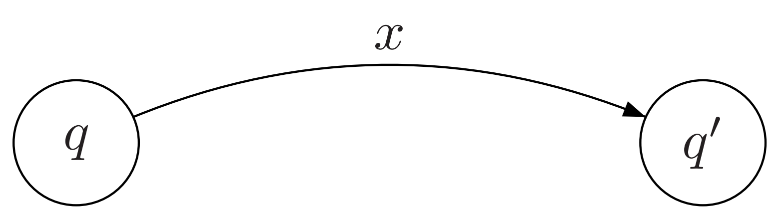
Il est possible de définir des algorithmes basés sur ces sous-ensembles $F_{k,q}$.



$L_q \cap x\Sigma^* = xL_{q'}$ dans les DFA

$L_q \supseteq xL_{q'}$ dans les AFER

approximé par



$F_{k,q} \cap x\Sigma^* = xF_{k,q'} \cap \Sigma^{\leq k}$

ou $F_{k,q} \supseteq xF_{k,q'} \cap \Sigma^{\leq k}$