



Algoritmuskélet 5. témakör

Pusztai Pál
pusztai@sze.hu

Tartalom

- A mintaillesztési probléma
 - Fogalmak és jelölések
 - Az egyszerű mintaillesztő
 - A Rabin-Karp illesztő
 - Mintaillesztés véges automatákkal
 - Átmeneti függvény számítás
 - Illesztés
 - A Knuth-Morris-Pratt illesztő
 - Prefix függvény számítás



Mintaillesztés

■ A mintaillesztési probléma

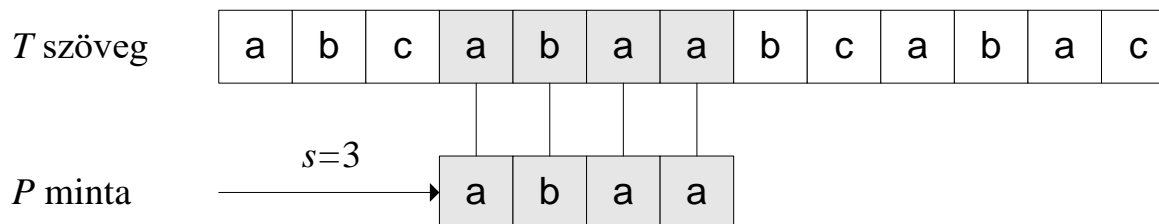
Tegyük fel, hogy a **szöveget**, amelyben keresünk a $T[1..n]$ tömbben, a **mintát** pedig, amit a szövegben keresünk a $P[1..m]$ tömbben tároljuk, ahol $m \leq n$ és mindkét tömb elemei a Σ **véges ábécé** jelei.

Lehetséges ábécék pl: $\Sigma = \{0, 1\}$, $\Sigma = \{a, b, \dots, z\}$.

A P minta **előfordul s eltolással** a T szövegben (más szavakkal a P minta a T szöveg $s+1$. **pozíciójára illeszkedik**), ha $0 \leq s \leq n-m$ és $T[s+1..s+m]=P[1..m]$ (azaz $T[s+j]=P[j]$, $1 \leq j \leq m$).

Ha P előfordul s eltolással T -ben, akkor s **érvényes eltolás**, ellenkező esetben s **érvénytelen eltolás**.

Feladat: Egy adott P minta összes érvényes eltolását megtalálni egy adott T szövegben.



A mintaillesztési probléma

Mintaillesztés

■ Fogalmak és jelölések

Jelölje a Σ ábécé jeleiből képzett összes véges hosszúságú sorozatok halmazát Σ^* .

A nulla hosszú, **üres sorozat** jele ε , Σ^* eleme.

Az x és y sorozatok **konkatenációja**, jele xy , egy olyan sorozat, amelyben x jeleit y jelei követik, és a hossza $|x| + |y|$.

A w sorozat az x sorozat **prefixe**, jele $w \sqsubset x$, ha van olyan $y \in \Sigma^*$ sorozat, hogy $x = wy$.

A w sorozat az x sorozat **szuffixe**, jele $w \sqsupset x$, ha van olyan $y \in \Sigma^*$ sorozat, hogy $x = yw$.

Pl: $ab \sqsubset abcca$, $cca \sqsupset abcca$.

Jelölje P_k a $P[1..m]$ minta k hosszúságú $P[1..k]$ prefixét. Ekkor $P_0 = \varepsilon$ és $P_m = P[1..m] = P$.

Jelölje T_k a $T[1..n]$ szöveg k hosszúságú $T[1..k]$ prefixét.

A **mintaillesztési probléma** (ezekkel a jelölésekkel): megtalálni az összes olyan s eltolási értéket a $0 \leq s \leq n-m$ tartományban, amelyre $P \sqsupset T_{s+m}$ teljesül.

Mintaillesztés

EGYSZERŰ-MINTAILLESZTŐ(T, P)

```

1   $n \leftarrow \text{hossz}[T]$ 
2   $m \leftarrow \text{hossz}[P]$ 
3  for  $s \leftarrow 0, n-m$ 
4      if  $P[1..m] = T[s+1..s+m]$ 
5          Ki: „A minta illeszkedik az”,  $s+1$ , „. pozícióra”
    
```

Hatékonyság: $O((n-m+1)m)$ (ez az illesztési idő, mivel előfeldolgozás nincs).

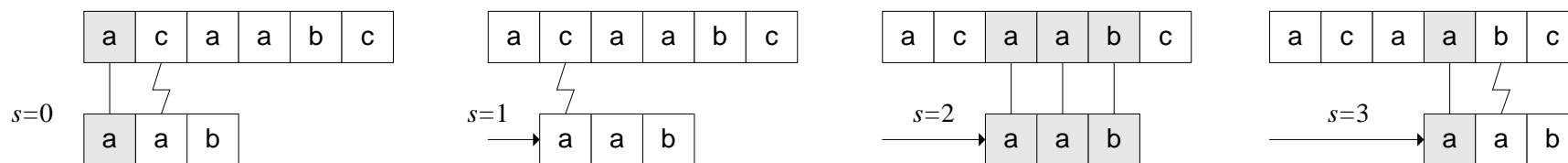
Ez a **brute-force** algoritmus egy adott eltolás esetén nem használja azokat az információkat, amelyeket a korábbi eltolások során már felderített. Pl. ha $P = \text{aaab}$ és $s = 0$ érvényes eltolás, akkor az 1, 2, és 3 értékű eltolások egyike sem lehet érvényes (hiszen $P[4] = \text{b}$).

A későbbiekben feltesszük, hogy azonos méretű sorozatok egyenlőségének vizsgálata megengedett, ún. primitív művelet.

Ha a sorozatokat balról-jobbra haladva hasonlítjuk össze, és megállunk amikor két jel nem egyezik, akkor a vizsgálat idejét az egyező jelek számának lineáris függvényeként kapjuk.

Pontosabban, az $x=y$ feltétel vizsgálatának ideje $\Theta(t+1)$, ahol t a leghosszabb olyan z sorozat hossza, amelyre $z \sqsubset x$ és $z \sqsubset y$.

Mintaillesztés



Az EGYSZERŰ-MINTAILLESZTŐ működése

Feladatok

- Legyen $T=000010001010001$, a minta $P=0001$. Milyen érvényes s eltolási értékeket kapunk az EGYSZERŰ-MINTAILLESZTŐ algoritmus végrehajtása során?
- Tegyük fel, hogy a P minta összes jele különböző. Hogyan gyorsítható fel az EGYSZERŰ-MINTAILLESZTŐ algoritmus úgy, hogy a futási ideje $O(n)$ legyen, ha a T szöveg hossza n ?



Mintaillesztés

Tegyük fel, hogy $\Sigma = \{0, 1, 2, \dots, 9\}$. Ekkor egy k hosszúságú jelsorozatot tekinthetünk egy k jegyű decimális számnak.

Egy adott $P[1..m]$ minta decimális értékét jelöljük p -vel, és egy $T[1..n]$ szöveg $T[s+1..s+m]$ m hosszúságú részsorozatának decimális értékét t_s -sel, bármely $s=0, 1, 2, \dots, n-m$ esetén.

Ekkor $t_s=p$ akkor és csak akkor, ha $T[s+1..s+m] = P[1..m]$, azaz s akkor és csak akkor érvényes eltolás, ha $t_s=p$.

A p értéke kiszámítható $\Theta(m)$ időben a **Horner-séma** segítségével:

$$p = P[m] + 10(P[m-1] + 10(P[m-2] + \dots + 10(P[2] + 10(P[1])) \dots).$$

A t_0 értéke hasonlóan meghatározható $T[1..m]$ segítségével $\Theta(m)$ idő alatt.

A t_1, t_2, \dots, t_{n-m} értékek $\Theta(n-m)$ időben kiszámíthatók, hiszen:

$$t_{s+1} = 10(t_s - 10^{m-1}T[s+1]) + T[s+m+1].$$

A $P[1..m]$ minta összes előfordulása tehát megtalálható a $T[1..n]$ szövegben $\Theta(m)$ előfeldolgozási és $\Theta(n-m+1)$ illesztési idő alatt.

Példa: Legyen $m=5$ és $t_s=31415$. Ha ki szeretnénk léptetni a legmagasabb helyiértékű $T[s+1]=3$ számjegyet, és beléptetni egy új (pl. $T[s+5+1]=2$ számjegyet a legalacsonyabb helyiértékre, akkor $t_{s+1} = 10(31415 - 10000 \cdot 3) + 2 = 14152$.



Mintaillesztés

Probléma: p és t_s értékek nagyok is lehetnek.

Megoldás: A p és t_s értékeket modulo q számítjuk, ahol q egy alkalmas modulus.

Általában, amikor az ábécé jeleinek száma d , és ezeket a $\{0, 1, \dots, d-1\}$ értékekkel azonosítjuk, akkor q értékének olyan prímszámot választunk, amelyre dq még ábrázolható a számítógépen.

Számolás:

$$t_{s+1} = 10(t_s - 10^{m-1}T[s+1]) + T[s+m+1] \quad (10\text{-es számrendszerben})$$

$$t_{s+1} = (d(t_s - T[s+1]h) + T[s+m+1]) \bmod q, \quad (d \text{ alapú számrendszerben } q \text{ modulussal})$$

ahol $h = d^{m-1} \pmod{q}$, egy m szélességű ablak legmagasabb helyiértékén szereplő „1”-nek megfelelő érték.

Probléma: A $t_s \equiv p \pmod{q}$ fennállásából nem következik a $t_s = p$ teljesülése, viszont ha a kongruencia nem teljesül, abból következik, hogy $t_s \neq p$, azaz s érvénytelen eltolás.

Megoldás: Minden olyan s eltolási érték, amelyre $t_s \equiv p \pmod{q}$ fennáll, további ellenőrzésre szorul, hogy eldönthessük s valóban érvényes, vagy csak egy **hamis találatot** határoz meg.

Ha q kellően nagy szám, akkor várhatóan ritkán lépnek fel hamis találatok, így az extra ellenőrzés költsége alacsony.

Mintaillesztés

RABIN-KARP-ILLESZTŐ(T, P, d, q)

```

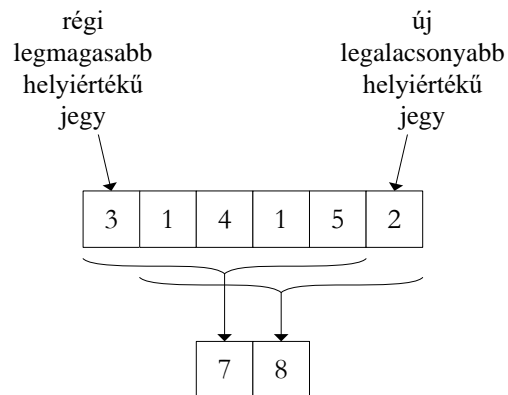
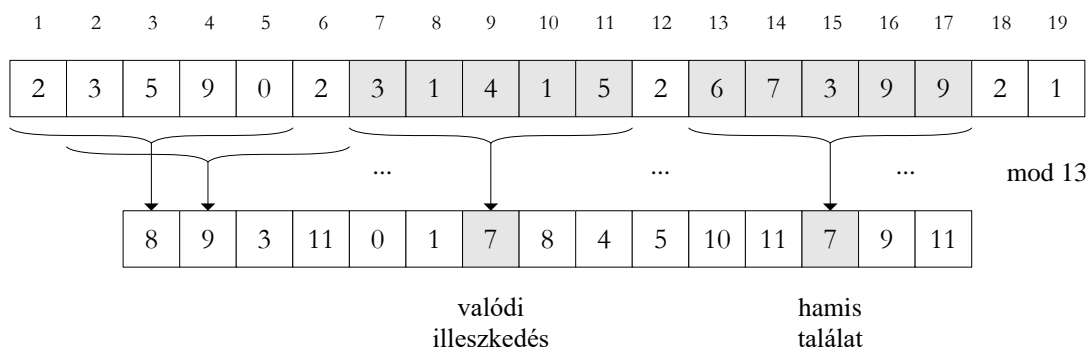
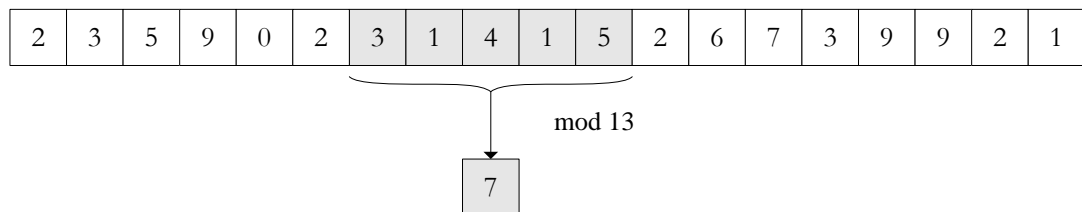
1   $n \leftarrow \text{hossz}[T]$ 
2   $m \leftarrow \text{hossz}[P]$ 
3   $h \leftarrow d^{m-1} \bmod q$ 
4   $p \leftarrow 0$ 
5   $t_0 \leftarrow 0$ 
6  for  $i \leftarrow 1, m$                                 /* Előfeldolgozás */
7       $p \leftarrow (dp + P[i]) \bmod q$ 
8       $t_0 \leftarrow (dt_0 + T[i]) \bmod q$ 
9  for  $s \leftarrow 0, n-m$                             /* Illesztés */
10     if  $p = t_s$ 
11         if  $P[1..m] = T[s+1..s+m]$ 
12             Ki: „A minta illeszkedik az”,  $s+1$ , „. pozícióra”
13     if  $s < n-m$ 
14          $t_{s+1} \leftarrow (d(t_s - T[s+1]h) + T[s+m+1]) \bmod q$ 
    
```

Hatékonyság: Az előfeldolgozás $\Theta(m)$, az illesztés $O((n-m+1)m)$ idejű.

Megjegyzés: Ha az érvényes eltolások száma várhatóan alacsony ($O(1)$) és q -nak a minta hosszánál nagyobb prímet választunk, akkor az illesztés várható ideje $O(n)$.



Mintaillesztés



régi legmagasabb helyiértékű jegy új legalacsonyabb helyiértékű jegy

léptetés

$$\begin{aligned}
 14152 &\equiv (31415 - 3 \cdot 10000) \cdot 10 + 2 \pmod{13} \\
 &\equiv (7 - 3 \cdot 3) \cdot 10 + 2 \pmod{13} \\
 &\equiv 8 \pmod{13}
 \end{aligned}$$

Feladatok

- Elhagyhatók-e a t értékek indexei a RABIN-KARP-ILLESZTŐ algoritmusban?
- Mi a legrosszabb eset a RABIN-KARP-ILLESZTŐ algoritmusnak?
- Legyen $T=314159265$, a minta $P=26$, és $q=11$. Mennyi a hamis találatok száma a RABIN-KARP-ILLESZTŐ algoritmus végrehajtása során?



Mintaillesztés

Egy M véges automata egy rendezett ötös, $(Q, q_0, A, \Sigma, \delta)$, ahol

- Q véges halmaz, az **állapotok** halmaza,
- $q_0 \in Q$ a **kezdőállapot**,
- $A \subseteq Q$ a **végállapotok** (vagy **elfogadó állapotok**) halmaza,
- Σ véges halmaz, a **bemeneti jelek** halmaza (vagy **bemeneti ábécé**),
- $\delta : Q \times \Sigma \rightarrow Q$ az automata **átmeneti függvénye**.

A véges automata **működése**:

- Kezdetben az automata a q_0 állapotban van.
- Az automata egyesével olvassa a bemeneti sorozatról a jeleket.
- Ha az automata a q állapotban az a jelet olvassa be, akkor „átmegy” a $\delta(q, a)$ állapotba.
- Ha az automata aktuális állapota végállapot, akkor M **elfogadja** az addig beolvasott sorozatot, különben **elveti** azt.

Mintaillesztés

Az M véges automata meghatároz egy \emptyset **végállapot függvényt**:

$\emptyset : \Sigma^* \rightarrow Q$, egy w sorozat esetén $\emptyset(w)$ az az állapot, amelybe M kerül w elolvasása után.

M akkor és csak akkor fogad el egy w sorozatot, ha $\emptyset(w) \in A$.

A \emptyset függvény rekurzív definíciója:

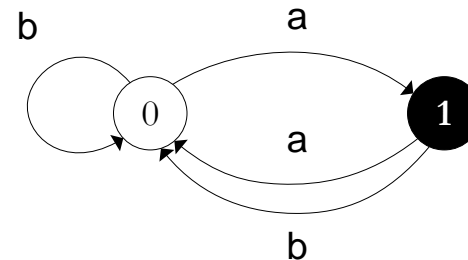
$$\emptyset(\varepsilon) = q_0,$$

$$\emptyset(wa) = \delta(\emptyset(w), a) \quad (w \in \Sigma^*, a \in \Sigma).$$

A δ átmeneti függvény

állapot	bemenet	
	a	b
0	1	0
1	0	0

Az állapot-átmenet diagram



Egy egyszerű, kétállapotú véges automata

Mintaillesztés

Minden P mintához létezik mintaillesztő automata, amelyet az alábbi módon konstruálhatunk meg:

- Meghatározzuk a $P[1..m]$ mintához tartozó σ **suffix függvényt**:
 - $\sigma : \Sigma^* \rightarrow \{0, 1, \dots, m\}$, és $\sigma(x)$ a leghosszabb olyan P -beli prefix hossza, amely suffixe x -nek:
 $\sigma(x) = \max\{k: P_k \sqsupseteq x\}$.
 - Pl: Ha $P=ab$, akkor $\sigma(\varepsilon)=0$, $\sigma(ccaca)=1$, $\sigma(ccab)=2$.
 - **Tulajdonságok:**
 - Egy m hosszúságú minta esetén $\sigma(x)=m$ akkor és csak akkor, ha $P \sqsupseteq x$.
 - Ha $x \sqsupseteq y$, akkor $\sigma(x) \leq \sigma(y)$.
- Az állapotok halmaza, Q , legyen $\{0, 1, \dots, m\}$. A q_0 kezdőállapot legyen 0, és az egyetlen végállapot m .
- Tetszőleges q állapotra és a jelre, a δ átmeneti függvényt az alábbi egyenlőség adja meg:
 $\delta(q, a) = \sigma(P_q a)$

Megjegyzés: a fenti választást az indokolja, hogy a $\emptyset(T_i) = \sigma(T_i)$ egyenlőség fennáll az automata működése során.

Mintaillesztés

ÁTMENETI-FÜGGVÉNY-SZÁMÍTÁS(P, Σ)

```

1   $m \leftarrow \text{hossz}[P]$ 
2  for  $q \leftarrow 0, m$ 
3      for minden  $a \in \Sigma$  jelre
4           $k \leftarrow \min(m+1, q+2)$ 
5          repeat
6               $k \leftarrow k-1$ 
7          until  $P_k \supset P_q a$ 
8           $\delta(q, a) \leftarrow k$ 
9  return  $\delta$ 
    
```

Hatékonyság: A **repeat** ciklus legfeljebb $m+1$ iterációt hajt végre, a $P_k \supset P_q a$ feltétel eldöntéséhez m jel összehasonlítására lehet szükség, így kapható az $O(m^3|\Sigma|)$.

Megjegyzés: δ meghatározható $O(m|\Sigma|)$ időben is (ha kihasználunk bizonyos, a P mintára kiszámolható információkat).

Feladatok

- Adjuk meg a $P=aabab$ mintához tartozó mintaillesztő automata δ átmeneti függvényét (táblázattal) és állapot-átmenet diagramját, ha $\Sigma=\{a, b\}$!

Mintaillesztés

VÉGES-AUTOMATA-ILLESZTŐ(T, δ, m)

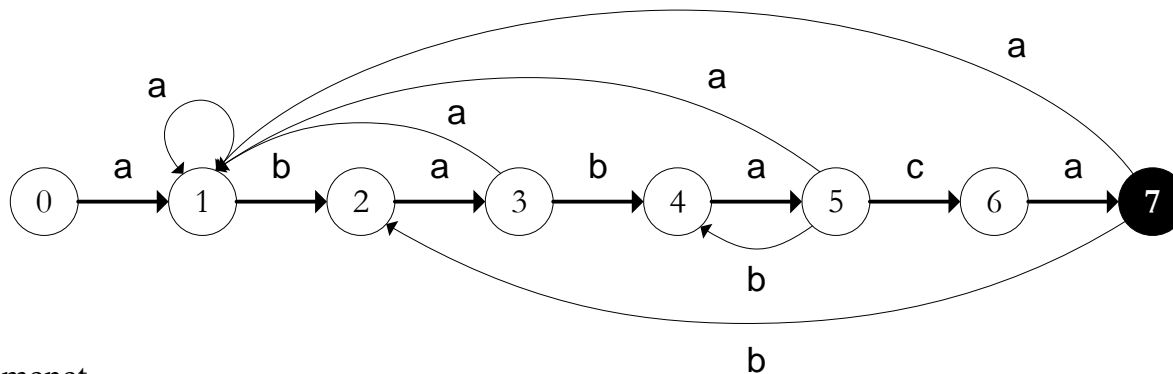
```

1   $n \leftarrow \text{hossz}[T]$ 
2   $q \leftarrow 0$ 
3  for  $i \leftarrow 1, n$ 
4       $q \leftarrow \delta(q, T[i])$ 
5      if  $q = m$ 
6          Ki: „A minta illeszkedik az”,  $i-m+1$ ,”. pozícióra”
    
```

Hatékonyság: Egy n hosszúságú szöveg esetén az illesztési idő $\Theta(n)$.

Összegezve: Egy m hosszúságú minta összes előfordulása egy n hosszúságú, Σ jeleiből álló sorozatban megtalálható $O(m|\Sigma|)$ előfeldolgozási idő és $\Theta(n)$ illesztési idő alatt.

Mintaillesztés



állapot	bemenet			P
	a	b	c	
0	1	0	0	a
1	1	2	0	b
2	3	0	0	a
3	1	4	0	b
4	5	0	0	a
5	1	4	6	c
6	7	0	0	a
7	1	2	0	

i	1	2	3	4	5	6	7	8	9	10	11	
$T[i]$	a	b	a	b	a	b	a	c	a	b	a	
$\emptyset(T_i)$ állapot	0	1	2	3	4	5	4	5	6	7	2	3

Egy mintaillesztő automata



Mintaillesztés

- Hogyan tudnánk gyorsítani az egyszerű mintaillesztést?
 - Lehetőleg ne egyesével léptessük a mintát a szövegen, az érvénytelen eltolásokat eleve ugorjuk át!
 - A már egyszer megvizsgált és egyező jeleket ne vizsgáljuk meg egy újabb eltolás vizsgálata esetén!
- Általában az alábbi kérdésre kellene ismerni a választ:
 - Ha a minta $P[1..q]$ jelei illeszkednek a szöveg $T[s+1..s+q]$ jeleire, mi a legkisebb olyan $s' > s$ eltolás, amelyre fennáll, hogy

$$P[1..k] = T[s'+1..s'+k], \quad \text{ahol } s' + k = s + q?$$
- Más szavakkal, ha tudjuk, hogy $P_q \sqsupset T_{s+q}$, akkor tudni szeretnénk, hogy mi az a leghosszabb P_k valódi prefixe P_q -nak ami szintén szuffixe T_{s+q} -nak.
 - $s' = s + q - k$ az első s -nél nagyobb eltolás, amely nem feltétlenül érvénytelen.
 - A legjobb esetben $k = 0$, azaz $s' = s + q$, így az $s+1, s+2, \dots, s+q-1$ eltolások azonnal kizárhatók.
 - Az s' eltolásra felesleges ellenőriznünk a P minta első k jelének illeszkedését, mert azok egyeznek a szöveg megfelelő jeleivel.
- Kiszámítjuk és felhasználjuk a minta prefix függvényét, amely megadja, hogyan illeszkedik a minta önmaga eltoltjaira.

Mintaillesztés

A π függvény a $P[1..m]$ mintához tartozó **prefix függvény**, ha $\pi : \{1, 2, \dots, m\} \rightarrow \{0, 1, \dots, m-1\}$ és

$$\pi[q] = \max\{k : k < q \text{ és } P_k \supset P_q\}.$$

Szavakban: $\pi[q]$ a leghosszabb olyan P -beli prefix hossza, amely valódi szuffixe P_q -nak.

PREFIX-FÜGGVÉNY-SZÁMÍTÁS(P)

```

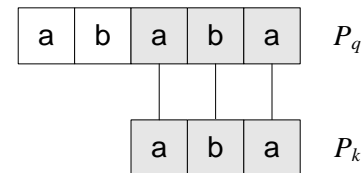
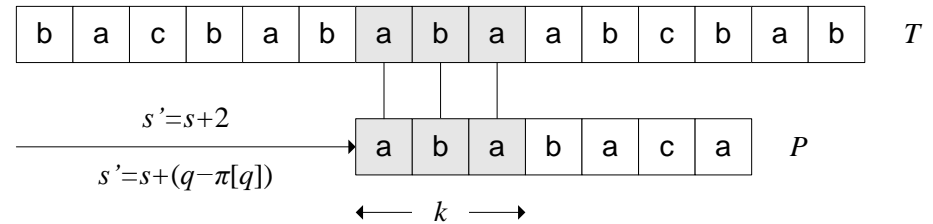
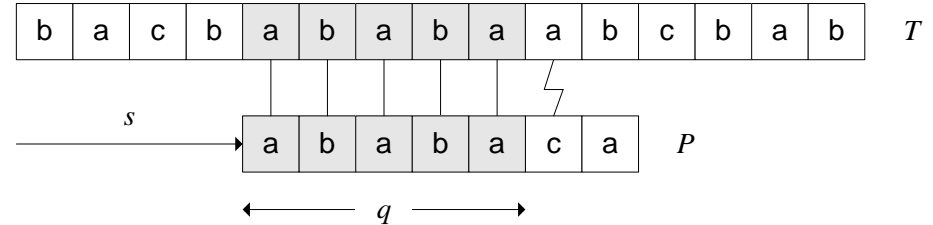
1   $m \leftarrow \text{hossz}[P]$ 
2   $\pi[1] \leftarrow 0$ 
3   $k \leftarrow 0$ 
4  for  $q \leftarrow 2, m$ 
5      while  $k > 0$  és  $P[k+1] \neq P[q]$ 
6           $k \leftarrow \pi[k]$ 
7      if  $P[k+1] = P[q]$ 
8           $k \leftarrow k+1$ 
9       $\pi[q] \leftarrow k$ 
10 return  $\pi$ 
```

Hatékonyság: $\Theta(m)$, mivel a belső ciklus $O(1)$ idejű (lásd jegyzet, amortizáló elemzés).



Mintaillesztés

i	1	2	3	4	5	6	7
$P[i]$	a	b	a	b	a	c	a
$\pi[i]$	0	0	1	2	3	0	1



A π prefix függvény és használata

Feladatok

- Adjuk meg a $P=ababbababbababbab$ mintához tartozó π prefix függvényt, ha $\Sigma=\{a, b\}$!



Mintaillesztés

KMP-ILLESZTŐ(T, P)

```

1   $n \leftarrow \text{hossz}[T]$ 
2   $m \leftarrow \text{hossz}[P]$ 
3   $\pi \leftarrow \text{PREFIX-FÜGGVÉNY-SZÁMÍTÁS}(P)$ 
4   $q \leftarrow 0$                                      /* Az illeszkedő jelek száma */
5  for  $i \leftarrow 1, n$                              /* A szöveg ellenőrzése balról jobbra */
6      while  $q > 0$  és  $P[q+1] \neq T[i]$ 
7           $q \leftarrow \pi[q]$                        /* A következő jel nem illeszkedik */
8      if  $P[q+1] = T[i]$ 
9           $q \leftarrow q+1$                            /* A következő jel illeszkedik */
10     if  $q = m$                                      /* A teljes minta illeszkedik? */
11         Ki: „A minta előfordul az”,  $i-m+1$ , „. pozíción”
12          $q \leftarrow \pi[q]$                        /* A következő illeszkedés keresése */
```

Hatékonyság: Előfeldolgozás (π kiszámítása) $\Theta(m)$, az illesztés $\Theta(n)$ idejű (mivel a belső ciklus $O(1)$ idejű itt is).

Megjegyzés: A KMP rövidítést a Knuth-Morris-Pratt nevek kezdőbetűi adják.



Feladatok

- Milyen értékeket vesz fel rendre a q változó a KMP-ILLESZTŐ algoritmus futása során, a $T=bbababababba$ szöveg, és a $P=ababa$ minta bemenő adatokra, ha $\Sigma=\{a, b\}$?



Mintaillesztés

Algoritmus	Előfeldolgozási idő	Illesztési idő
Egyszerű	0	$O((n-m+1)m)$
Rabin-Karp	$\Theta(m)$	$O((n-m+1)m)$
Véges automata	$O(m \Sigma)$	$\Theta(n)$
Knuth-Morris-Pratt	$\Theta(m)$	$\Theta(n)$

A mintaillesztő algoritmusok hatékonysága

