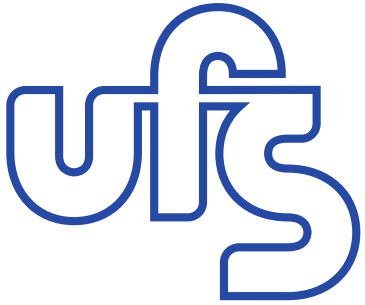
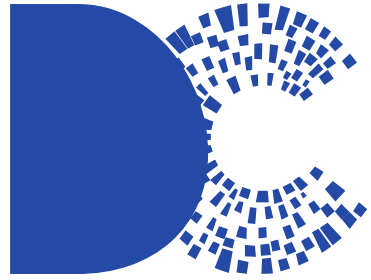


UNIVERSIDADE FEDERAL DE SERGIPE



DEPARTAMENTO DE COMPUTAÇÃO

Busca em cadeias (Knuth-Morris-Pratt)

Projeto e Análise de Algoritmos

Bruno Prado

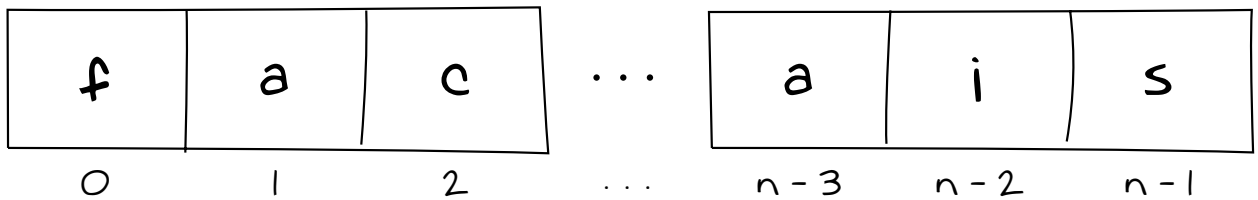
Departamento de Computação / UFS

Introdução

- O que é uma cadeia?
  - É uma sequência de símbolos T com tamanho n

Introdução

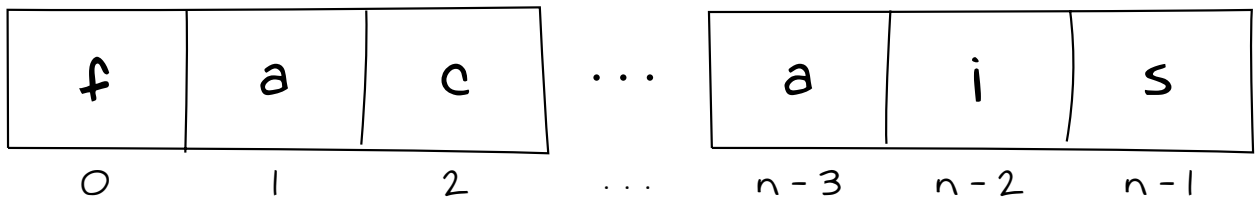
- O que é uma cadeia?
  - É uma sequência de símbolos T com tamanho n
  - Os símbolos são definidos por um alfabeto finito $\sum$



$$\sum = \{a, b, \ldots, y, z\}$$

Introdução

- O que é uma cadeia?
  - É uma sequência de símbolos T com tamanho n
  - Os símbolos são definidos por um alfabeto finito $\sum$

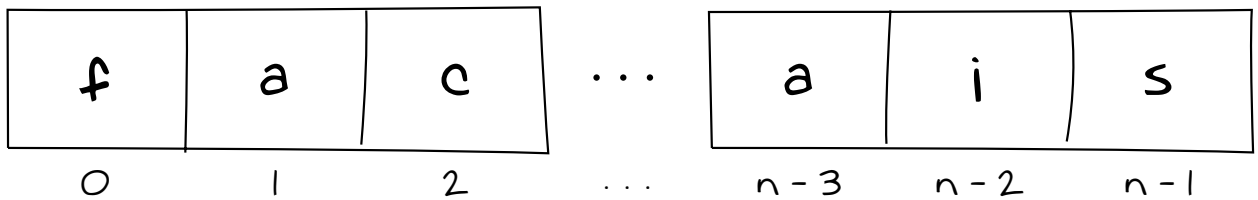


$$\sum = \{a, b, \ldots, y, z\}$$

  - Aplicações multidisciplinares
    - Biologia: representação da cadeia de DNA, sendo composta pelos símbolos A, C, G, T

Introdução

- O que é uma cadeia?
  - É uma sequência de símbolos T com tamanho n
  - Os símbolos são definidos por um alfabeto finito $\sum$

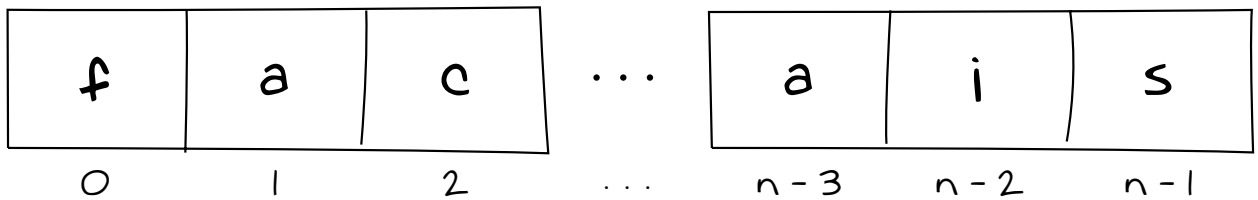


$$\sum = \{a, b, \ldots, y, z\}$$

  - Aplicações multidisciplinares
    - Biologia: representação da cadeia de DNA, sendo composta pelos símbolos A, C, G, T
    - Computação: armazenamento de texto através do tipo *string*, com o padrão de codificação ASCII

Introdução

- O que é uma cadeia?
  - É uma sequência de símbolos T com tamanho n
  - Os símbolos são definidos por um alfabeto finito $\sum$



$$\sum = \{a, b, \ldots, y, z\}$$

  - Aplicações multidisciplinares
    - Biologia: representação da cadeia de DNA, sendo composta pelos símbolos A, C, G, T
    - Computação: armazenamento de texto através do tipo *string*, com o padrão de codificação ASCII
    - ...

Introdução

- Notação e terminologia
  - Todos as cadeias de tamanho finito que podem ser construídas do alfabeto finito $\sum$ é definido por $\sum^*$

Introdução

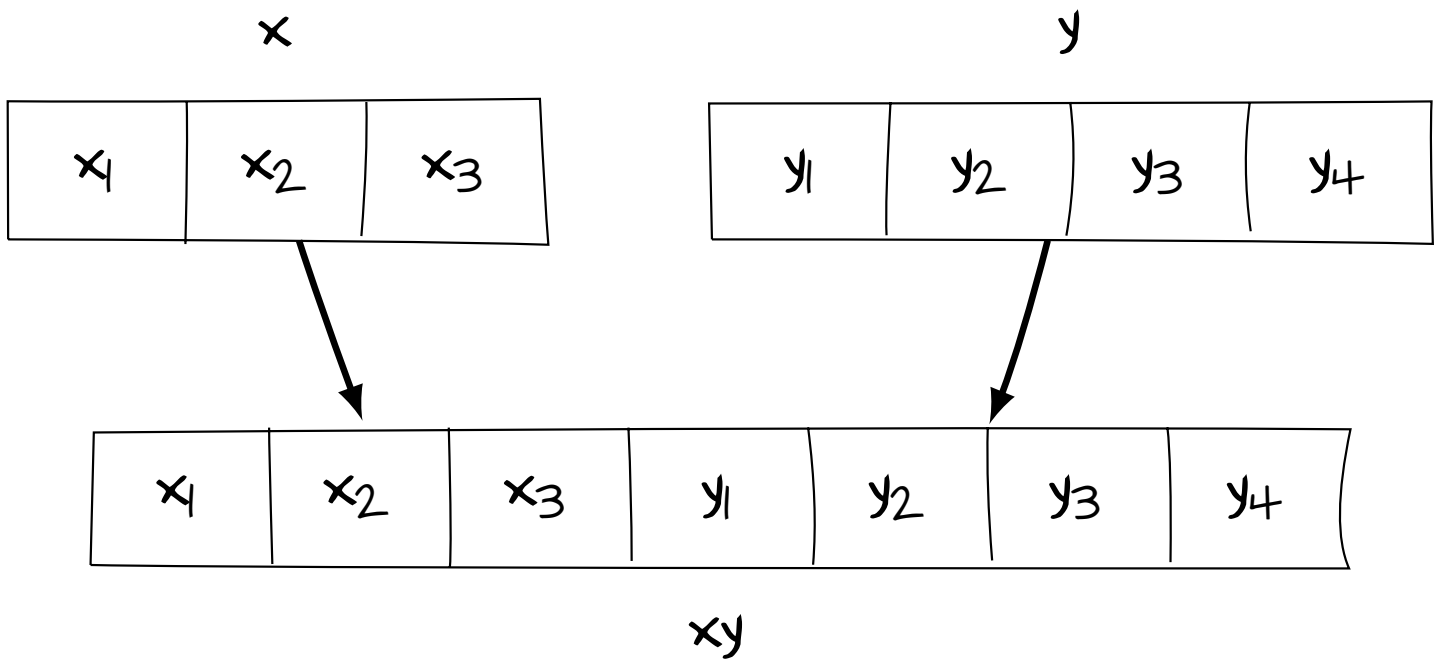
- Notação e terminologia
  - Todos as cadeias de tamanho finito que podem ser construídas do alfabeto finito $\sum$ é definido por $\sum^*$
  - Uma cadeia vazia é denotada pelo símbolo ε

Introdução

- Notação e terminologia
  - Todos as cadeias de tamanho finito que podem ser construídas do alfabeto finito $\sum$ é definido por $\sum^*$
  - Uma cadeia vazia é denotada pelo símbolo ε
  - O tamanho de uma cadeia x é definida por $|x|$

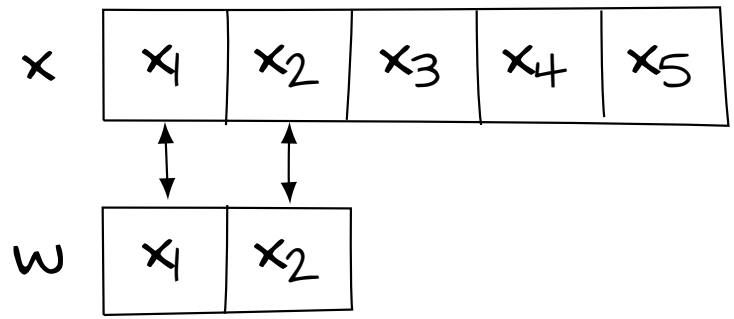
Introdução

- Notação e terminologia
  - Todos as cadeias de tamanho finito que podem ser construídas do alfabeto finito $\sum$ é definido por $\sum^*$
  - Uma cadeia vazia é denotada pelo símbolo ε
  - O tamanho de uma cadeia x é definida por $|x|$
  - A concatenação de duas cadeias x e y resulta em uma cadeia xy com os caracteres de x seguidos dos caracteres de y, com tamanho total de $|x| + |y|$



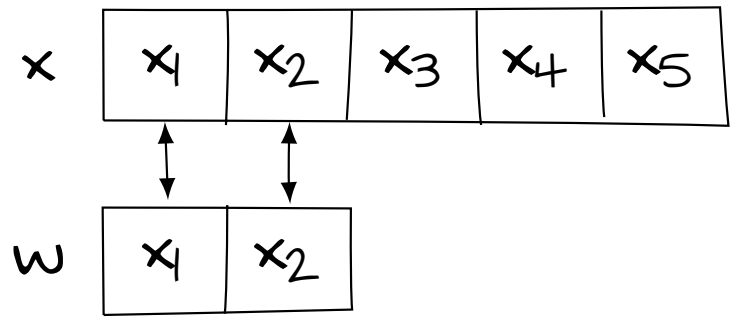
Introdução

- Notação e terminologia
  - Prefixo: a cadeia w é um prefixo da cadeia x ($w \sqsubset x$) se $x = wy$, com $y \in \sum^*$ e $|w| \leq |x|$

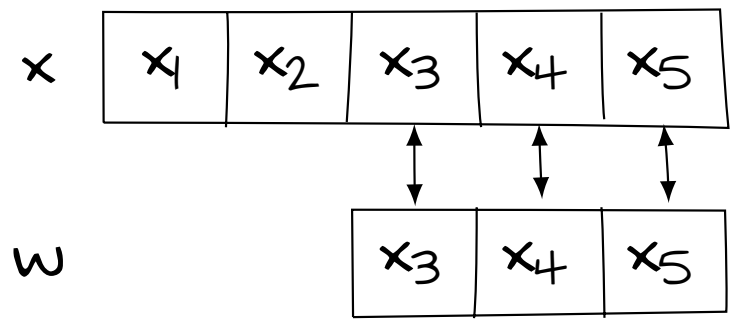


Introdução

- Notação e terminologia
  - Prefixo: a cadeia w é um prefixo da cadeia x ($w \sqsubset x$) se $x = wy$, com $y \in \sum^*$ e $|w| \leq |x|$

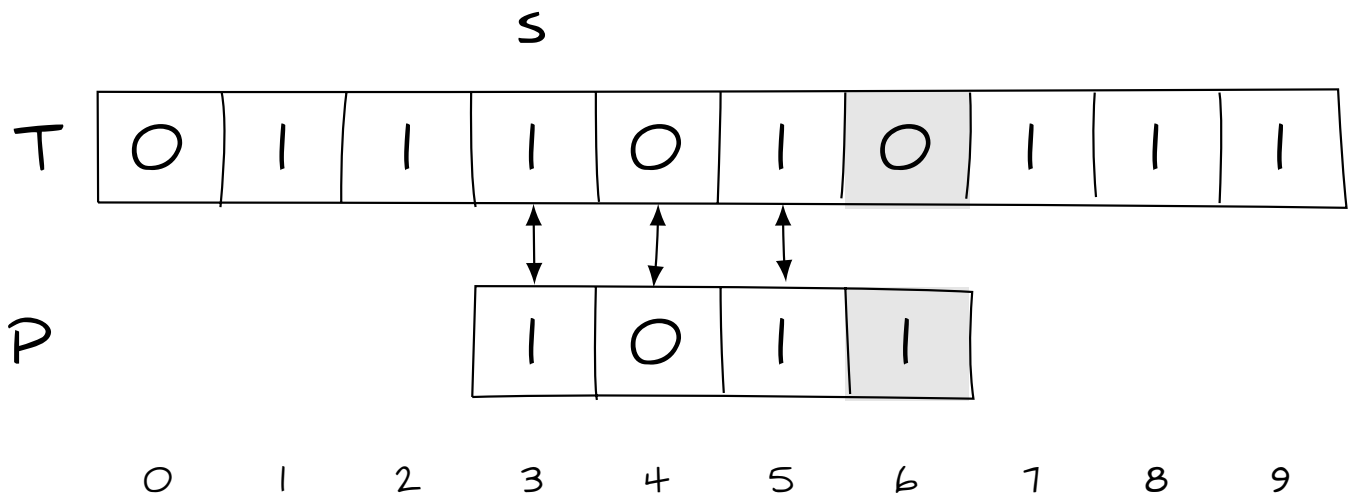


  - Sufixo: a cadeia w é sufixo da cadeia x ($w \sqsupset x$) se $x = yw$, com $y \in \sum^*$ e $|w| \leq |x|$



Introdução

- Como pode ser definida a busca em cadeias?
  - É o processo para encontrar todas as ocorrências de um padrão P em uma cadeia T que possuem m e n símbolos, respectivamente, onde $m \leq n$



$$\sum = \{0, 1\}$$
$$|P| = m = 4,\ |T| = n = 10$$
$$0 \leq s \leq n - m$$

Busca em cadeias

- Knuth-Morris-Pratt (KMP)
  - É um algoritmo linear para busca em cadeia que utiliza preprocessamento do padrão, armazenando uma tabela para comparação em tempo constante

Busca em cadeias

- Knuth-Morris-Pratt (KMP)
  - É um algoritmo linear para busca em cadeia que utiliza preprocessamento do padrão, armazenando uma tabela para comparação em tempo constante
  - O princípio de funcionamento é baseado em autômatos finitos e na tabela de transição

Busca em cadeias

- Knuth-Morris-Pratt (KMP)
  - É um algoritmo linear para busca em cadeia que utiliza preprocessamento do padrão, armazenando uma tabela para comparação em tempo constante
  - O princípio de funcionamento é baseado em autômatos finitos e na tabela de transição
  - Em cada posição da tabela é armazenado o comprimento do maior prefixo de P_i que é um sufixo de P_j através da função de prefixo k

$$k(i) = max\left\{(j-1) : (j < i) \wedge \left(P_j \sqsupset P_i\right)\right\}$$

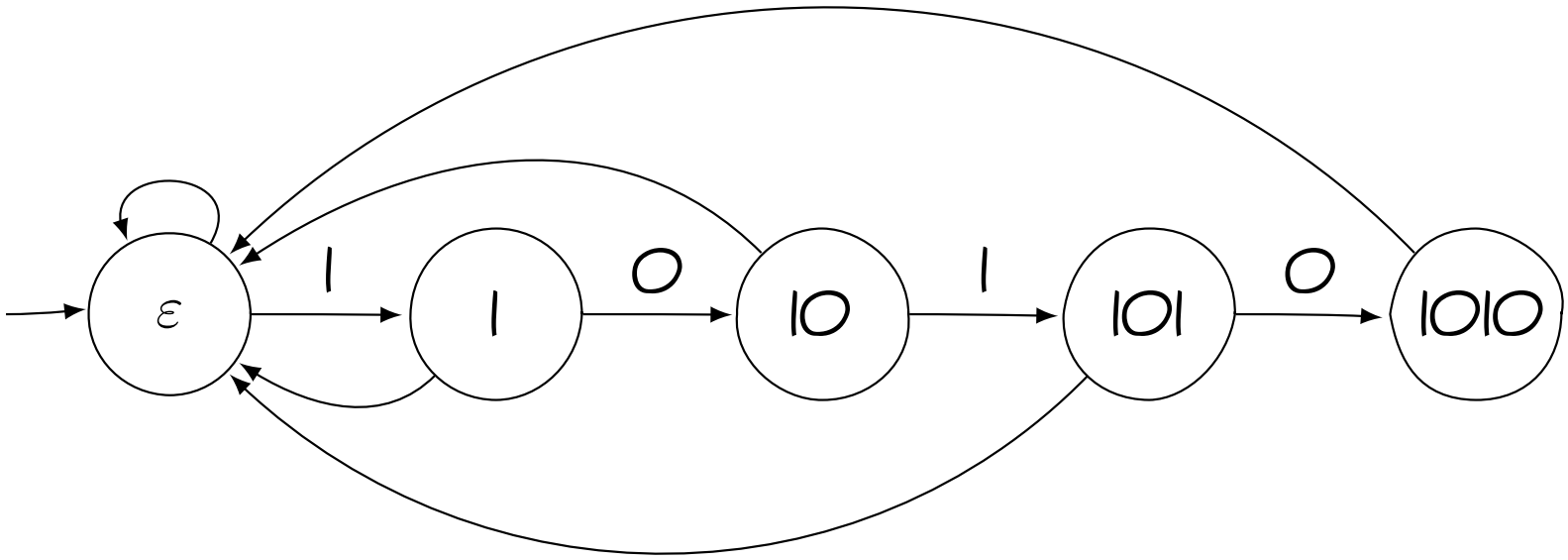
Busca em cadeias

- Knuth-Morris-Pratt (KMP)
  - Cálculo da tabela de transição

```
1  // Padrão de tipos por tamanho
2  #include <stdint.h>
3  // Procedimento de cálculo da tabela de transição
4  void calcular_tabela(int32_t* k, char* P) {
5      // i = sufixo, j = prefixo
6      for(int32_t i = 1, j = -1; i < strlen(P); i++) {
7          // Prefixo e sufixo diferentes
8          while(j >= 0 && P[j + 1] != P[i])
9              // Retorno de estado
10             j = k[j];
11         // Combinação de prefixo e sufixo
12         if(P[j + 1] == P[i])
13             // Avanço de estado
14             j++;
15         // Atualização da transição do estado
16         k[i] = j;
17     }
18 }
```

Busca em cadeias

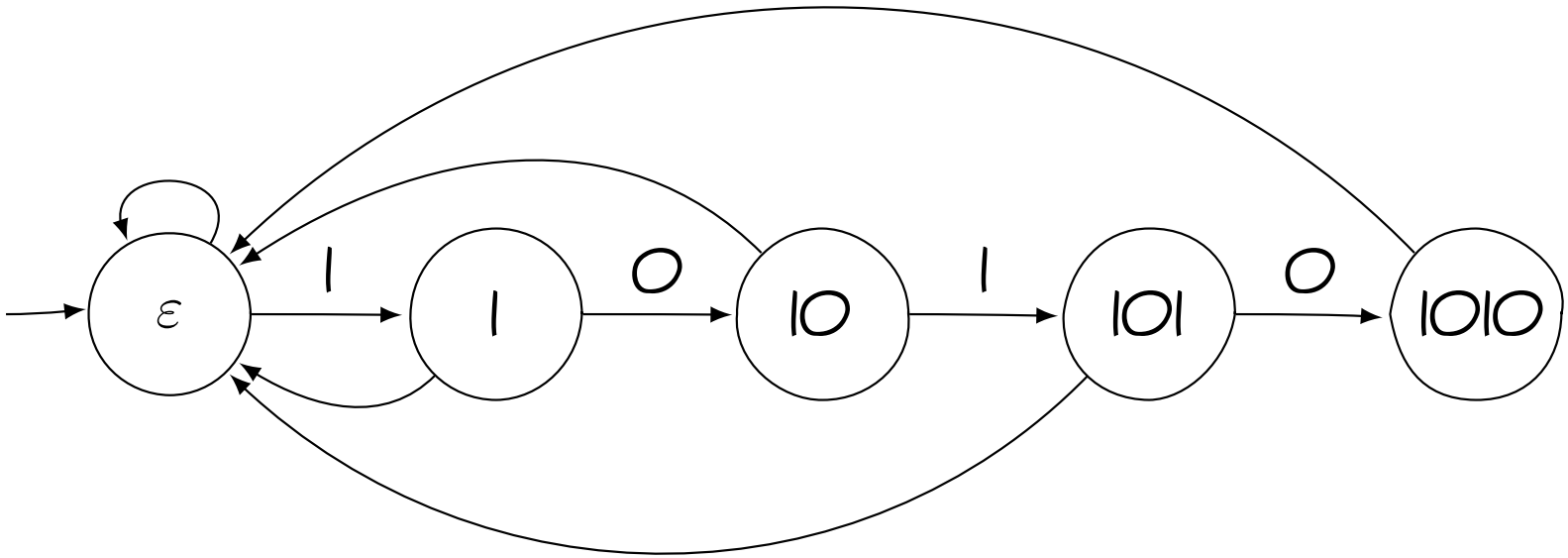
- Knuth-Morris-Pratt (KMP)
  - Cálculo da tabela de transição



P	1	0	1	0
k	-1	-1	-1	-1
-1	0	1	2	3
j		i		

Busca em cadeias

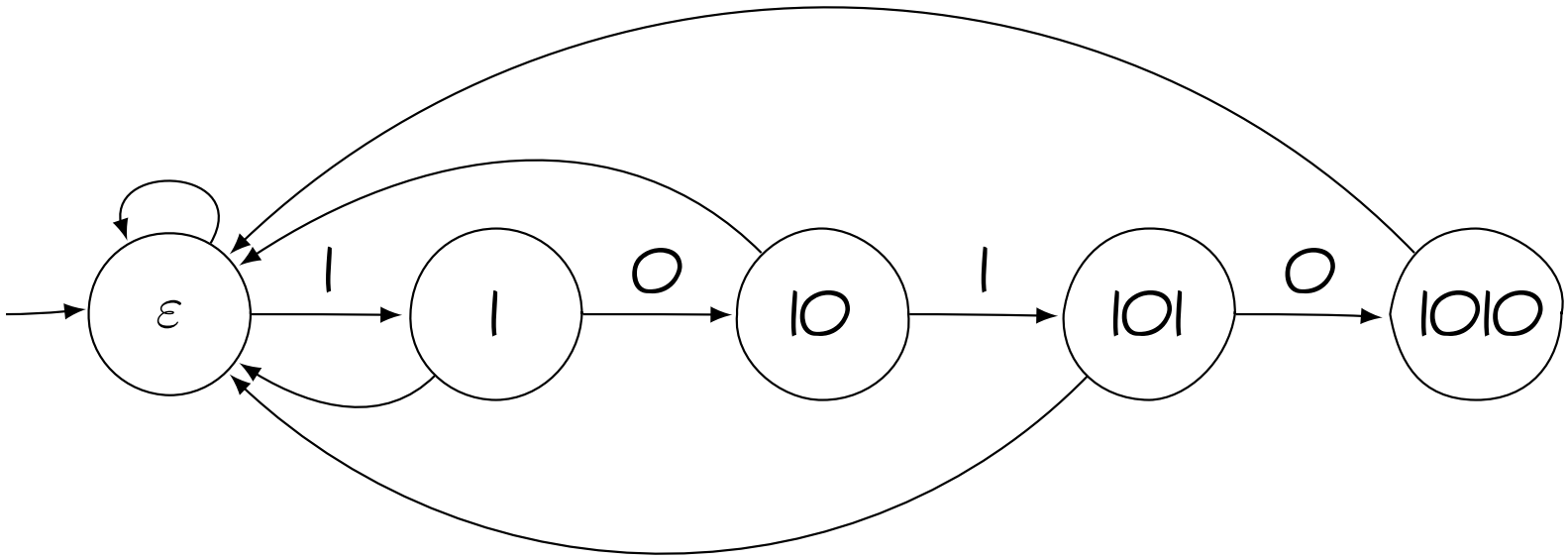
- Knuth-Morris-Pratt (KMP)
  - Cálculo da tabela de transição



P	1	0	1	0
k	-1	-1	-1	-1
-1	0	1	2	3
j		i		

Busca em cadeias

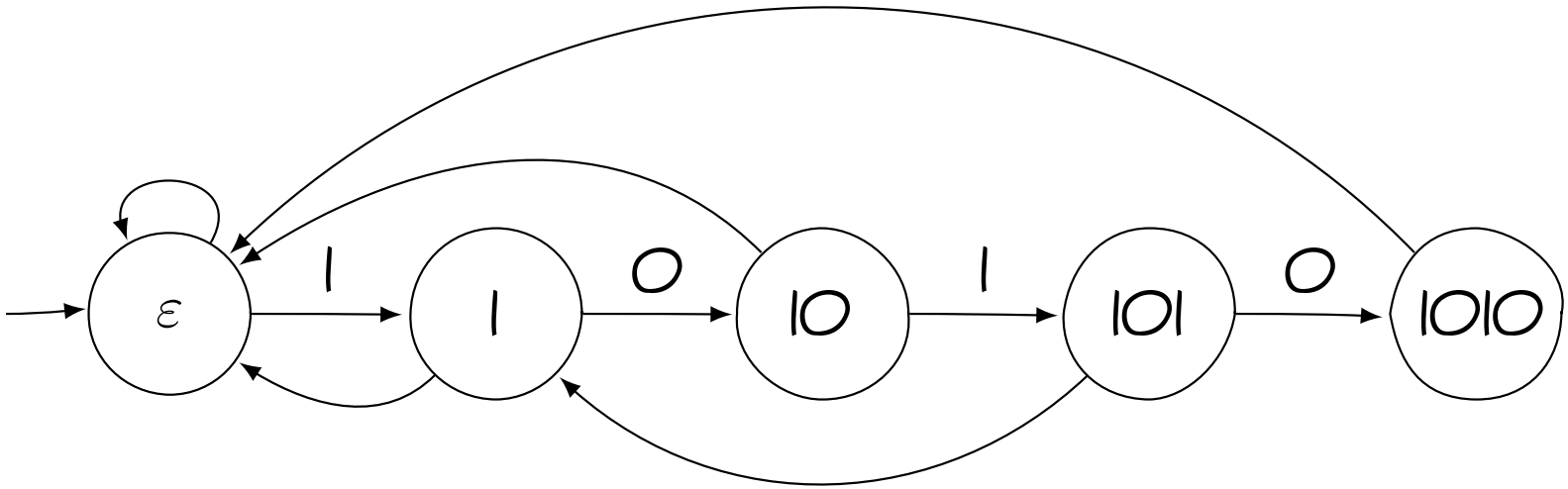
- Knuth-Morris-Pratt (KMP)
  - Cálculo da tabela de transição

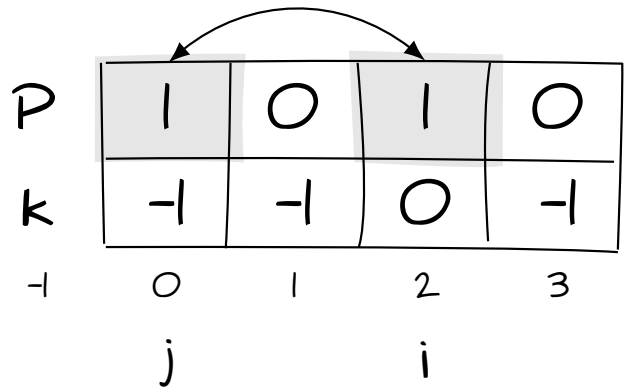


P	1	0	1	0
k	-1	-1	-1	-1
-1	0	1	2	3
j			i	

Busca em cadeias

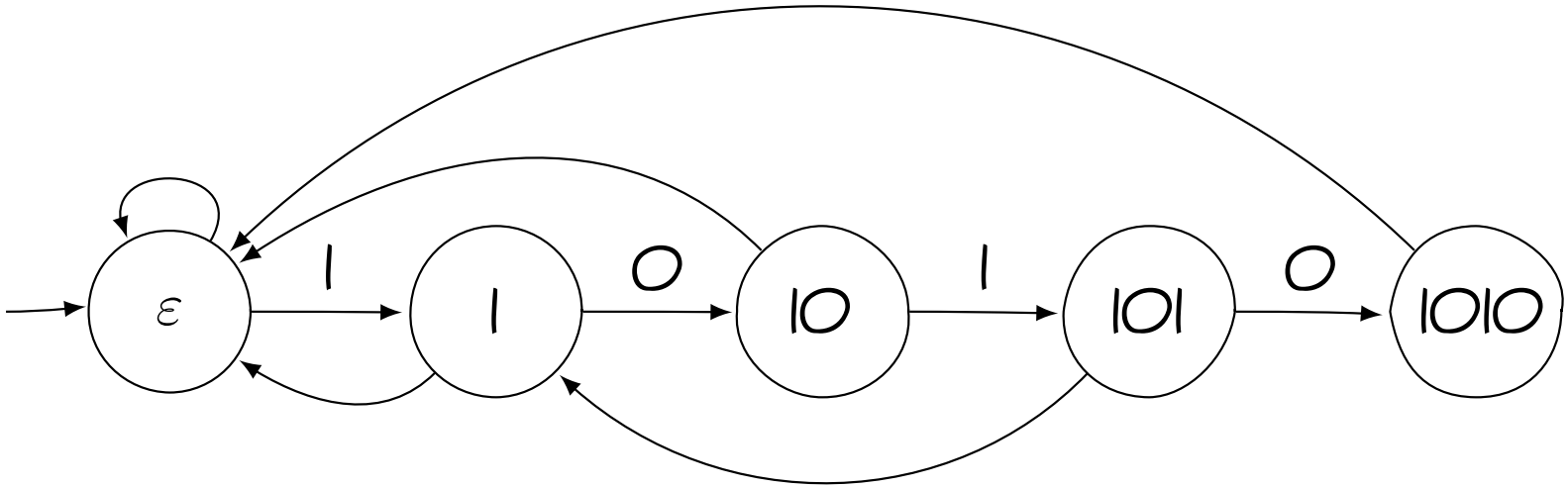
- Knuth-Morris-Pratt (KMP)
  - Cálculo da tabela de transição





Busca em cadeias

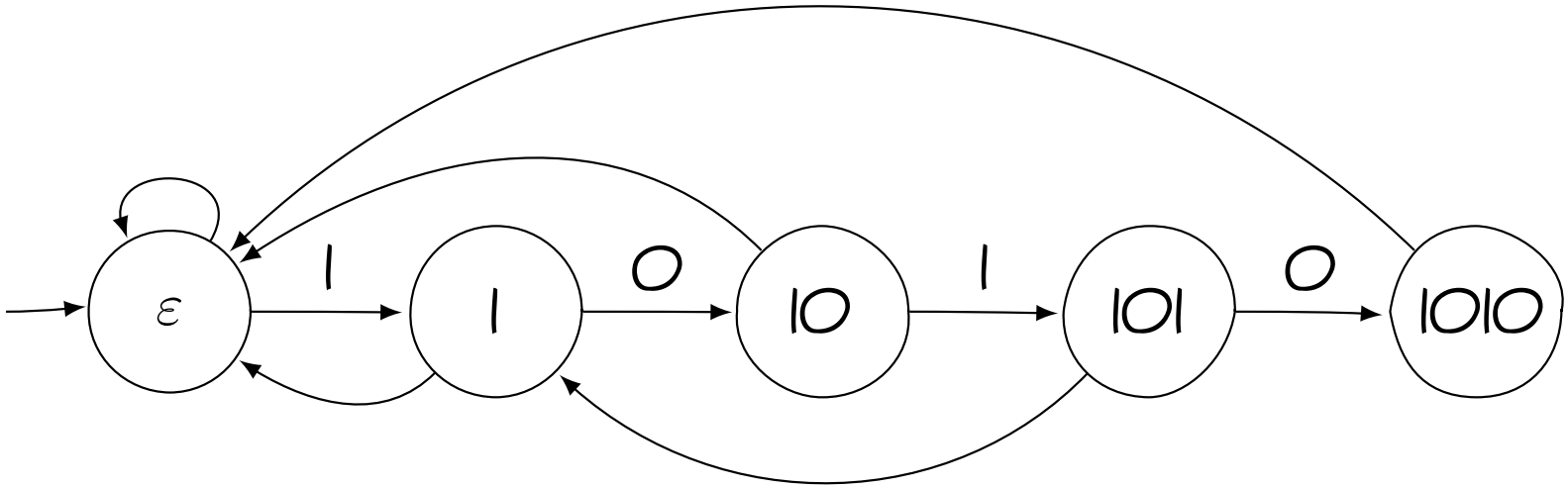
- Knuth-Morris-Pratt (KMP)
  - Cálculo da tabela de transição



P	1	0	1	0
k	-1	-1	0	-1
	-1	0	1	2	3
		j			i

Busca em cadeias

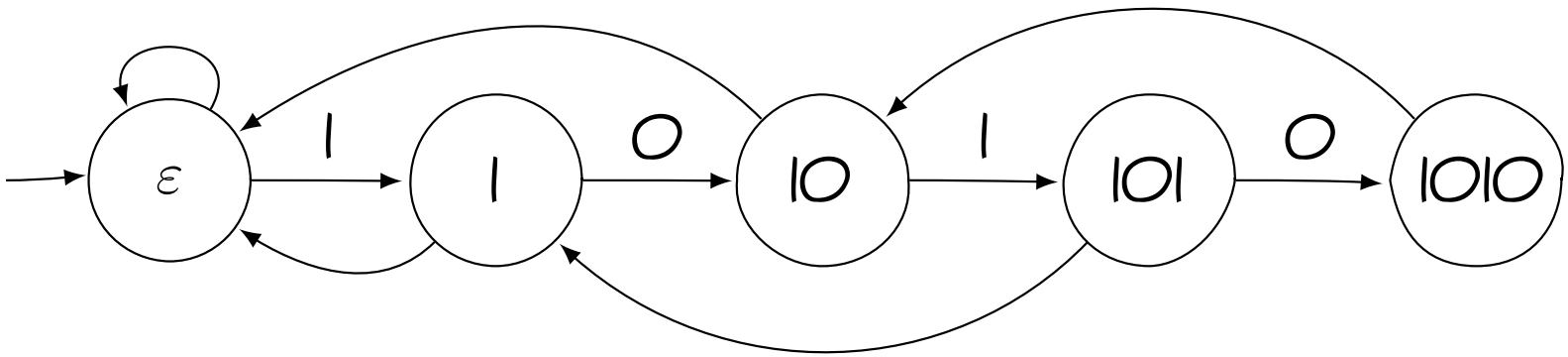
- Knuth-Morris-Pratt (KMP)
  - Cálculo da tabela de transição

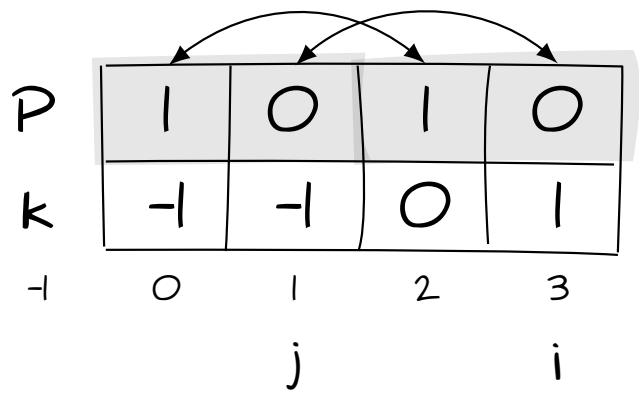


P	1	0	1	0
k	-1	-1	0	-1
	-1	0	1	2	3
		j			i

Busca em cadeias

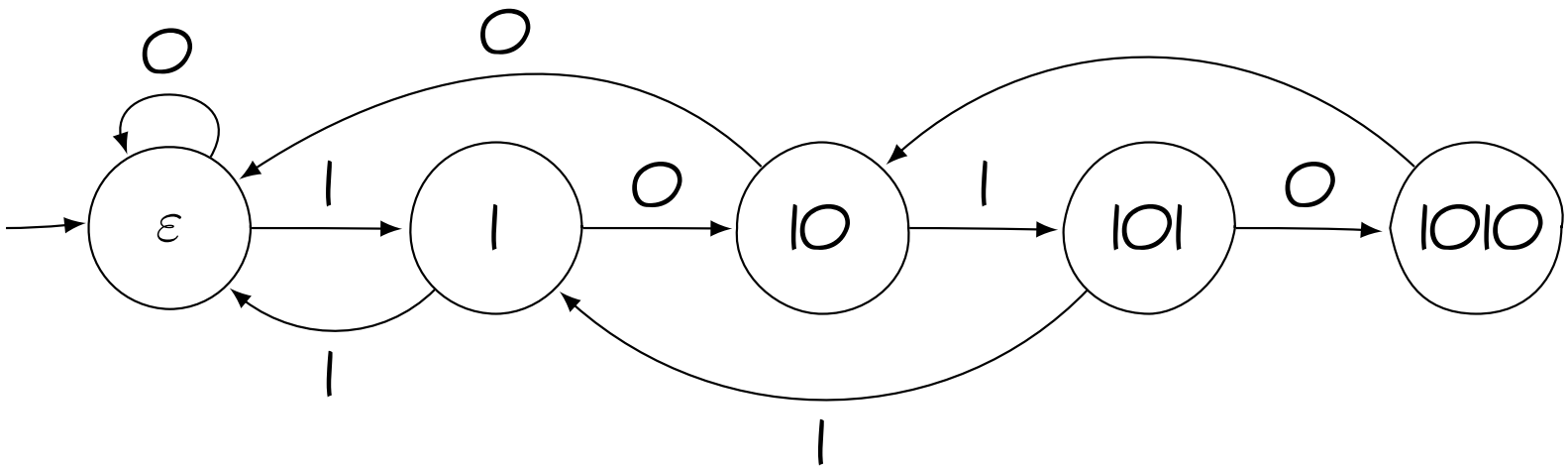
- Knuth-Morris-Pratt (KMP)
  - Cálculo da tabela de transição





Busca em cadeias

- Knuth-Morris-Pratt (KMP)
  - Autômato Finito Determinístico



P	1	0	1	0
k	-1	-1	0	1
	-1	0	1	2	3

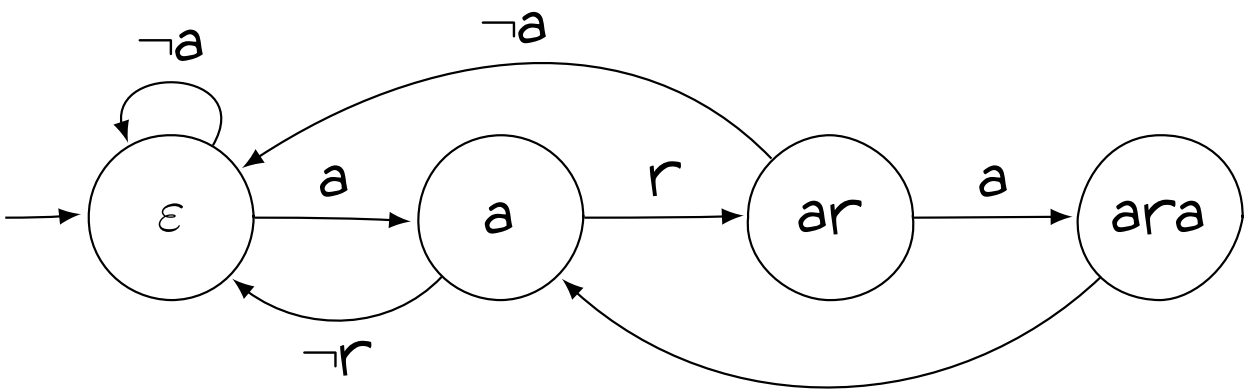
Busca em cadeias

- Knuth-Morris-Pratt (KMP)

```
1  // Padrão de tipos por tamanho
2  #include <stdint.h>
3  // Busca por KMP
4  void KMP(int32_t* k, int32_t* R, char* T, char* P) {
5      // Pré-processamento
6      int32_t n = strlen(T), m = strlen(P);
           calcular_tabela(k, P);
7      // Iterando na cadeia T
8      for(int32_t i = 0, j = -1; i < n; i++) {
9          // Retorno de estado
10         while(j >= 0 && P[j + 1] != T[i]) j = k[j];
11         // Avanço de estado
12         if(P[j + 1] == T[i]) j++;
13         // Combinação do padrão
14         if(j == m - 1) {
15             inserir(R, i - m + 1);
16             j = k[j];
17         }
18     }
19 }
```

Busca em cadeias

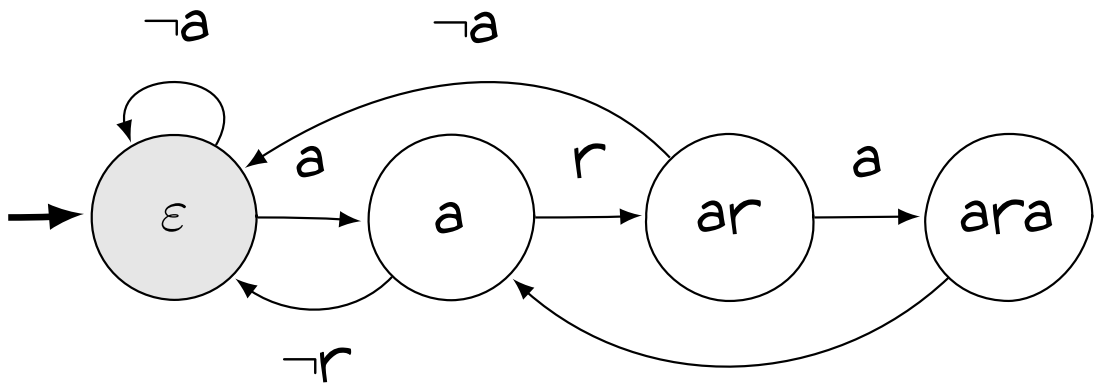
- Knuth-Morris-Pratt (KMP)
  - Tabela de transição do padrão $P = ara$

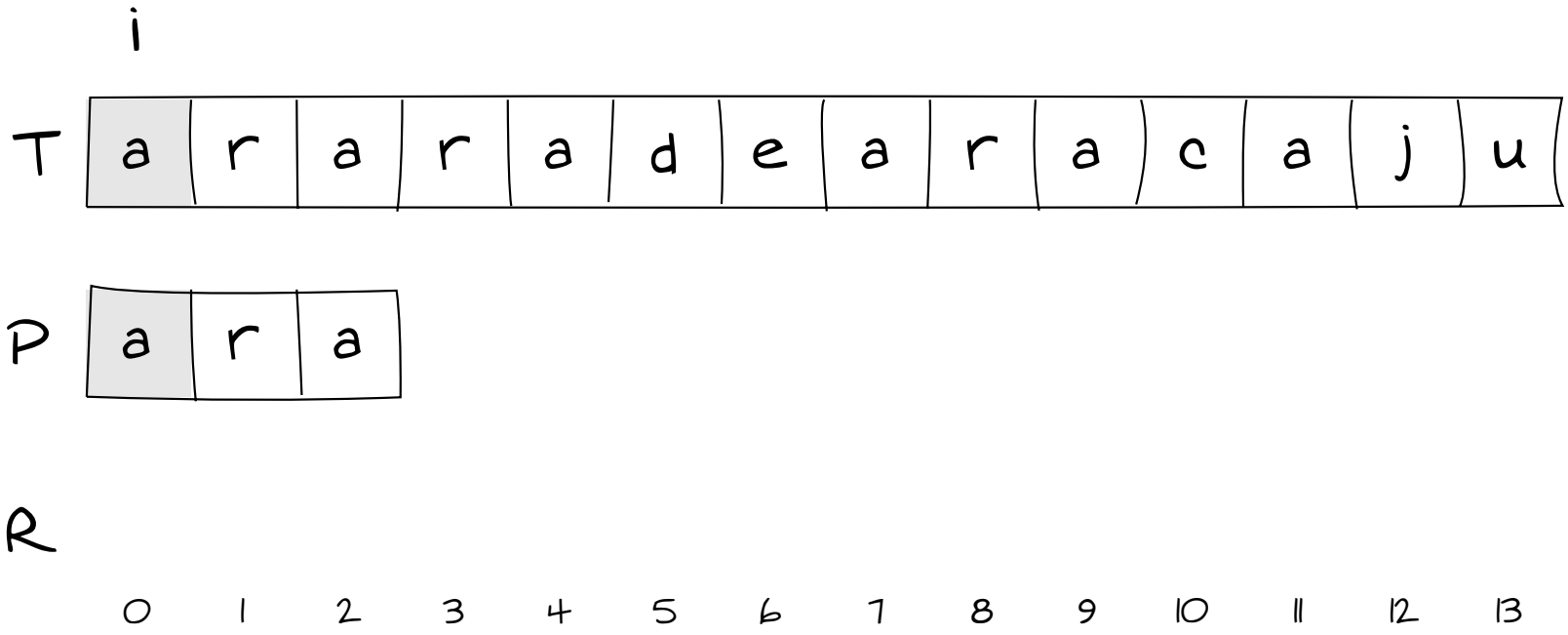


P	a	r	a
k	-1	-1	0
	-1	0	1	2

Busca em cadeias

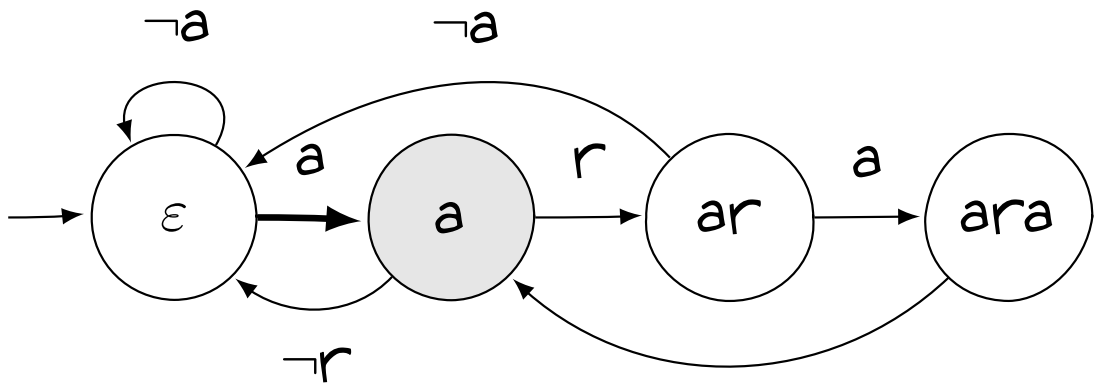
- Knuth-Morris-Pratt (KMP)
  - Buscando $P = ara$ em $T = araradearacaju$

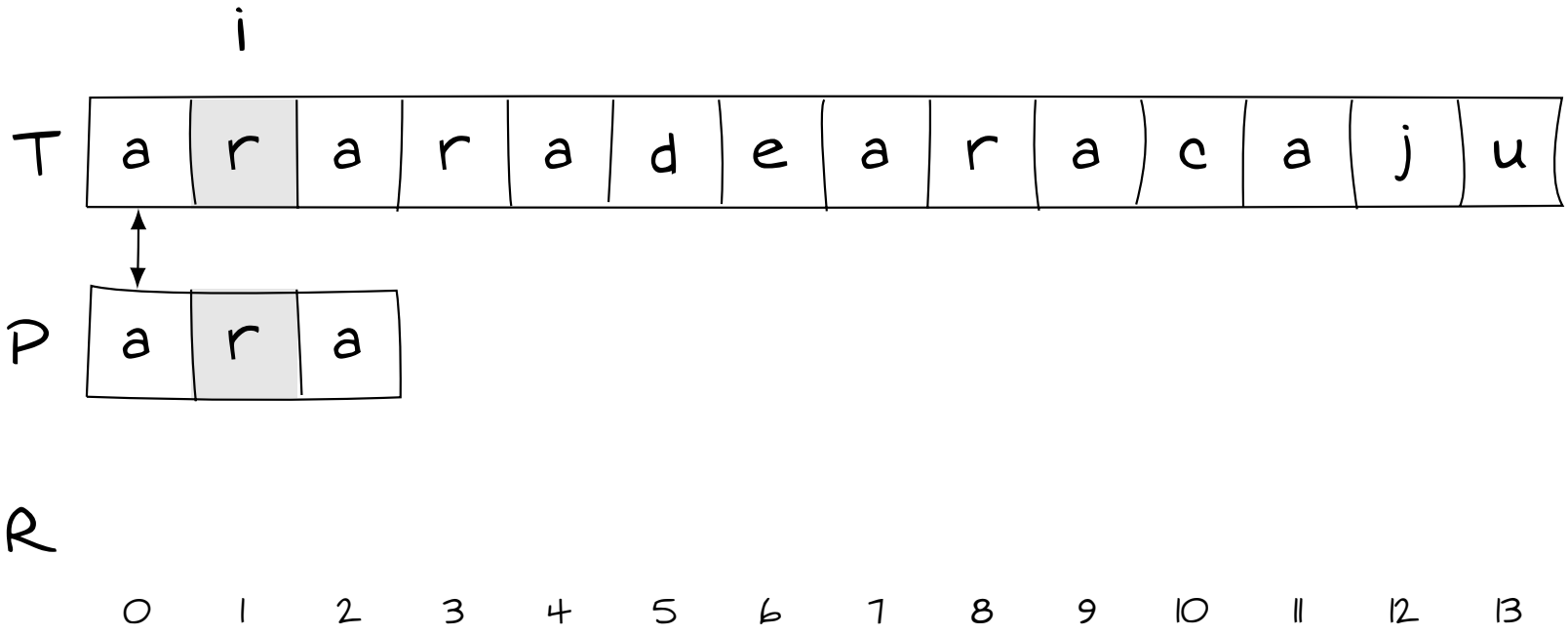




Busca em cadeias

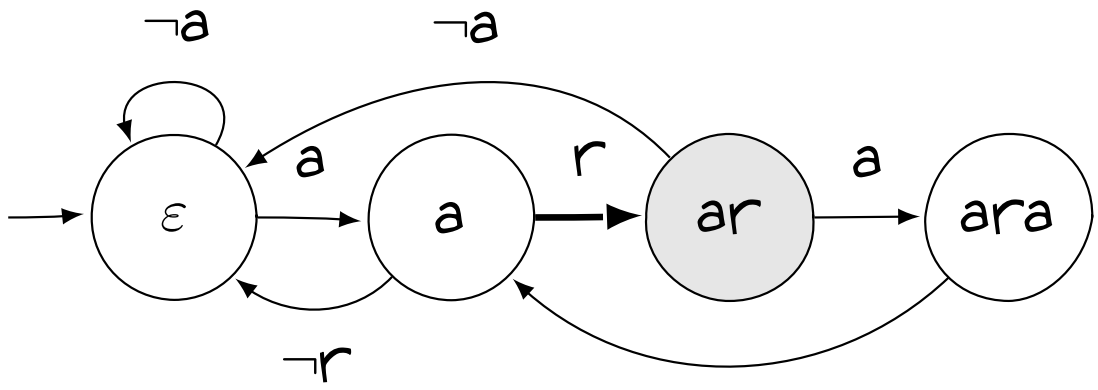
- Knuth-Morris-Pratt (KMP)
  - Buscando $P = ara$ em $T = araradearacaju$

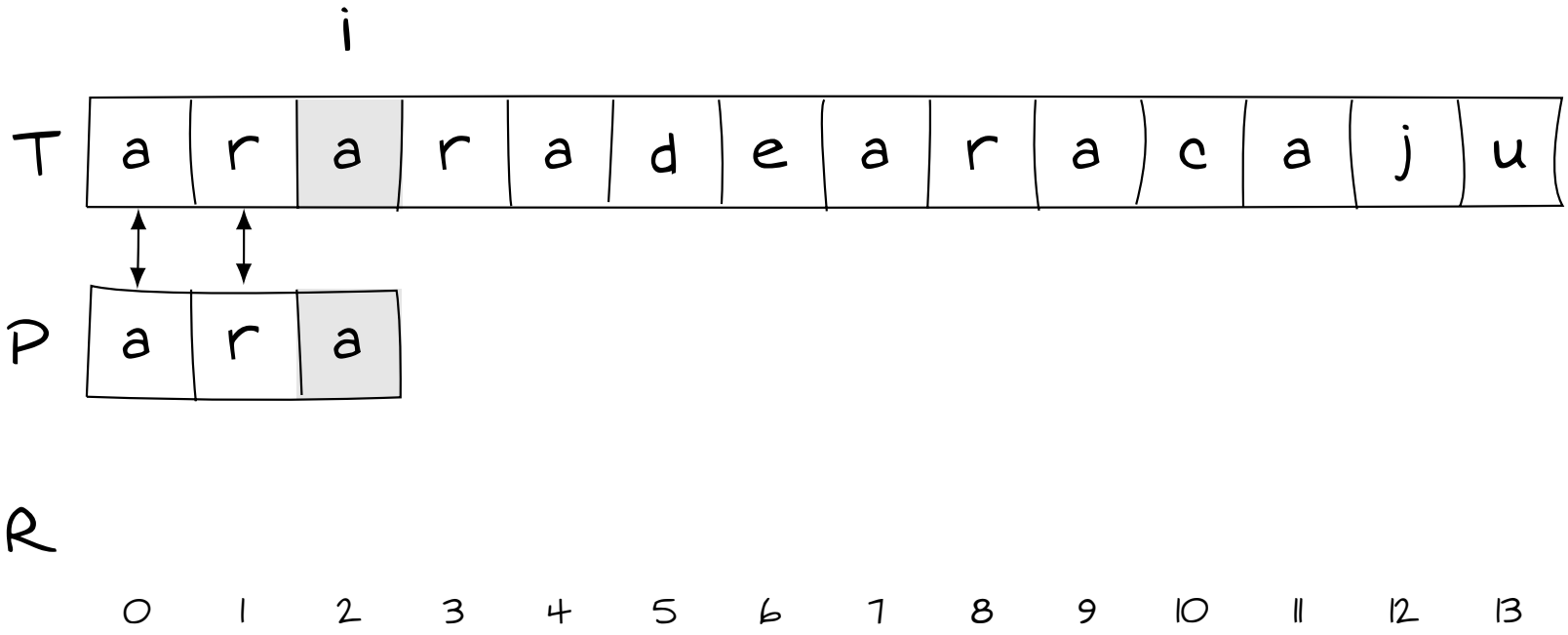




Busca em cadeias

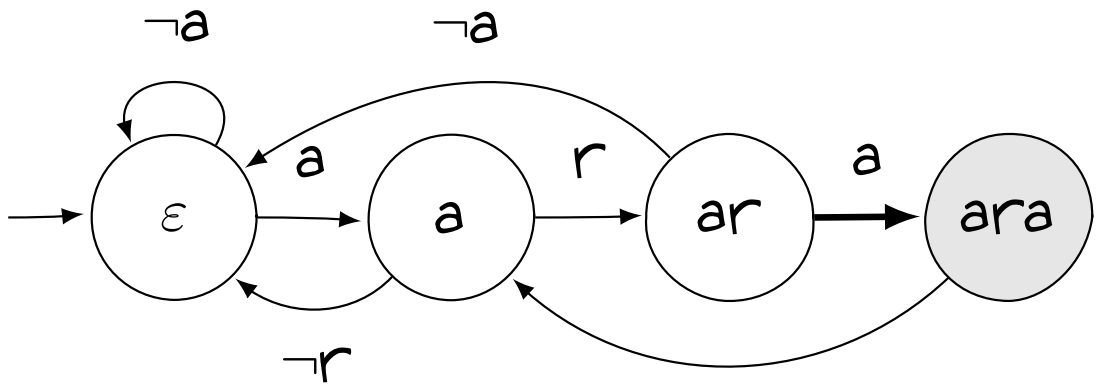
- Knuth-Morris-Pratt (KMP)
  - Buscando $P = ara$ em $T = araradearacaju$

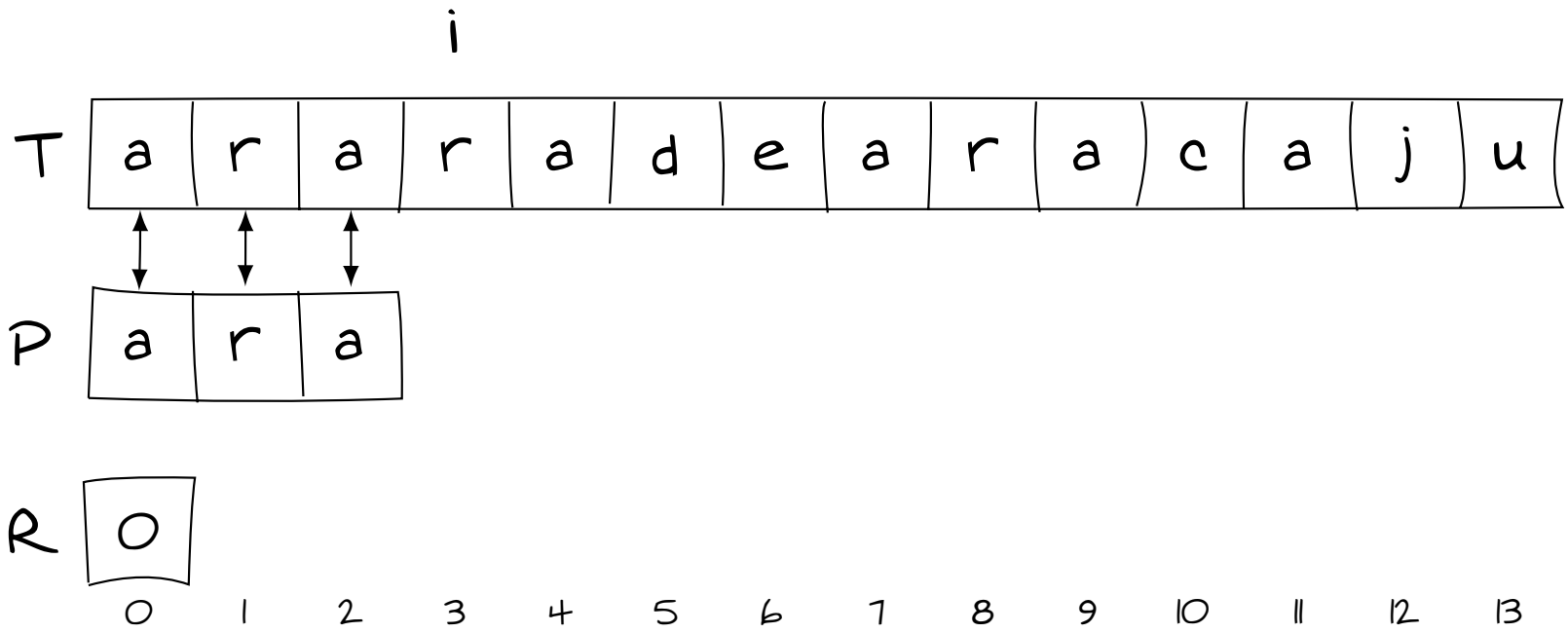




Busca em cadeias

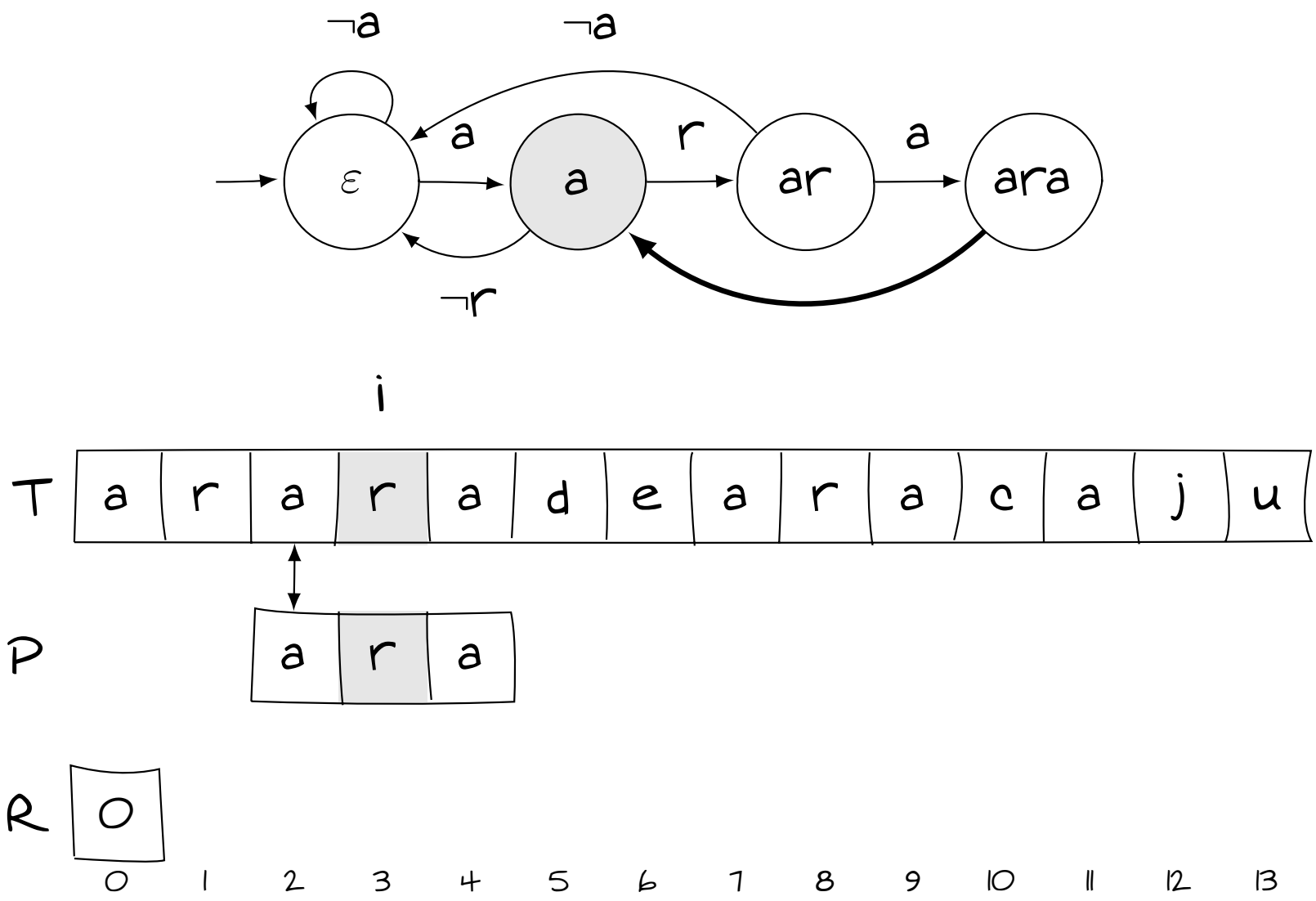
- Knuth-Morris-Pratt (KMP)
  - Buscando $P = ara$ em $T = araradearacaju$





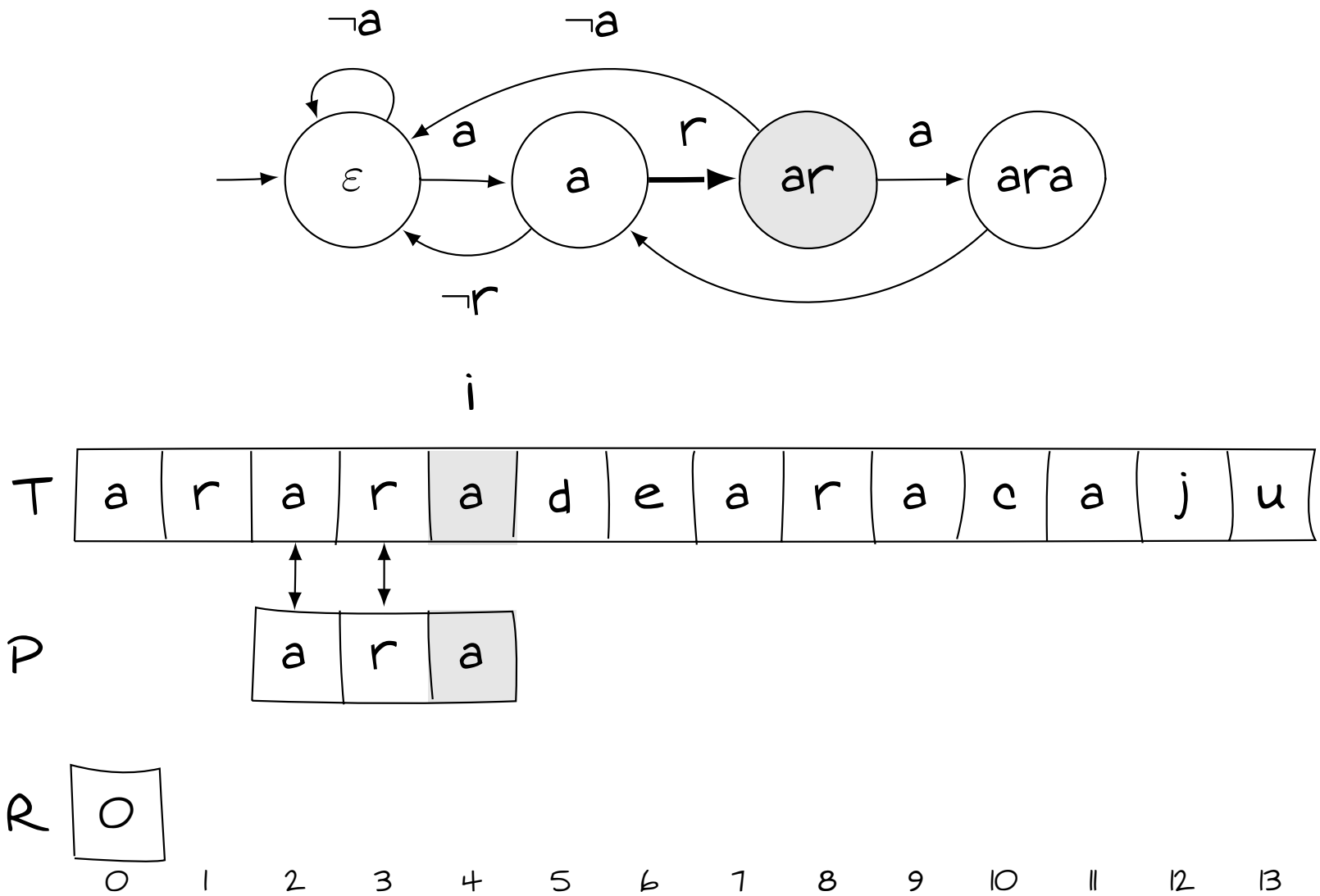
Busca em cadeias

- Knuth-Morris-Pratt (KMP)
  - Buscando $P = ara$ em $T = araradearacaju$



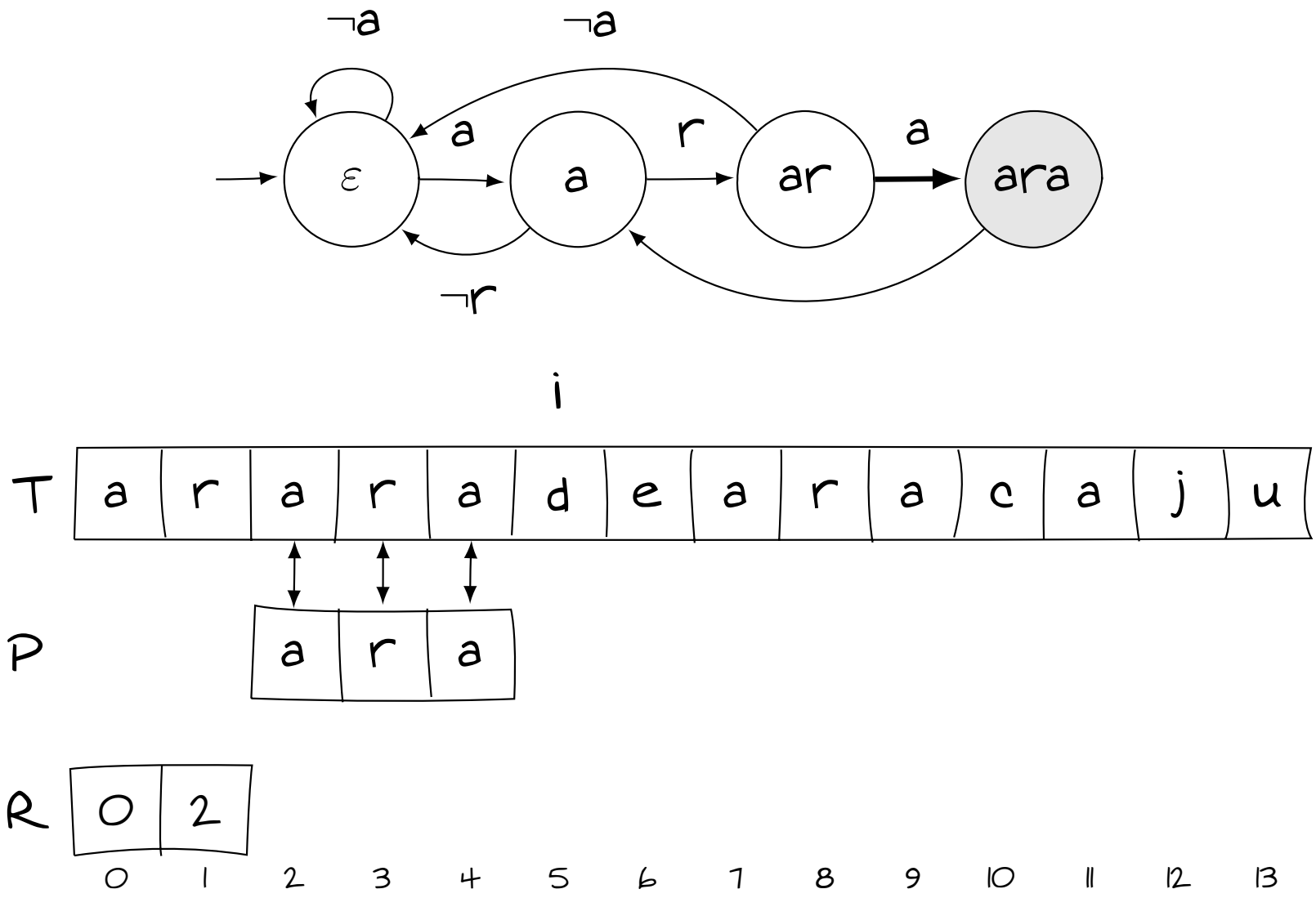
Busca em cadeias

- Knuth-Morris-Pratt (KMP)
  - Buscando $P = ara$ em $T = araradearacaju$



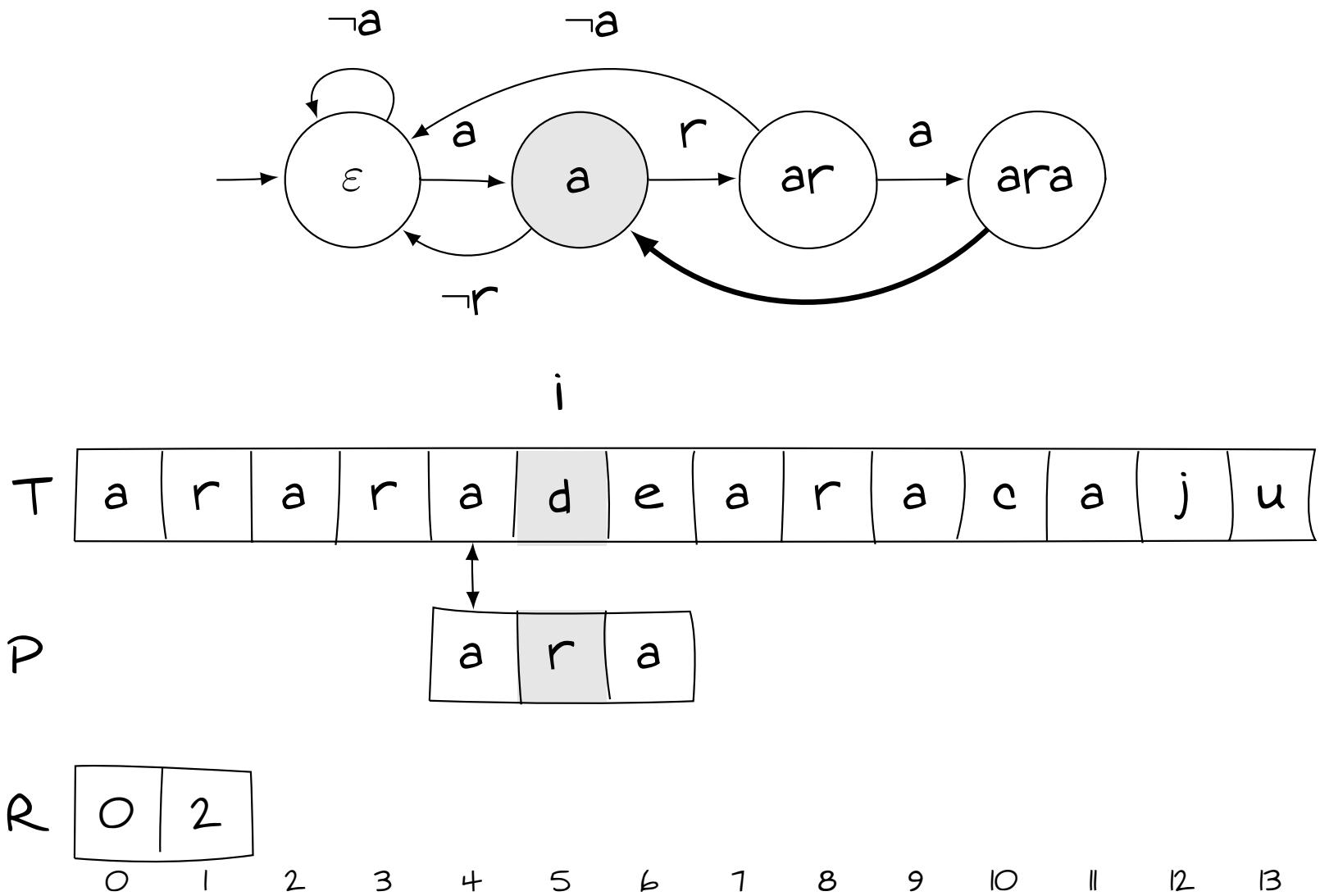
Busca em cadeias

- Knuth-Morris-Pratt (KMP)
  - Buscando $P = ara$ em $T = araradearacaju$



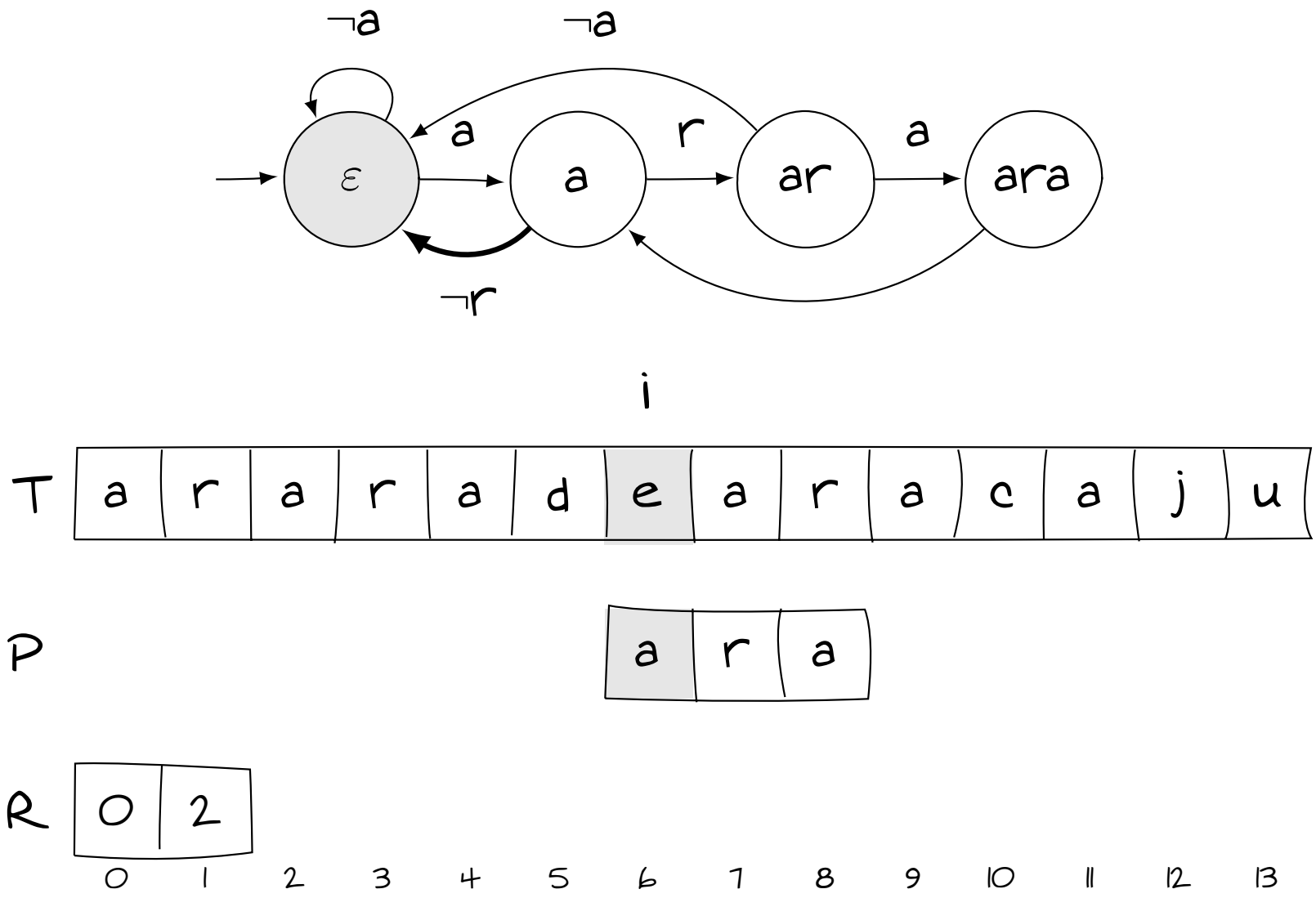
Busca em cadeias

- Knuth-Morris-Pratt (KMP)
  - Buscando $P = ara$ em $T = araradearacaju$



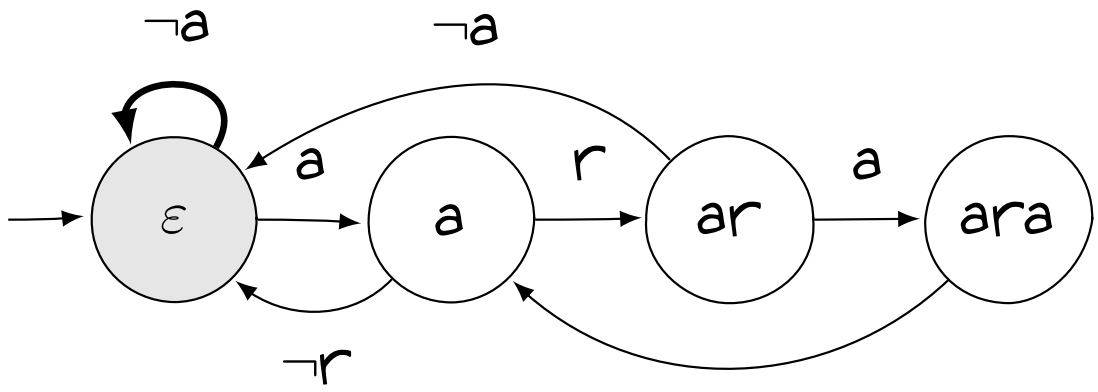
Busca em cadeias

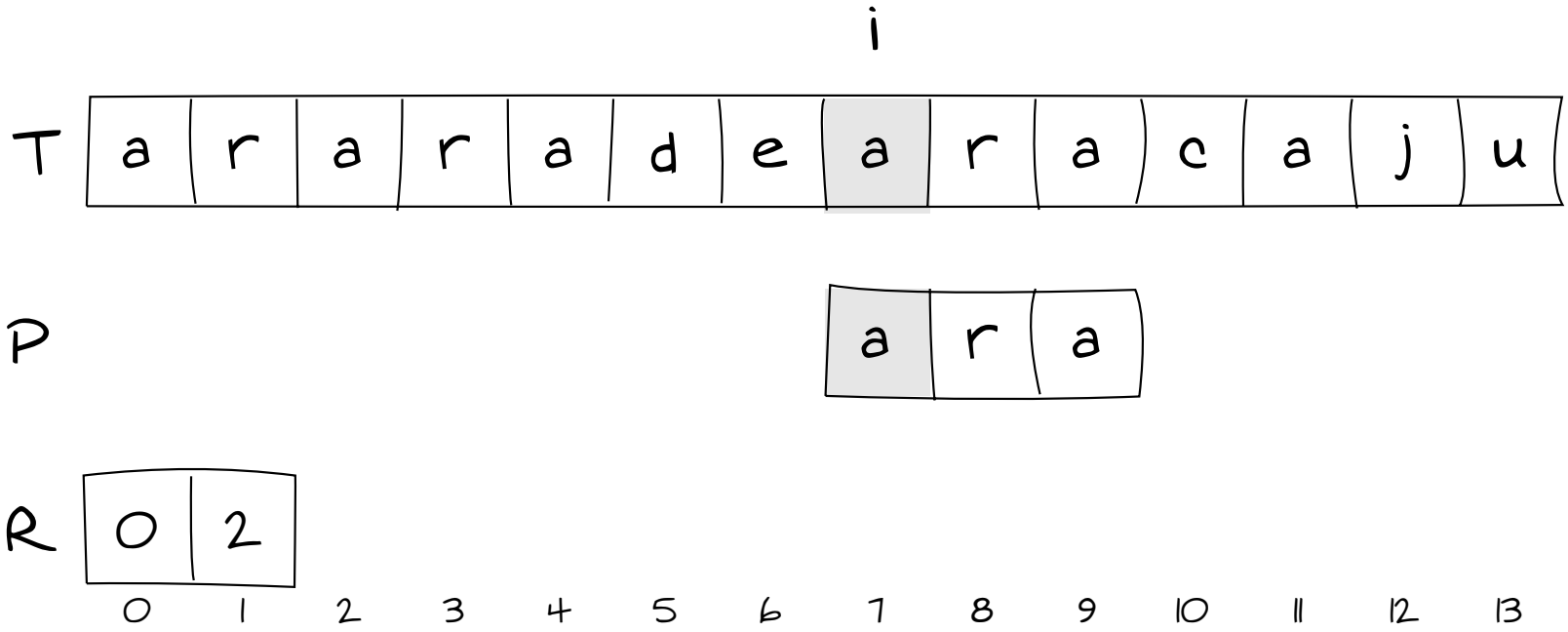
- Knuth-Morris-Pratt (KMP)
  - Buscando $P = ara$ em $T = araradearacaju$



Busca em cadeias

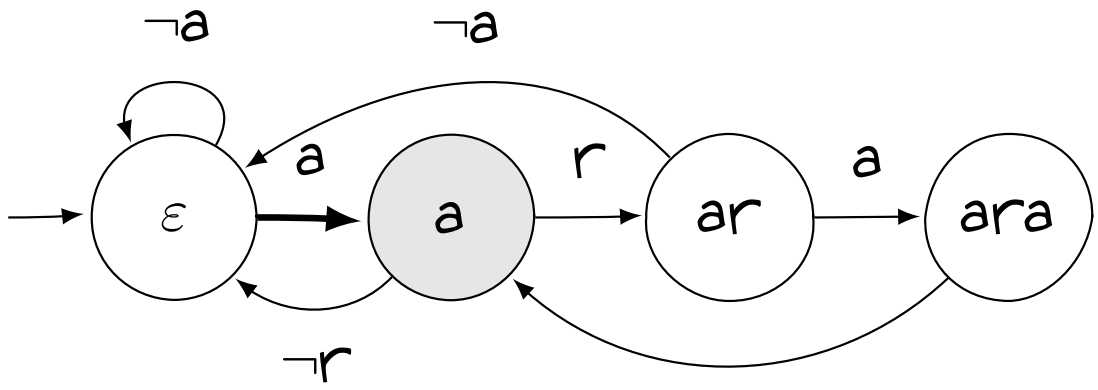
- Knuth-Morris-Pratt (KMP)
  - Buscando $P = ara$ em $T = araradearacaju$

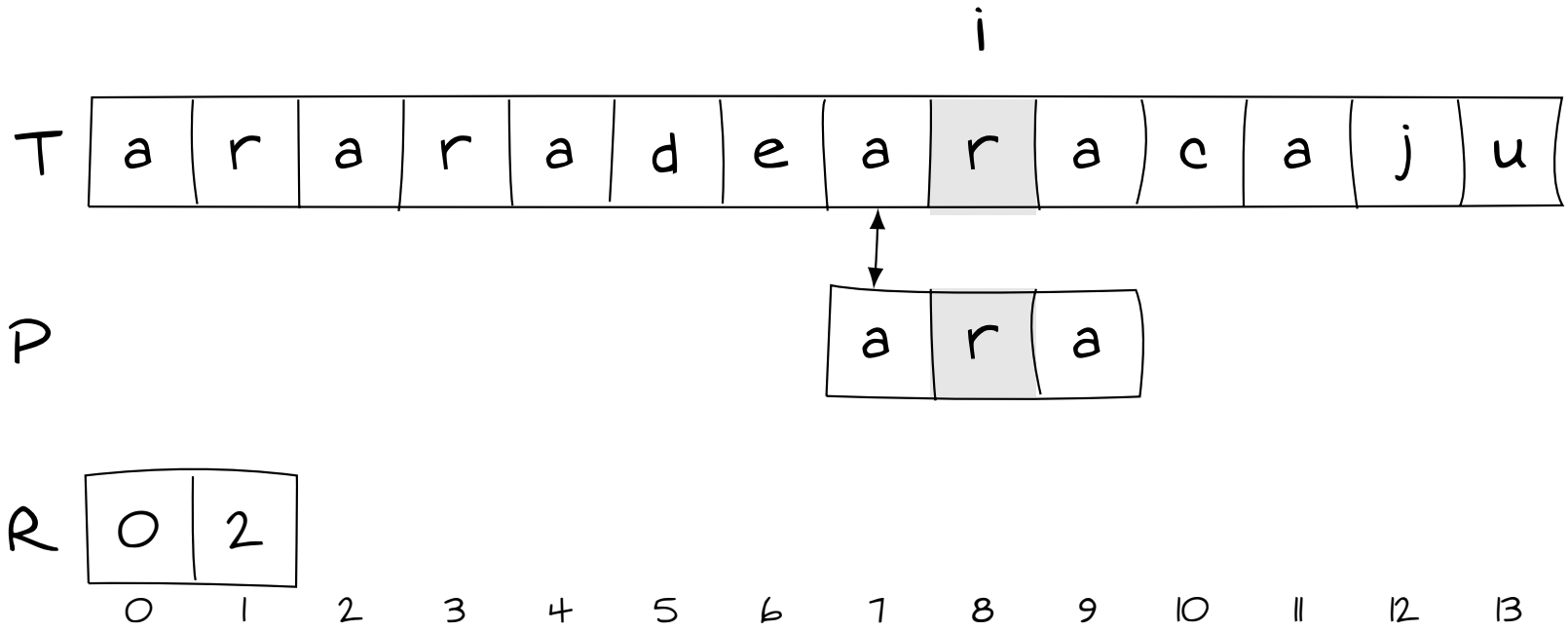




Busca em cadeias

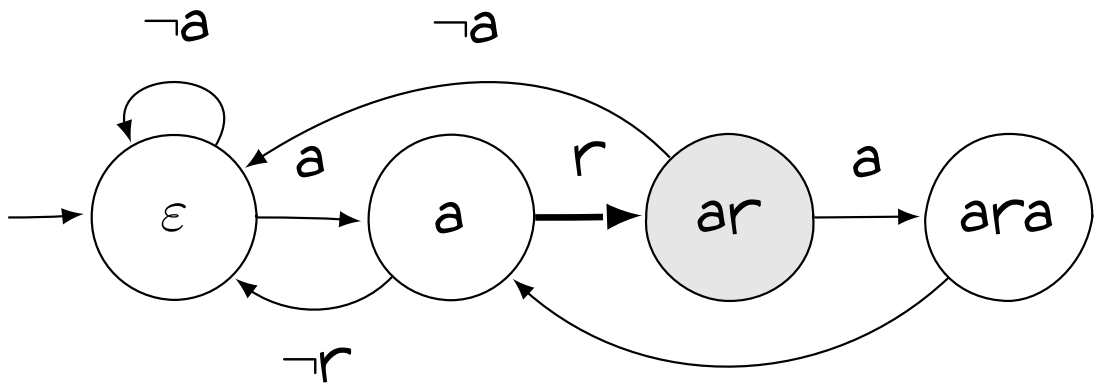
- Knuth-Morris-Pratt (KMP)
  - Buscando $P = ara$ em $T = araradearacaju$

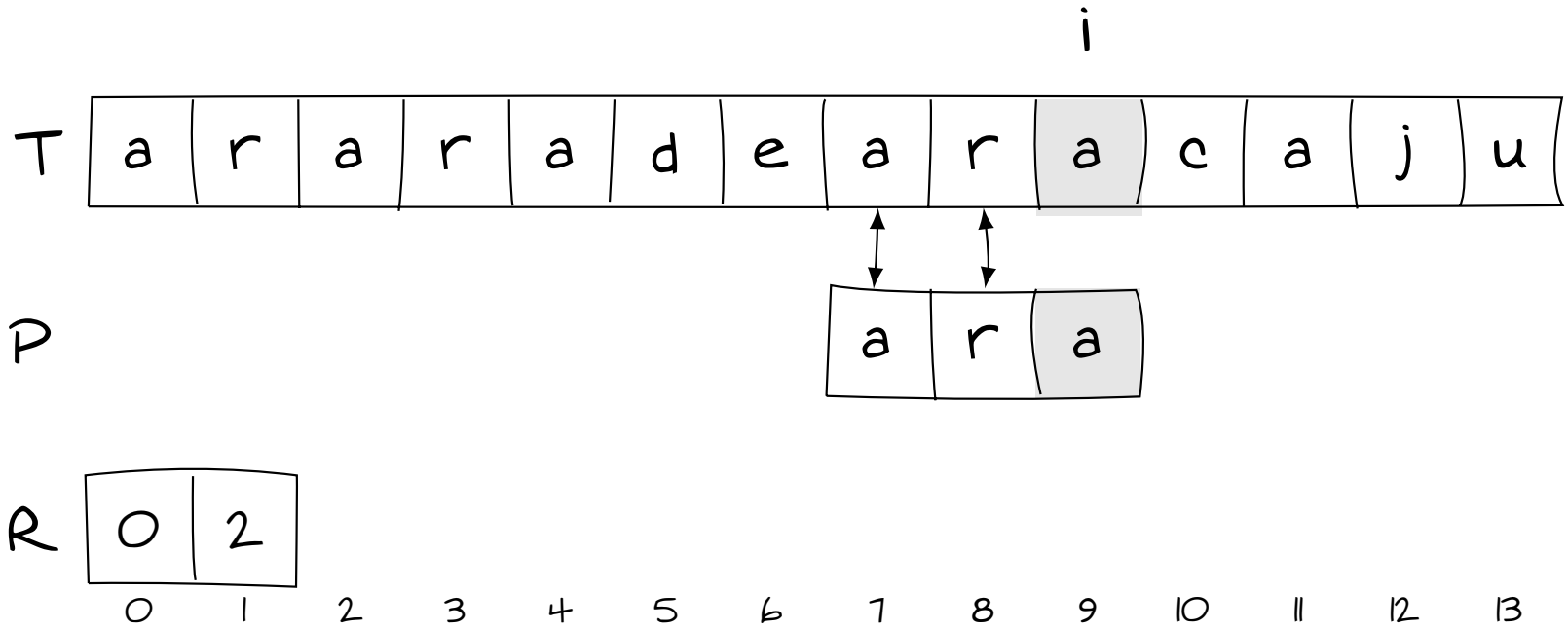




Busca em cadeias

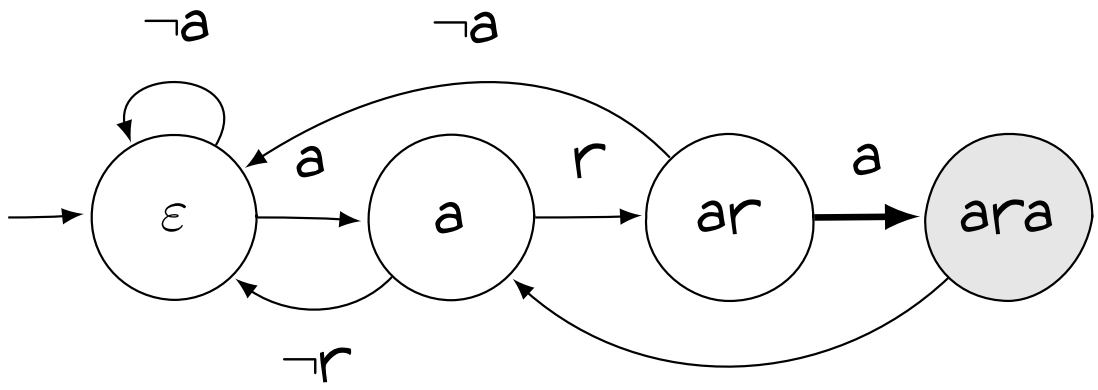
- Knuth-Morris-Pratt (KMP)
  - Buscando $P = ara$ em $T = araradearacaju$

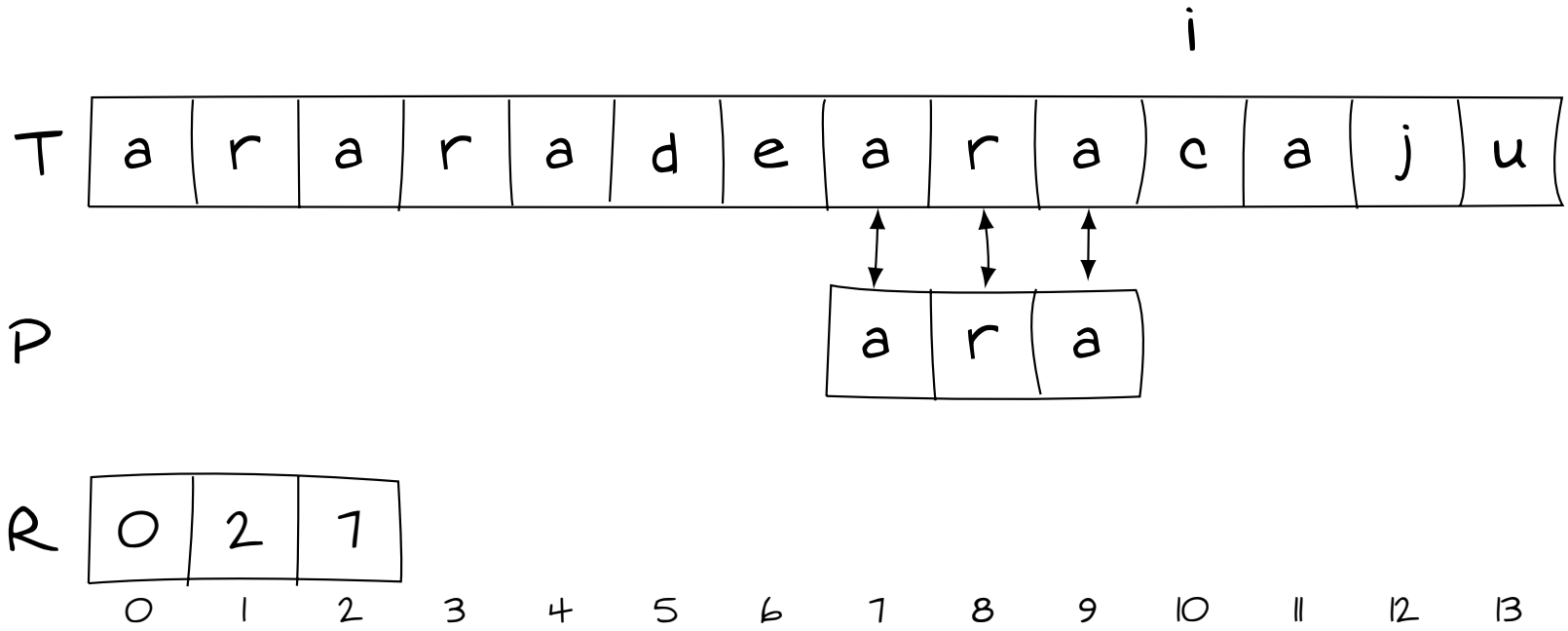




Busca em cadeias

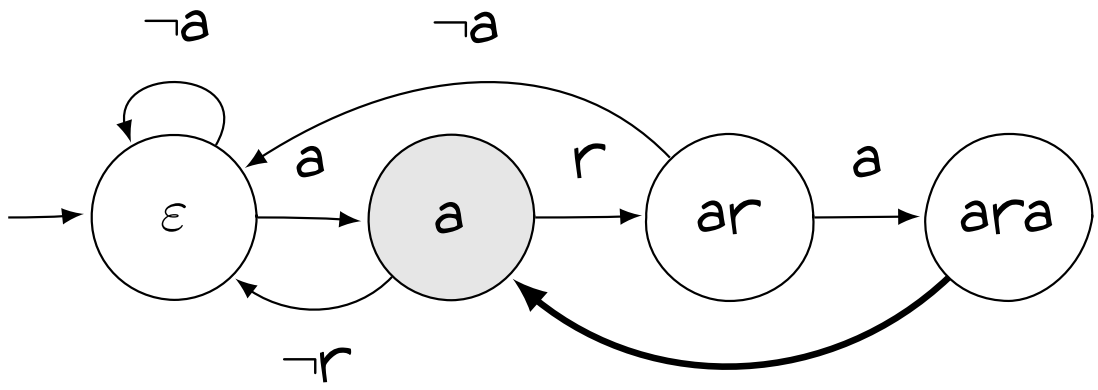
- Knuth-Morris-Pratt (KMP)
  - Buscando $P = ara$ em $T = araradearacaju$

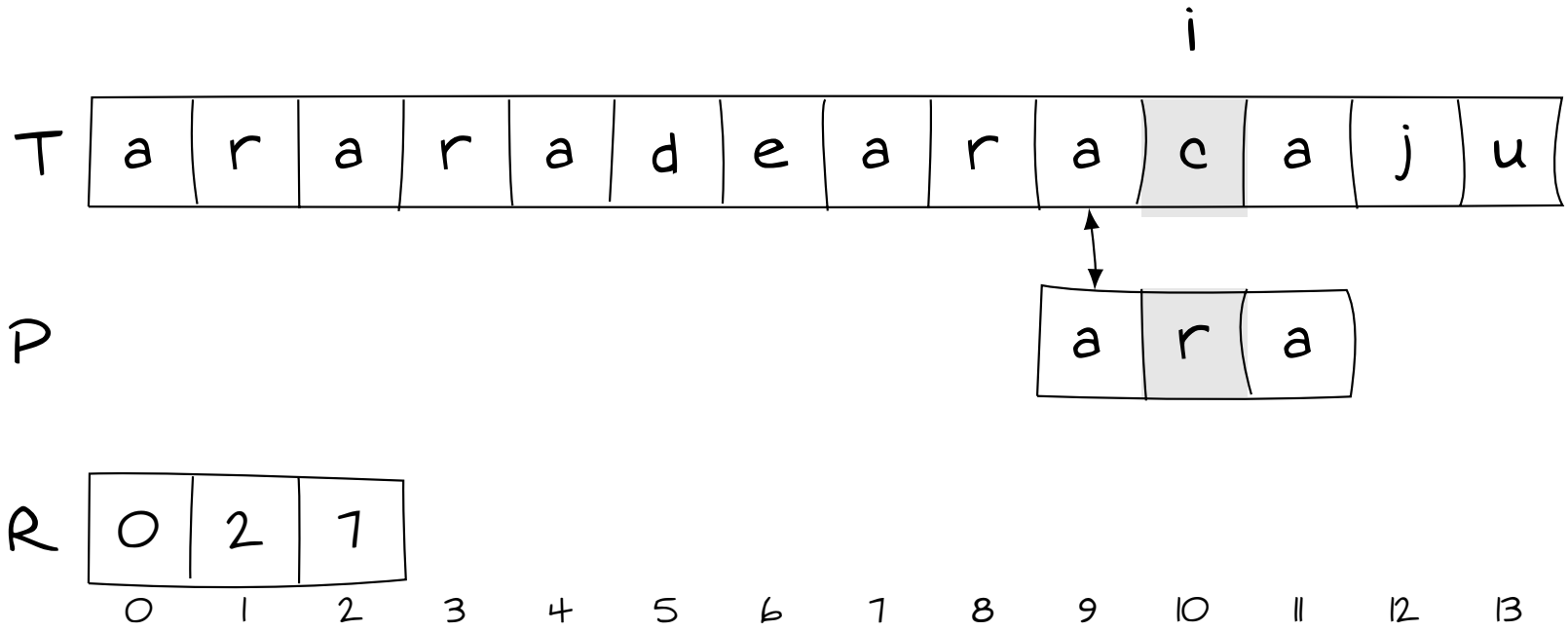




Busca em cadeias

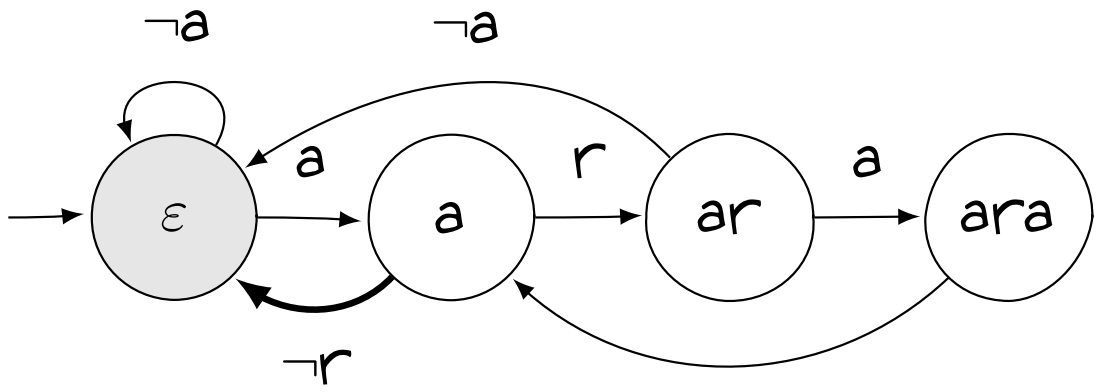
- Knuth-Morris-Pratt (KMP)
  - Buscando $P = ara$ em $T = araradearacaju$

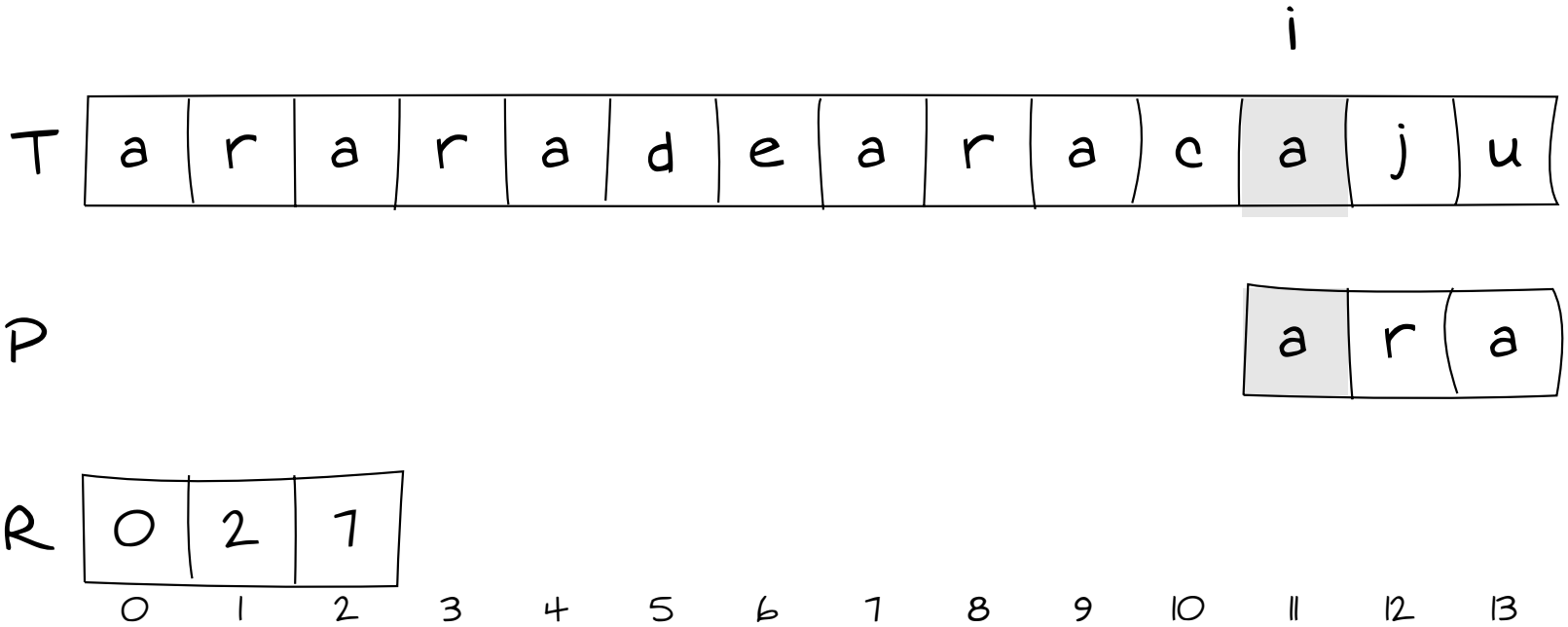




Busca em cadeias

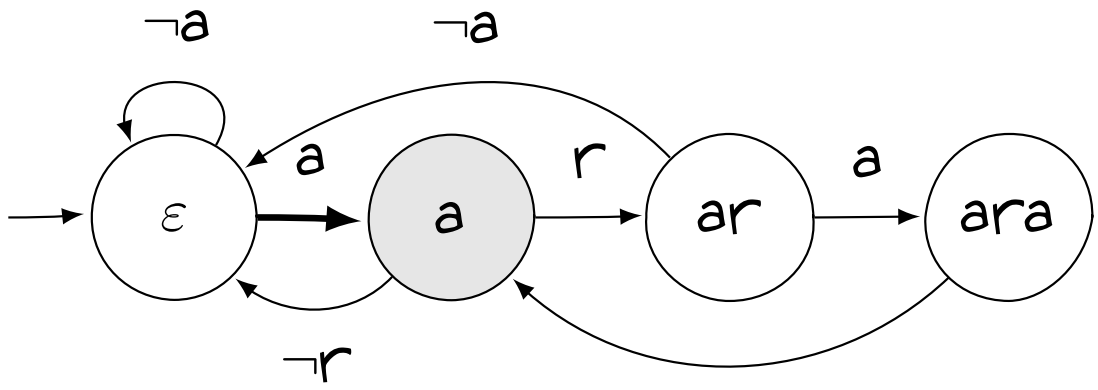
- Knuth-Morris-Pratt (KMP)
  - Buscando $P = ara$ em $T = araradearacaju$

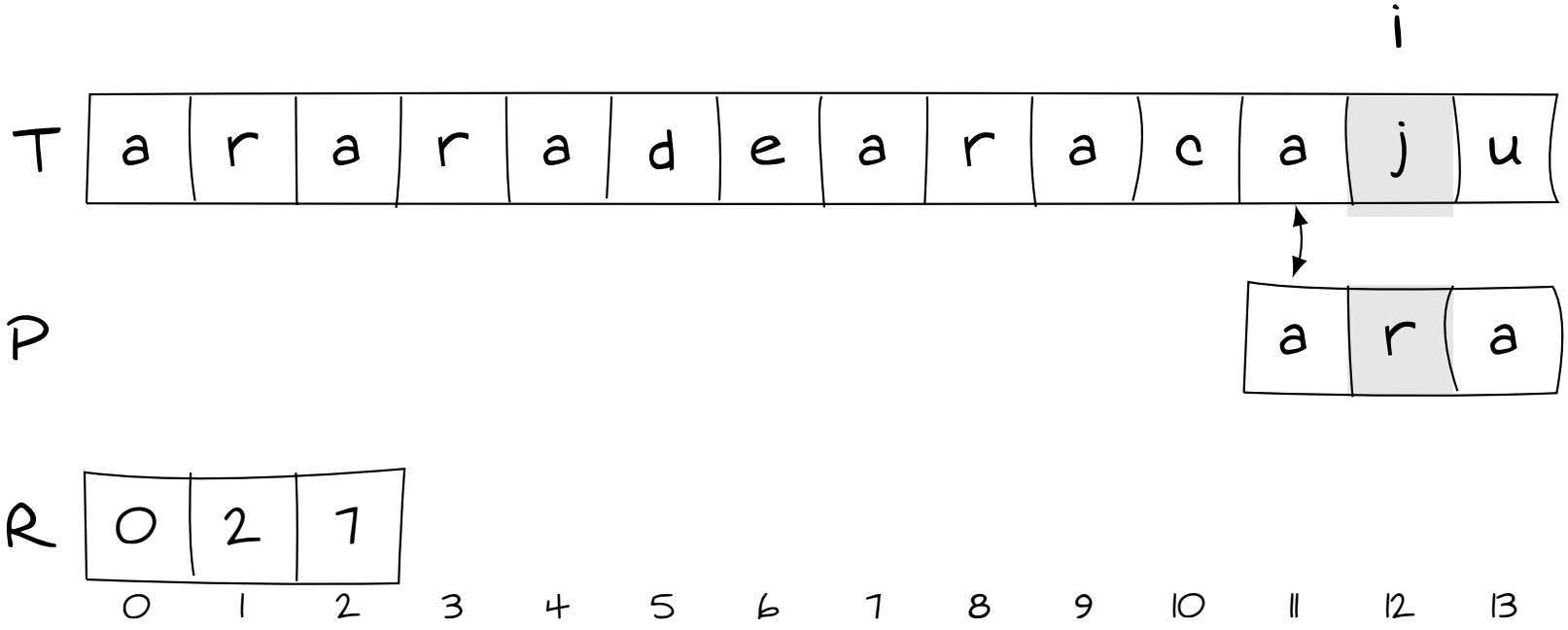




Busca em cadeias

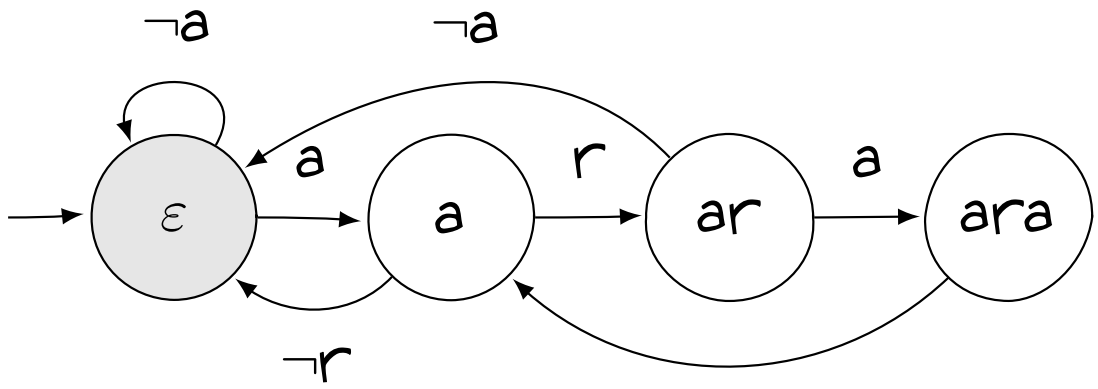
- Knuth-Morris-Pratt (KMP)
  - Buscando $P = ara$ em $T = araradearacaju$

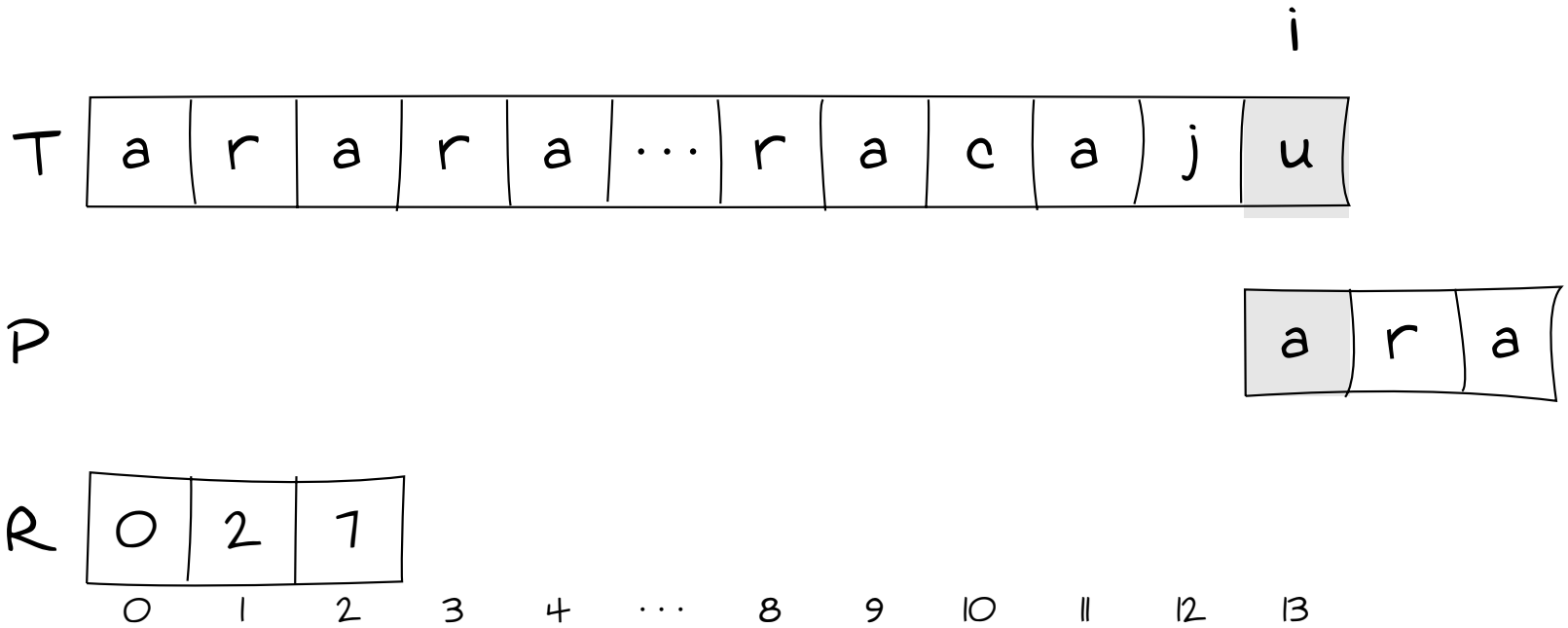




Busca em cadeias

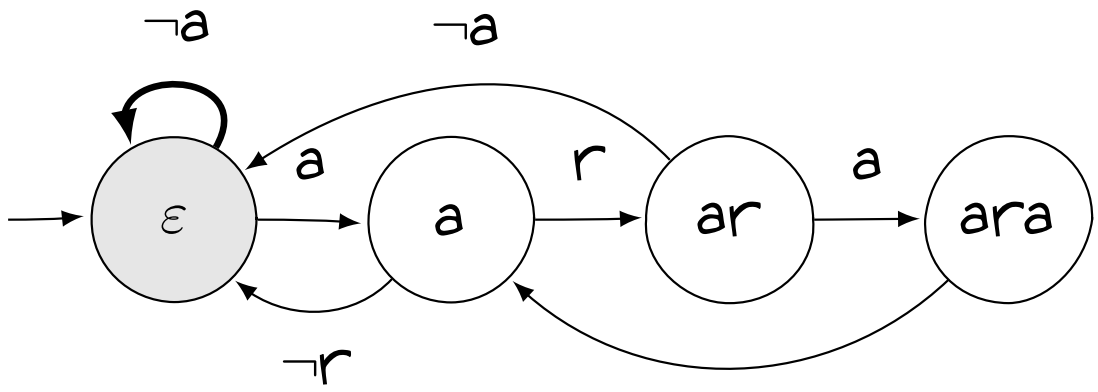
- Knuth-Morris-Pratt (KMP)
  - Buscando $P = ara$ em $T = araradearacaju$

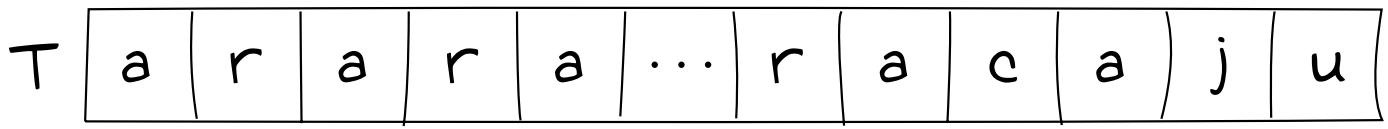


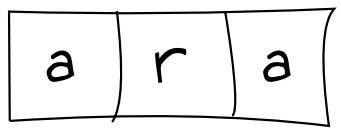


Busca em cadeias

- Knuth-Morris-Pratt (KMP)
  - Buscando $P = ara$ em $T = araradearacaju$







Busca em cadeias

- Knuth-Morris-Pratt (KMP)
  - Análise de complexidade
    - Espaço $O(n+m)$
    - Tempo $\Theta(m) + O(n+m) = O(n+m)$

Exemplo

- Aplique os algoritmos de busca em cadeias para encontrar o padrão 111000 na sequência binária 10111000110111100010101100011100001101101111
  - Execute passo a passo a busca na cadeia
  - Descreva seu princípio de funcionamento e as vantagens com relação aos algoritmos já vistos

Exercício

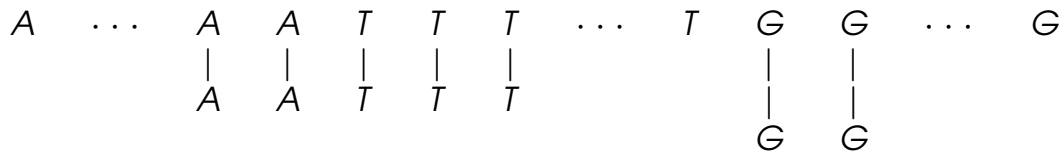
- A empresa de biotecnologia Poxim Tech está desenvolvendo um sistema de diagnóstico para doenças genéticas, comparando a sequência de DNA com genes conhecidos
  - A sequência de DNA é composta somente pelos símbolos *A*, *C*, *G* e *T* para codificação dos genes
  - Uma doença genética possui até 10 genes associados, cada um deles com sequências de tamanho entre 100 até 1000, denotados por letras maiúsculas e números entre 4 e 8 caracteres
  - Para tratar os efeitos da mutação nos genes que alteram sua codificação, é feita a busca por combinações que possuam o tamanho mínimo de subcadeia, com pelo menos 90% de compatibilidade total para manifestação da doença
  - No diagnóstico será calculada a probabilidade de manifestação da doença, de acordo com a quantidade de genes detectados

Exercício

- Diagnóstico da doença CRTLF4 com genes *AATTTGGCCC* e *GGGGGGGGGG*
  - DNA: *AAAAAAAAAATTTTTTTTTTGGGGGGGGGG*
  - Tamanho da subcadeia: 3

Exercício

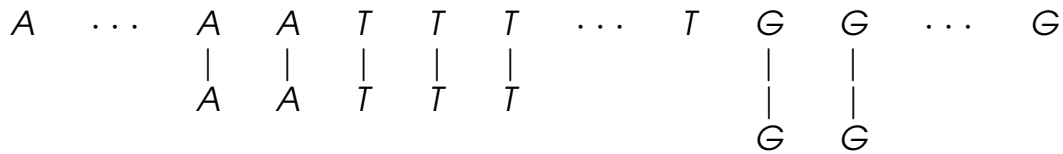
- Diagnóstico da doença CRTLF4 com genes *AATTTGGCCC* e *GGGGGGGGGG*
  - DNA: *AAAAAAAAAATTTTTTTTTTGGGGGGGGGG*
  - Tamanho da subcadeia: 3



AATTTGGCCC : 5 *combinações* = 50%

Exercício

- Diagnóstico da doença CRTLF4 com genes *AATTTGGCCC* e *GGGGGGGGGG*
  - DNA: *AAAAAAAAAATTTTTTTTTTGGGGGGGGGG*
  - Tamanho da subcadeia: 3

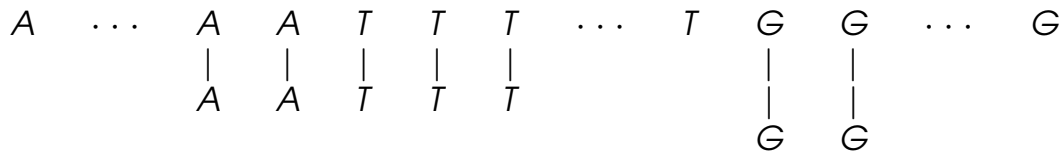


AATTTGGCCC : 5 *combinações* = 50%

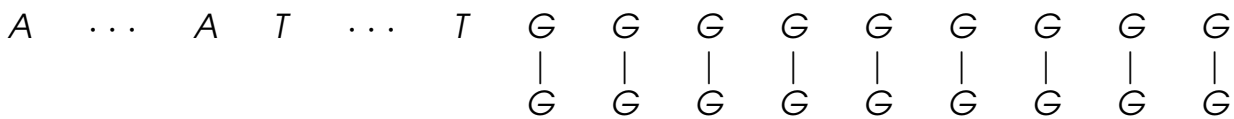
GGGGGGGGGG : 9 *combinações* = 90%

Exercício

- Diagnóstico da doença CRTLF4 com genes *AATTTGGCCC* e *GGGGGGGGGG*
  - DNA: *AAAAAAAAAATTTTTTTTTTGGGGGGGGGG*
  - Tamanho da subcadeia: 3



AATTTGGCCC : 5 *combinações* = 50%



GGGGGGGGGG : 9 *combinações* = 90%

Chance de 50% de ocorrência da doença CRTLF4

Exercício

- Formato do arquivo de entrada
  - [#*Tamanho da subcadeia*]
  - $[B_0 \ldots B_{N-1}]$
  - [#*Número de doenças*]
  - $[Código_0]$ $[\#Genes_0]$ $[G_{0_0}] \ldots [G_{0_{i-1}}]$
  - $\vdots$
  - $[Código_{M-1}]$ $[\#Genes_{M-1}]$ $[G_{M-1_0}] \ldots [G_{M-1_{j-1}}]$

```
1 3
2 AAAATTTCGTTAAATTTGAACATAGGGATA
3 4
4 ABCDE 3 AAA AAT AAAG
5 XY1WZ2AB 1 TTTTTTGGGG
6 H1N1 4 ACTG AACCGGTT AATAAT AAAAAAAAGA
7 HUEBR 1 CATAGGGATT
```

Exercício

- Formato do arquivo de saída
  - É feita a ordenação estável em ordem decrescente dos resultados, utilizando como critério de ordenação a probabilidade de ocorrência da doença e fazendo o arredondamento dos percentuais para fins de comparação e impressão

```
1 XY1WZ2AB->100%
2 HUEBR->100%
3 ABCDE->67%
4 H1N1->25%
```