Projeto e Análise de Algoritmos

Busca em cadeias (KMP)

Bruno Prado

Departamento de Computação / UFS

Introdução

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

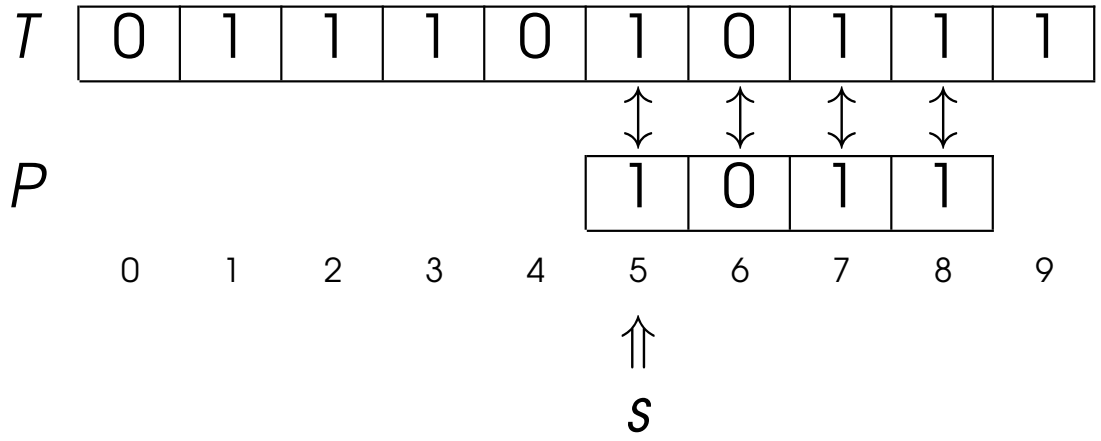
b	o	m	...	o	b	a
0	1	2	...	n - 3	n - 2	n - 1

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

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

Introdução

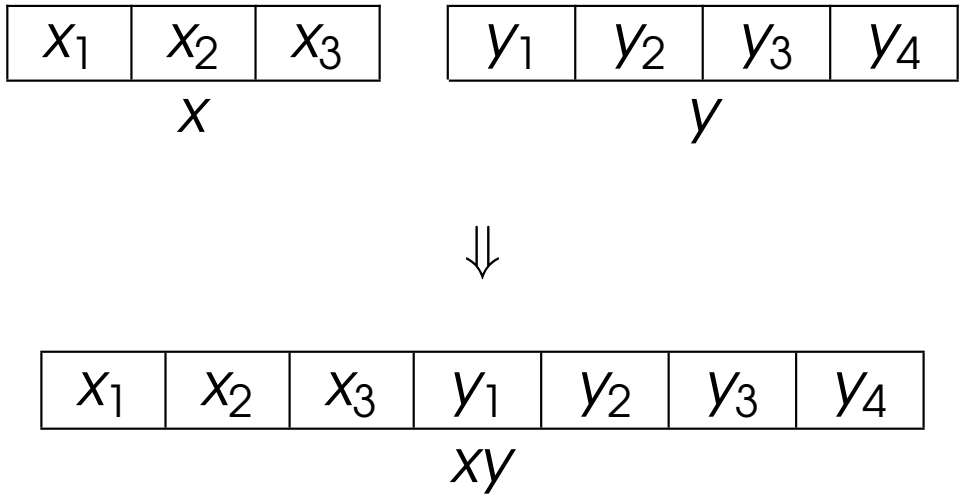
- O que é uma busca em cadeia?
  - É o processo para encontrar todas as ocorrências de um padrão em uma cadeia T que possui n símbolos
  - Para a busca é utilizada a cadeia de padrão P com quantidade de símbolos $m \leq n$
  - As cadeias P e T utilizam um alfabeto finito $\sum$



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

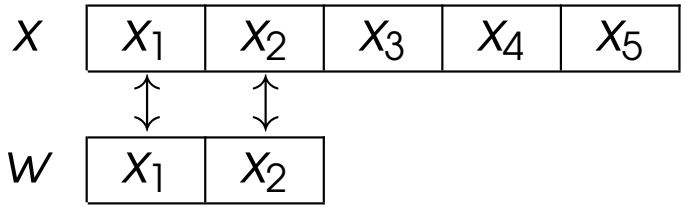
Introdução

- Notação e terminologia
  - É definido por $\sum^*$ todos os conjuntos de cadeias de tamanho finito que podem ser construídas do alfabeto finito $\sum$
  - 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 $|x| + |y|$

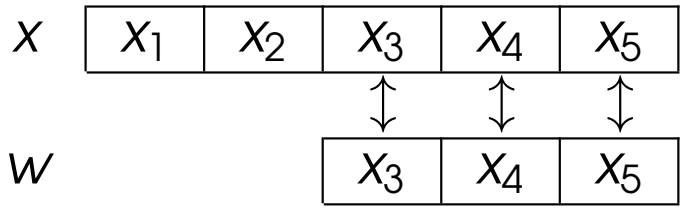


Introdução

- Notação e terminologia
  - Prefixo
    - A cadeia w é um prefixo da cadeia x se $x = wy$, para alguma cadeia $y \in \sum^*$
    - Denotado por $w \sqsubset x$, com $|w| \leq |x|$

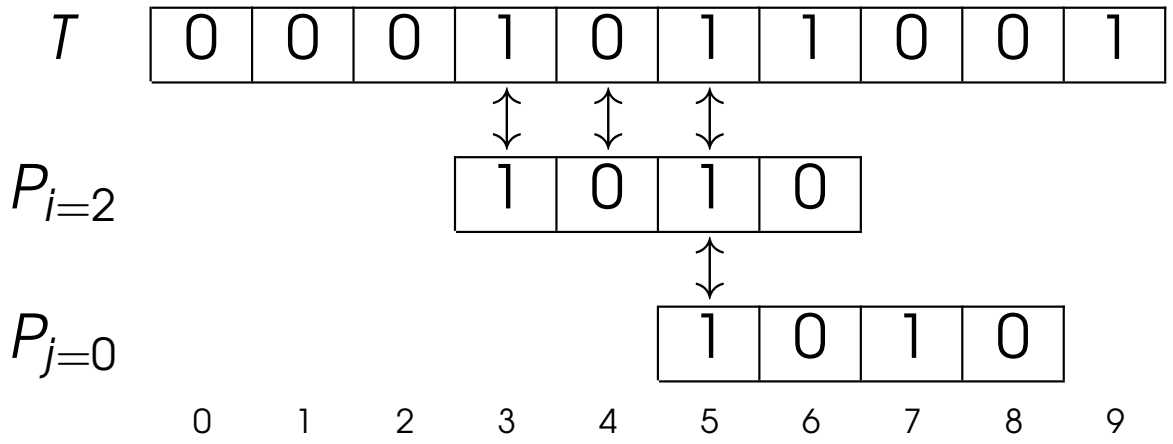


  - Sufixo
    - A cadeia w é sufixo da cadeia x se $x = yw$, para alguma cadeia $y \in \sum^*$
    - Denotado por $w \sqsupset x$, com $|w| \leq |x|$



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
  - Seu princípio de funcionamento está baseado em autômatos finitos e em sua função 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 π



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

Busca em cadeias

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

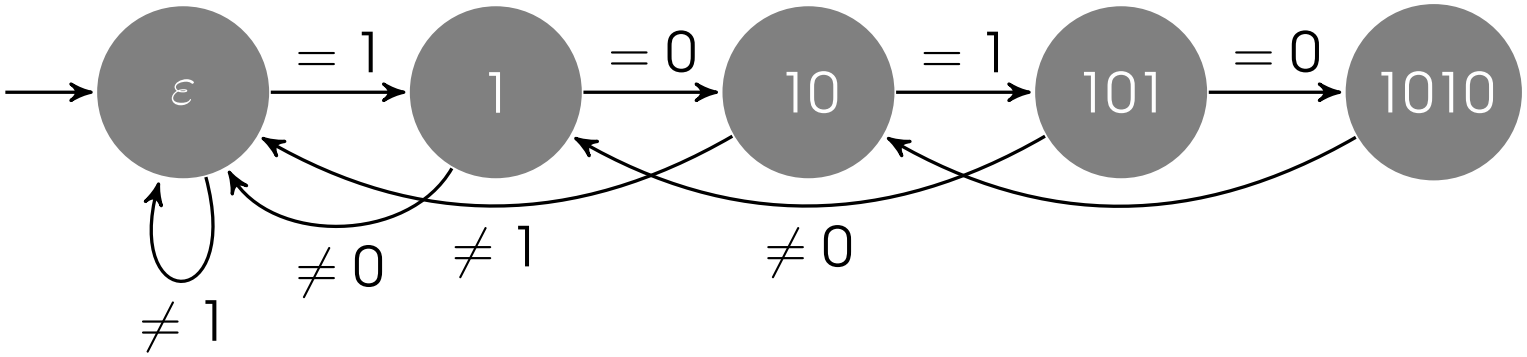
	Prefixo								
P_3	1	0	1	0					$\pi[3] = 1$
P_2		1	0	1	0				$\pi[2] = 0$
P_1			1	0	1	0			$\pi[1] = -1$
P_0				1	0	1	0		$\pi[0] = -1$
P_{-1}				ε	1	0	1	0	

$P[i]$	1	0	1	0
$\pi[i]$	-1	-1	0	1
-1	0	1	2	3

O armazenamento do comprimento do maior prefixo, que também é sufixo dele mesmo, evita que o algoritmo precise retroceder na comparação dos símbolos

Busca em cadeias

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



$P[i]$	1	0	1	0
$\pi[i]$	-1	-1	0	1
-1	0	1	2	3

Busca em cadeias

- Knuth-Morris-Pratt (KMP)
  - Implementação em C

```
void calcular_tabela(int tab(), char P()) {
    unsigned int i, m = strlen(P);
    int j = -1;
    inicializar(tab, m);
    for(i = 1; i < m; i++) {
        while(j >= 0 && P(j + 1) != P(i))
            j = tab(j);
        if(P(j + 1) == P(i))
            j++;
        tab(i) = j;
    }
}
```

Busca em cadeias

- Knuth-Morris-Pratt (KMP)
  - Implementação em C

```
void busca_kmp(int pos(), int tab(), char T(), char P()) {
    unsigned int i, n = strlen(T);
    unsigned int m = strlen(P);
    int j = -1;
    calcular_tabela(tab, P);
    for(i = 0; i < n; i++) {
        while(j >= 0 && P(j + 1) != T(i)) j = tab(j);
        if(P(j + 1) == T(i)) j++;
        if(j == m - 1) {
            inserir(pos, i - m + 1);
            j = tab(j);
        }
    }
}
```

Busca em cadeias

- Knuth-Morris-Pratt (KMP)
  - Execução do algoritmo de cálculo da tabela

	Prefixo						
P_2	a	r	$\underline{\text{a}}$				$\pi[2] = 0$
P_1		a	r	a			$\pi[1] = -1$
P_0			a	r	a		$\pi[0] = -1$
P_{-1}			ε	a	r	a	

$P[i]$	a	r	a
$\pi[i]$	-1	-1	0
-1	0	1	2

Busca em cadeias

- Knuth-Morris-Pratt (KMP)
  - Execução do algoritmo de busca

	i													
T	a	r	a	r	a	d	e	a	r	a	c	a	j	u

j			
P	a	r	a
π	-1	-1	0

pos														
-1	0	1	2	3	4	5	6	7	8	9	10	11	12	13

Busca em cadeias

- Knuth-Morris-Pratt (KMP)
  - Execução do algoritmo de busca

		i												
T	a	r	a	r	a	d	e	a	r	a	c	a	j	u

		j	
P	a	r	a
π	-1	-1	0

pos														
-1	0	1	2	3	4	5	6	7	8	9	10	11	12	13

Busca em cadeias

- Knuth-Morris-Pratt (KMP)
  - Execução do algoritmo de busca

			i											
T	a	r	a	r	a	d	e	a	r	a	c	a	j	u

		j	
P	a	r	a
π	-1	-1	0

pos														
-1	0	1	2	3	4	5	6	7	8	9	10	11	12	13

Busca em cadeias

- Knuth-Morris-Pratt (KMP)
  - Execução do algoritmo de busca

				i										
T	a	r	a	r	a	d	e	a	r	a	c	a	j	u

	j		
P	a	r	a
π	-1	-1	0

pos	0

-1	0	1	2	3	4	5	6	7	8	9	10	11	12	13

Busca em cadeias

- Knuth-Morris-Pratt (KMP)
  - Execução do algoritmo de busca

					i									
T	a	r	a	r	a	d	e	a	r	a	c	a	j	u

		j	
P	a	r	a
π	-1	-1	0

pos	0

-1	0	1	2	3	4	5	6	7	8	9	10	11	12	13

Busca em cadeias

- Knuth-Morris-Pratt (KMP)
  - Execução do algoritmo de busca

						i								
T	a	r	a	r	a	d	e	a	r	a	c	a	j	u

	j		
P	a	r	a
π	-1	-1	0

pos	0	2

-1	0	1	2	3	4	5	6	7	8	9	10	11	12	13

Busca em cadeias

- Knuth-Morris-Pratt (KMP)
  - Execução do algoritmo de busca

						i								
T	a	r	a	r	a	d	e	a	r	a	c	a	j	u

j			
P	a	r	a
π	-1	-1	0

pos	0	2

-1	0	1	2	3	4	5	6	7	8	9	10	11	12	13

Busca em cadeias

- Knuth-Morris-Pratt (KMP)
  - Execução do algoritmo de busca

							i							
T	a	r	a	r	a	d	e	a	r	a	c	a	j	u

j			
P	a	r	a
π	-1	-1	0

pos	0	2

-1	0	1	2	3	4	5	6	7	8	9	10	11	12	13

Busca em cadeias

- Knuth-Morris-Pratt (KMP)
  - Execução do algoritmo de busca

								i						
T	a	r	a	r	a	d	e	a	r	a	c	a	j	u

j			
P	a	r	a
π	-1	-1	0

pos	0	2

-1	0	1	2	3	4	5	6	7	8	9	10	11	12	13

Busca em cadeias

- Knuth-Morris-Pratt (KMP)
  - Execução do algoritmo de busca

									i					
T	a	r	a	r	a	d	e	a	r	a	c	a	j	u

	j		
P	a	r	a
π	-1	-1	0

pos	0	2

-1	0	1	2	3	4	5	6	7	8	9	10	11	12	13

Busca em cadeias

- Knuth-Morris-Pratt (KMP)
  - Execução do algoritmo de busca

										i				
T	a	r	a	r	a	d	e	a	r	a	c	a	j	u

		j	
P	a	r	a
π	-1	-1	0

pos	0	2

-1	0	1	2	3	4	5	6	7	8	9	10	11	12	13

Busca em cadeias

- Knuth-Morris-Pratt (KMP)
  - Execução do algoritmo de busca

											i			
T	a	r	a	r	a	d	e	a	r	a	c	a	j	u

	j		
P	a	r	a
π	-1	-1	0

pos	0	2	7

-1	0	1	2	3	4	5	6	7	8	9	10	11	12	13

Busca em cadeias

- Knuth-Morris-Pratt (KMP)
  - Execução do algoritmo de busca

												i		
T	a	r	a	r	a	d	e	a	r	a	c	a	j	u
j														
P	a	r	a											
π	-1	-1	0											
pos	0	2	7											
	-1	0	1	2	3	4	5	6	7	8	9	10	11	12	13

Busca em cadeias

- Knuth-Morris-Pratt (KMP)
  - Execução do algoritmo de busca

													i	
T	a	r	a	r	a	d	e	a	r	a	c	a	j	u

	j		
P	a	r	a
π	-1	-1	0

pos	0	2	7

-1	0	1	2	3	4	5	6	7	8	9	10	11	12	13

Busca em cadeias

- Knuth-Morris-Pratt (KMP)
  - Execução do algoritmo de busca

														i
T	a	r	a	r	a	d	e	a	r	a	c	a	j	u

j			
P	a	r	a
π	-1	-1	0

pos	0	2	7

-1	0	1	2	3	4	5	6	7	8	9	10	11	12	13

Busca em cadeias

- Knuth-Morris-Pratt (KMP)
  - Análise de complexidade
    - Espaço $O(n - m + 1)$
    - 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 exclusivamente 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

$$
\begin{array}{ccccccccccccc}
A & \cdots & A & A & T & T & T & \cdots & T & G & G & \cdots & G \\
 & & | & | & | & | & | & & & | & | & & \\
 & & A & A & T & T & T & & & | & | & & \\
 & & & & & & & & & G & G & &
\end{array}
$$

$$AATTTGGCCC: \ 5 \ combinações = 50\%$$

Exercício

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

$$\begin{array}{ccccccccccccc} A & \cdots & A & A & T & T & T & \cdots & T & G & G & \cdots & G \\ & & | & | & | & | & | & & & | & | & & \\ & & A & A & T & T & T & & & | & | & & \\ & & & & & & & & & G & G & & \end{array}$$

$$AATTTGGCCC: \ 5 \ combinações = 50\%$$

$$\begin{array}{ccccccccccccccc} A & \cdots & A & T & \cdots & T & G & G & G & G & G & G & G & G & G \\ & & & & & & | & | & | & | & | & | & | & | & | \\ & & & & & & G & G & G & G & G & G & G & G & G \end{array}$$

$$GGGGGGGGGG: \ 9 \ combinações = 90\%$$

Exercício

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

$$\begin{array}{ccccccccccccc} A & \cdots & A & A & T & T & T & \cdots & T & G & G & \cdots & G \\ & & | & | & | & | & | & & & | & | & & \\ & & A & A & T & T & T & & & | & | & & \\ & & & & & & & & & G & G & & \end{array}$$

$$AATTTGGCCC: \ 5 \ combinações = 50\%$$

$$\begin{array}{ccccccccccccccc} A & \cdots & A & T & \cdots & T & G & G & G & G & G & G & G & G & G \\ & & & & & & | & | & | & | & | & | & | & | & | \\ & & & & & & G & G & G & G & G & G & G & G & G \end{array}$$

$$GGGGGGGGGG: \ 9 \ combinações = 90\%$$

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*]
  - $[Código_0]$ $[\#Genes_0]$ $[G_{0_0}] \ldots [G_{0_{i-1}}]$
  - $\vdots$
  - $[Código_{M-1}]$ $[\#Genes_{M-1}]$ $[G_{M-1_0}] \ldots [G_{M-1_{j-1}}]$

```
3
AAAATTTCGTTAAATTTGAACATAGGGATA
4
ABCDE 3 AAA AAT AAAG
XY1WZ2AB 1 TTTTTTGGGG
H1N1 4 ACTG AACCGGTT AATAAT AAAAAAAAAGA
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

```
XY1WZ2AB: 100%
HUEBR: 100%
ABCDE: 67%
H1N1: 25%
```