

Sobre Matching en Hileras

Victor David Coto Solano

Diego Quirós Artiñano

Derek Rojas Mendoza

Universidad Nacional de Costa Rica

EIF-203: Estructuras Discretas Grupo 01-10am

Carlos Loria-Saenz

Ciclo I 2022



Índice

Introducción General	1
<i>String Matching</i>	1
¿Qué es String Matching?	1
Tipos de String Matching	1
Algoritmos	1
Introducción	1
Fuerza Bruta (Naive Algorithm)	2
Introducción al Algoritmo de Fuerza Bruta	2
Implementación del Algoritmo de Fuerza Bruta	2
Análisis del Algoritmo de Fuerza Bruta	2
Paso 1	2
Paso 2	2
Paso 3	2
Paso 4	2
Paso 5	2
Paso 6	2
Knuth-Morris-Pratt (KMP)	2
Introducción al Algoritmo de Knuth-Morris-Pratt	2
Implementación del Algoritmo de Knuth-Morris-Pratt	3
Análisis del Algoritmo de Knuth-Morris-Pratt	4
Paso 1: Establecer el tamaño n de los datos	4
Paso 2: Determinar las operaciones de interés	4

Paso 3: Establecer la relacion de recurrencia para $T_{KMP}(n, m)$ con $n = \text{len}(\text{txt}) \wedge$ $m = \text{len}(\text{Pat})$	4
Paso 4: Resolver la ecuacion de $T_{KMP}(n, m)$ eliminando la recursion	4
Paso 5: Determinar el orden de crecimiento asintotico de $T_{KMP}(n, m)$	4
Paso 6	4
Algoritmo de Boyer-Moore	4
Introducción del Algoritmo de Boyer-Moore	4
Implementación del Algoritmo de boyer-Moore	6
Análisis del Algoritmo de Boyer-Moore	8
Paso 1	8
Paso 2	8
Paso 3	8
Paso 4	8
Paso 5	8
Paso 6	8
Algoritmo de Karp-Rabin	8
Introducción del Algoritmo de Karp-Rabin	8
Implementación del Algoritmo de Karp-Rabin	8
Análisis del Algoritmo de Karp-Rabin	9
Paso 1	9
Paso 2	9
Paso 3	9
Paso 4	9
Paso 5	9
Paso 6	9
ADN	9

Importancia del String Matching con el ADN**9****Referencias****11**

Índice de figuras

1. Causes de Huntington	10
-----------------------------------	----

Índice de tablas

Índice de algoritmos

1.	Algoritmo de fuerza bruta	2
2.	Algoritmo de Knuth-Morris-Pratt	3
3.	Algoritmo de Boyer_Moore	6
4.	Algoritmo de Karp-Rabin	8

Introducción General

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris...

String Matching

¿Qué es String Matching?

String Matching es cuando se agarra un patrón y se buscan todas las instancias de ese patrón dentro de un texto específico. String es el término que se usan para las cadenas de texto y matching la palabra en inglés que dice que dos cosas concuerdan, en este caso la hilera del patrón con la hilera del texto.

String Matching tiene varios usos en la vida real. Situaciones tan variadas como bases de datos musicales (Faro y Lecroq, 2013) y a la misma vez se puede usar para detección de plagio, forense digital, chequeo de palabras al escribir, detección de intrusión y muchísimos otros casos posibles (*Applications of String Matching Algorithms* - *GeeksforGeeks* (2020)). Para esta investigación nos vamos a basar en secuencias de ADN como uso del String Matching.

Tipos de String Matching

String Matching se subdivide en dos categorías: exactas y semejantes. Las dos de las categorías tienen sus propios algoritmos diferentes. En esta investigación no solo le daremos enfoque a la categoría de algoritmos de String Matching exactos, sino que nos vamos a analizar 4 algoritmos de este, los cuales serán: Fuerza Bruta (el más básico), Knuth-Morris-Pratt, Boyer-Moore y finalmente el Rabin-Karp. A cada uno se le va a dar un enfoque mayor en la siguiente parte de la investigación, pero en general lo que los destaca es que los primeros tres son con revisión de caracteres y el último utiliza hashing (que convierte uno o varios elementos de entrada a una función en otro elemento).

Algoritmos

Introducción

En esta parte de la investigación se van a analizar los cuatro algoritmos previamente mencionados. Para esto se va a explicar en términos generales como sirve el programa, se va a generar un algoritmo y finalmente se va a analizar este algoritmo con los seis pasos vistos en clase.

Fuerza Bruta (Naive Algorithm)

Introducción al Algoritmo de Fuerza Bruta

El algoritmo de fuerza bruta es un algoritmo que analiza de izquierda a derecha revisando por caracteres si un patrón existe dentro de un texto.

Implementación del Algoritmo de Fuerza Bruta

Algoritmo 1 Algoritmo de fuerza bruta

```

1: procedure FUERZABRUTA((texto, patron))
2:    $n \leftarrow \text{len}(\text{texto})$ 
3:    $m \leftarrow \text{len}(\text{patron})$ 
4:   for  $i \leq (n - m)$  do                                     ▷ Despues de  $n - m$  no puede ser el patron
5:     for  $j < m$  do                                           ▷ Evalua el patron caracter por caracter
6:       if  $\text{texto}[i + j] \neq \text{patron}[j]$  then                 ▷ Evalua si los caracteres son los mismos
7:         break
8:       end if
9:     end for
10:    if  $j = m$  then                                           ▷ Si llega al final entonces existe y devuelve la posición
11:      return  $i$ 
12:    end if
13:  end for
14:  return -1                                                  ▷ Si pasa por todo y no encuentra no existe devuelva -1
15: end procedure

```

Análisis del Algoritmo de Fuerza Bruta

Paso 1

Paso 2

Paso 3

Paso 4

Paso 5

Paso 6

Knuth-Morris-Pratt (KMP)

Introducción al Algoritmo de Knuth-Morris-Pratt

Es un algoritmo que hace las comparaciones de izquierda a derecha, este realiza la búsqueda usando información basada en fallos previos obtenidos del patrón, esto se hace en una fase de “procesamiento” que crea una tabla de valores sobre su propio contenido. Esta tabla se crea para ver donde podría darse la siguiente coincidencia sin buscar más de 1 vez los caracteres del texto o cadena de caracteres donde se realiza la búsqueda, el algoritmo es de tiempo $O(n + m)$ siendo n = el tiempo de la fase de “searching” y m = el tiempo de la fase de procesamiento.

La tabla de fallos se encarga de evitar que cada carácter del texto sea analizado más de 1 vez, esto lo logra comparando el patrón consigo mismo para ver que partes se repiten. Este método guarda una lista con números le indican al algoritmo cuando debe devolverse desde la posición actual una vez que el patrón no coincida con el texto.

El texto y el patrón van avanzando simultáneamente mientras ambos coincidan, si una vez coinciden del todo pero la letra siguiente sigue cumpliendo con el patrón, entonces el algoritmo mueve el patrón 1 a la derecha, si no coinciden, entonces el patrón se empieza a devolver para intentar hacerlo coincidir con el texto.

Implementación del Algoritmo de Knuth-Morris-Pratt

Algoritmo 2 Algoritmo de Knuth-Morris-Pratt

```

1: procedure KNUTH_MORRIS_PUTH((texto, patron))
2:    $n \leftarrow \text{len}(\text{texto})$ 
3:    $m \leftarrow \text{len}(\text{patron})$ 
4:    $\text{resultado} = \text{False}$ 
5:    $\text{listaDeIndices} = []$ 
6:    $\text{tablaDeFallo} = [0] * m$   $\triangleright$  Tabla que se va a usar para devoluciones en los procesamientos
7:    $i = 0$ 
8:    $j = 0$ 
9:   procedure PROCESAMIENTO((patron, m, tablaDeFallo))
10:     $\text{longitud} = 0$   $\triangleright$  longitud del sufijo del prefijo mas largo
11:     $i = 1$ 
12:    while  $i < m$  do
13:      if  $\text{patron}[i] == \text{patron}[\text{longitud}]$  then
14:         $\text{longitud}++ = 1$ 
15:         $\text{tablaDeFallo}[i] = \text{longitud}$ 
16:         $i++ = 1$ 
17:      else
18:        if  $\text{longitud}! = 0$  then
19:           $\text{longitud} = \text{tablaDeFallo}[\text{longitud} - 1]$ 
20:        else
21:           $\text{tablaDeFallo}[i] = 0$ 
22:           $i++ = 1$ 
23:        end if
24:      end if
25:    end while
26:  end procedure
27:  procedure BÚSQUEDA(())
28:     $i = 0$ 
29:     $j = 0$ 
30:    while  $i < n$  do
31:      if  $\text{patron}[j] == \text{texto}[i]$ 
32:         $i++ = 1$ 
33:         $j++ = 1$ 
34:      end if
35:      if  $j == m$  then
36:         $\text{listaDeIndices} += [i - j]$ 
37:         $\text{resultado} = \text{True}$ 
38:         $j = \text{tablaDeFallo}[j - 1]$ 
39:      else if  $i < n \wedge \text{patron}[j]! = \text{texto}[i]$  then

```

```

40:         if  $j! = 0$  then
41:              $j = \text{tablaDeFallo}[j - 1]$ 
42:         else
43:              $i+ = 1$ 
44:         end if
45:     end if
46: end while
47: end procedure
48: return listaDeIndices
49: end procedure

```

Análisis del Algoritmo de Knuth-Morris-Pratt

Paso 1: Establecer el tamaño n de los datos

El tamaño esta dado por la cantidad de elementos en el texto * la cantidad de elementos en el patron

$$n = \text{len}(\text{txt}) * \text{len}(\text{pat})$$

Paso 2: Determinar las operaciones de interés

Comparaciones entre Pat y Txt y construcción de lps[]

Suponemos el peor caso: todas son iguales a la mas grande de todas

Asumimos que la mas grande vale 1

Paso 3: Establecer la relacion de recurrencia para $T_{KMP}(n, m)$ con $n = \text{len}(\text{txt}) \wedge m = \text{len}(\text{Pat})$

Se obtuvo que:

$$\begin{aligned}
 T_{KMP}(n, m) &= m & \sin &= 0 \\
 &= (2 + t_{kmpS}(n - 1)) + (1 + t_{kmpP}(m - 1)) & \sin &> 0
 \end{aligned}$$

con kmpS = KMP search y kmpP = MKP Processing

Paso 4: Resolver la ecuacion de $T_{KMP}(n, m)$ eliminando la recursion

$$T_{kmp}(n, m) = 2n + m$$

Paso 5: Determinar el orden de crecimiento asintotico de $T_{KMP}(n, m)$

$$T_{kmp}(n, m) \sim O(n + m)$$

con n = el tiempo de la fase de “searching” y m = el tiempo de la fase de procesamiento.

Paso 6

*Instrumentación

Algoritmo de Boyer-Moore

Introducción del Algoritmo de Boyer-Moore

El algoritmo de Boyer-Moore realiza su búsqueda escaneando el patrón deseado desde la posición derecha hasta la posición izquierda. Este utiliza dos heurísticas para completar su

cometido, llamadas 'Bad Character Rule' y 'Good Suffix Rule'. Para ambas heurísticas el algoritmo realiza un "preprocesamiento" del patrón, donde se elaboran dos tablas de interés:

- La tabla BC (Bad Character) basa su elaboración en el desplazamiento necesario para ir desde el carácter que se encuentra más a la derecha (del patrón) hasta la primera ocurrencia de los caracteres específicos que lo componen. Si
- La GS (Good Suffix) se elabora a partir de las coincidencias encontradas entre los sufijos del patrón y los caracteres restantes. Nos da el desplazamiento necesario, desde la izquierda, para encontrar dichas coincidencias. Estas pueden ser exactas (el prefijo aparece exactamente igual) o aproximadas (aparece una parte del sufijo).

Cuando el algoritmo va realizando las comparaciones entre el patrón y el texto y encuentre una discordancia, la decisión entre usar el desplazamiento recomendado por la tabla BC o la tabla GS dependerá netamente de cual de las dos le ofrezca un mayor "salto" o ventaja.

Implementación del Algoritmo de Boyer-Moore

Algoritmo 3 Algoritmo de Boyer_Moore

```

1: procedure BOYER-MOORE((patron, texto))
2:    $sizeP = len(P)$ 
3:    $sizeT = len(T)$ 
4:    $boyerMooreBadChar = [0] * 256$  ▷ 256 es el número generalmente aceptado como
   alfabeto
5:   for  $0 \leq i < sizeP - 1$  do
6:      $boyerMooreBadChar[ord(patron[i])] = sizeP - i - 1$ 
7:   end for
8:    $suff = [0] * sizeP$ 
9:    $f = 0$ 
10:   $g = sizeP - 1$ 
11:   $suff[sizeP - 1] = sizeP$ 
12:  for  $sizeP - 2 \geq i > -1$  do
13:    if  $i > g \wedge suff[i + sizeP - 1 - f] < i - g$  then
14:       $suff[i] = suff[i + sizeP - 1 - f]$ 
15:    else
16:      if  $i < g$  then
17:         $g = i$ 
18:      end if
19:       $f = i$ 
20:      while  $g \geq 0 \wedge P[g] == P[g + sizeP - 1 - f]$  do
21:         $g = g - 1$ 
22:      end while
23:       $suff[i] = f - g$ 
24:    end if
25:  end for
26:   $boyerMooreGoodSuffix = [sizeP] * sizeP$ 
27:  for  $0 \leq i < sizeP$  do
28:    if  $suff[i] == i + 1$  then
29:      for  $0 \leq j < sizeP - 1 - i$  do
30:        if  $boyerMooreGoodSuffix[j] == sizeP$  then
31:           $boyerMooreGoodSuffix[j] = sizeP - 1 - i$ 
32:        end if
33:      end for
34:    end if
35:  end for
36:  for  $0 \leq i < sizeP - 1$  do
37:     $boyerMooreGoodSuffix[sizeP - 1 - suff[i]] = sizeP - 1 - i$ 
38:  end for
39:   $i = 0$ 

```

```

40:   $j = 0$ 
41:  while  $j \leq \text{size}T - \text{size}P$  do
42:     $i = \text{size}P - 1$ 
43:    while  $i \neq -1 \wedge \text{patron}[i] == \text{texto}[i + j]$  do
44:       $i - = 1$ 
45:    end while
46:    if  $i < 0$  then
47:       $\text{print}(j)$ 
48:       $j + = \text{boyerMooreGoodSuffix}[0]$ 
49:    else
50:       $j + = \max(\text{boyerMooreGoodSuffix}[i], \text{boyerMooreBadChar}[\text{ord}(T[i + j]) -$ 
       $\text{size}P + 1 + i])$ 
51:    end if
52:  end while
53: end procedure

```

Análisis del Algoritmo de Boyer-Moore

Paso 1

Paso 2

Paso 3

Paso 4

Paso 5

Paso 6

Algoritmo de Karp-Rabin (o Rabin-Karp)

Introducción del Algoritmo de Karp-Rabin

Implementación del Algoritmo de Karp-Rabin

Algoritmo 4 Algoritmo de Karp-Rabin

```

1: procedure KARP_RABIN((patron, texto))
2:    $n = \text{len}(\text{patron})$ 
3:    $m = \text{len}(\text{texto})$ 
4:    $d = 256$       ▷ Este es el alfabeto defecto que sale en varios analisis (caracteres alfabeto
                    inglés)
5:    $q = 33554393$       ▷ Cualquier número primo
                    sirve, pero preferiblemente alto porque los pequeños solo hacen que el algoritmo corra como
                    fuerza bruta porque más hashes concuerdan
6:    $h = d^{m-1} \bmod(q)$ 
7:    $\text{ValorHashPatron} = 0$ 
8:    $\text{ValorHashVentanaTexto} = 0$ 
9:    $\text{listaIndices} = []$ 
10:  for  $i = 0 < n$  do
11:     $\text{ValorHashPatron} = (d * \text{ValorHashPatron} + \text{patron}[i]) \bmod(q)$       ▷ Esta
                                operación sirve con un abecedario normal como tabla ascii, para usarlo en Python el accesor
                                se tiene que meter como parametro de ord()
12:     $\text{ValorHashVentanaTexto} = (d * \text{ValorHashVentanaTexto} + \text{texto}[i]) \bmod(q)$ 
13:  end for
     $j = 0$  ▷ definirla afuera para poder usarla dentro del for sin tener que redefinirla cada vez
    que empiece el for otra vez
14:  for  $i = 0 \leq m - n$  do
15:    if  $\text{ValorHashPatron} == \text{ValorHashVentanaTexto}$  then ▷ Solo hacer fuerza Bruta cuando
    los valores hash concuerdan
16:      for  $j = 0 < n$  do
17:        if  $\text{patron}[j] \neq \text{texto}[i + j]$  then      ▷ Si la fuerza bruta se incumple salga
18:          break
19:        else
20:           $j++ = 1$ 
21:        end if
22:      end for
23:      if  $j == n$  then

```

```

24:         listaIndices += [i]           ▷ Al llegar al final de la fuerza bruta registra el indice
25:     end if
26: end if
27: if i < m - n then
28:     ValorHashVentanaTexto = (d*(ValorHashVentanaTexto - texto[i]*h) + texto[i +
n]) mod(q)
29:     if ValorHashVentanaTexto < 0 then  ▷ GeeksForGeeks recomienda en caso de
que den hashes negativos
30:         ValorHashVentanaTexto += q
31:     end if
32: end if
33: end for
34: return listaIndices
35: end procedure

```

Análisis del Algoritmo de Karp-Rabin

Paso 1

Paso 2

Paso 3

Paso 4

Paso 5

El tiempo de este algoritmo es $O(n+m)$ normalmente, pero en el peor de los casos termina siendo igual que fuerza bruta o $O(nm)$ siendo n y m longitudes de texto y patron

Paso 6

ADN

Importancia del String Matching con el ADN

El Ácido desoxirribonucleico o ADN es un ácido nucleico compuesto de 4 nucleótidos, estos son la adenina (A), timina (T), guanina (G) y citosina (C), además de estos 4 nucleótidos, existe un 5 que es el uracilo (U), pero este se encuentra en el ARN reemplazando a la timina (T). Este contiene las instrucciones y la información genética usadas en el desarrollo y funcionamiento de todo ser vivo, incluyendo algunos virus, además, es el responsable de la herencia genética, como el color de ojos, el color de piel, el cabello y todo tipo de cosas que podemos heredar de nuestros padres o antepasados, sin embargo, también es responsable de heredar ciertas enfermedades genéticas, como por ejemplo, la fibrosis quística, la hemofilia, la enfermedad de Huntington, etc.

Para efectos de esta investigación, hemos decidido utilizar la enfermedad de Huntington como ejemplo de string matching en una cadena de ADN.

La enfermedad de Huntington es una enfermedad hereditaria que da la instrucción al cuerpo de producir una proteína llamada Huntingtina (HTT). Aunque se desconoce la función de dicha proteína, se cree que juega un papel importante en las neuronas. Esta enfermedad es

provocada por una mutación en un segmento del ADN conocido como una repetición del trinucleótido CAG (citosa, adenina y guanina). Este segmento CAG normalmente se repite de 10 a 35 veces en personas sanas, sin embargo, en personas con la enfermedad de Huntington, dicho segmento se repite de 36 a as de 120 veces.

Como se puede ver en la siguiente imagen:

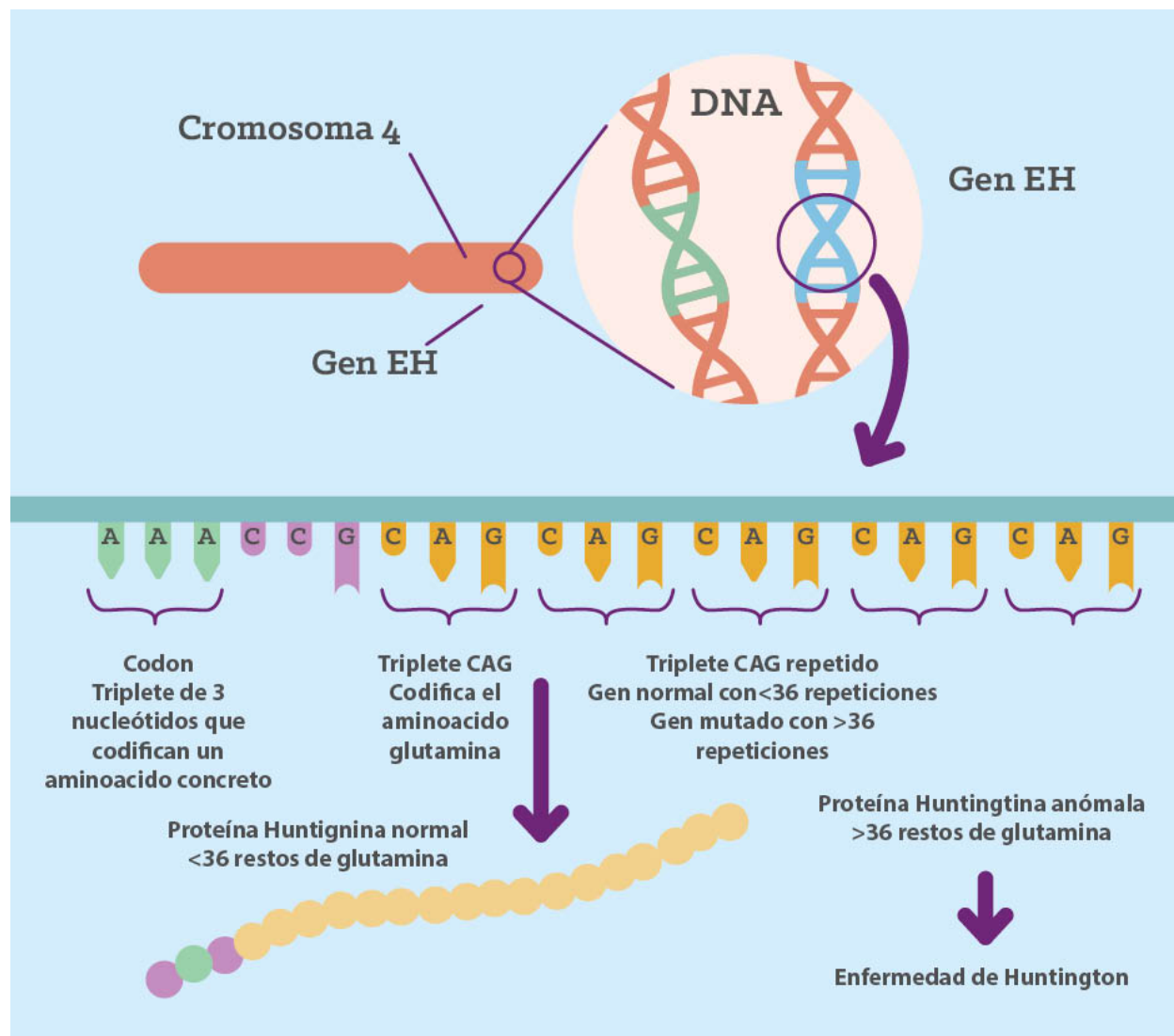


Figura 1
Causes de Huntington

Para efectos del string matching, vamos a comprobar cuantas veces se repite el segmento CAG en la cadena de ADN para comprobar si una persona padece de esta enfermedad.

Conclusión

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris...

Referencias

- Applications of String Matching Algorithms - GeeksforGeeks.* (2020, septiembre). Descargado de <https://www.geeksforgeeks.org/applications-of-string-matching-algorithms> ([Online; accessed 7. May 2022])
- Faro, S., y Lecroq, T. (2013). The exact online string matching problem: A review of the most recent results. *ACM computing surveys*, 45(2), 1–42.