**HỌC VIỆN CÔNG NGHỆ BƯU CHÍNH VIỄN THÔNG**

**KHOA CÔNG NGHỆ THÔNG TIN**



BÁO CÁO CHUYÊN ĐỀ CÔNG NGHỆ PHẦN MỀM

**GIẢNG VIÊN : TS.Nguyễn Duy Phương**

**SINH VIÊN : Phạm Thị Hạnh**

**MSV : B12DCCN315**

**Lớp : D12CNPM2**

*Tháng 6/2016*

MỤC LỤC

[1. GIỚI THIỆU CHUNG 5](#_Toc454527288)

[2. Phân loại các thuật toán đối sánh mẫu 6](#_Toc454527289)

[2.1.Thuật toán tìm kiếm mẫu từ trái qua phải 7](#_Toc454527290)

[2.1.1.Brute Force algorithm 7](#_Toc454527291)

[2.1.2. Search with an automaton 10](#_Toc454527292)

[2.1.3.Karp-Rabin algorithm 15](#_Toc454527293)

[2.1.4.Shift Or Algorithm 18](#_Toc454527294)

[2.1.5.Morris-Pratt algorithm 21](#_Toc454527295)

[2.1.6.Knuth-Morris-Pratt algorithm 25](#_Toc454527296)

[2.1.7.Apostolico-Crochemore algorithm 28](#_Toc454527297)

[2.1.8. Not so naive algorithm 32](#_Toc454527298)

[2.2. Trình bày các thuật toán tìm kiếm mẫu từ phải qua trái 35](#_Toc454527299)

[2.2.1. Boyer Moore algorithm 35](#_Toc454527300)

[2.2.2. Zhu Kataoka algorithm: 40](#_Toc454527301)

[2.2.3. Berry-Ravindran algorithm: 43](#_Toc454527302)

[2.2.4. Turbo-BM algorithm 47](#_Toc454527303)

[2.2.5. Tuned boyer-moore algorithm 49](#_Toc454527304)

[2.2.6. Apostolico-giancarlo algorithm: 53](#_Toc454527305)

[2.3. Trình bày các thuật toán tìm kiếm mẫu từ vị trí cụ thể: 56](#_Toc454527306)

[2.3.1. Colussi algorithm: 56](#_Toc454527307)

[2.3.2 Skip Search Algorithm 61](#_Toc454527308)

[2.4. Trình bày các thuật toán tìm kiếm mẫu từ vị trí bất kỳ: 63](#_Toc454527309)

[2.4.1. Horspool algorithm: 63](#_Toc454527310)

[2.4.2. Quick Search algorithm 66](#_Toc454527311)

[2.4.3. Smith algorithm 68](#_Toc454527312)

[2.4.4. Raita algorithm 72](#_Toc454527313)

# GIỚI THIỆU CHUNG

Để một máy tính tìm kiếm hoạt động hiệu quả, ngoài kỹ thuật thu thập thông tin và tạo chỉ mục cho thông tin, chung ta cũng cần quan tâm đến việc sử dụng các thuật toán tối ưu để tìm kiếm dữ liệu.

Dữ liệu trong máy tính được lưu trữ dưới rất nhiều dạng khác nhau nhưng phổ biến nhất vẫn là dạng chuỗi. Một phép toán cơ bản trên chuỗi là đối sánh mẫu (pattern matching), bài tóa yêu cầu ta tìm ra một hoặc nhiều vị trí xuất hiện của mẫu trên một văn bản. Trong đó mẫu có độ dài m và văn bản có độ dài n (m ≤ n ), tập các kỹ tự được dùng gọi là bảng chữ cái ∑, có số lượng là δ.

Để tăng tốc độ tìm kiếm dữ liệu thì đối sánh đa mẫu được sử dụng. Trong phương pháp đối sánh đơn mẫu thì tại một thời điểm chỉ có một mẫu được đối sánh. Còn trong phương pháp đối sánh đa mẫu, tại một thời điểm nhiều mẫu được đồng thời đối sánh.

Việc đối sánh mẫu diễn ra với nhiều lần thử trên các đoạn khác nhau của văn bản. Trong đó cửa sổ là một chuỗi m ký tự liên tiếp trên văn bản. Mỗi lần thử chương trình sẽ kiểm tra sự giống nhau giữa mẫ với cửa sổ hiện thời. Tùy theo kết quả kiểm tra cửa sổ sẽ được dịch đi sang phải trên văn bản cho lần thử tiếp theo.

# Phân loại các thuật toán đối sánh mẫu

Ta có thể phân thuật toán thành 4 nhóm dựa trên cách thức tìm kiếm của những thuật toán đó.

**Nhóm 1: Nhóm các thuật toán duyệt từ trái sang phải:**

1. Brute force algorithm

2. Search with an automaton

3. Karp-Rabin algorithm

4. Shift Or algorithm

5. Morris-Pratt algorithm

6. Knuth-Morris-Pratt algorithm

7. Apostolico-Crochemore algorithm

8. Not so naive algorithm

**Nhóm 2: Nhóm các thuật toán duyệt từ phải sang trái**

1. Boyer-Moore algorithm

2. Turbo-BM algorithm

3. Tuned boyer-moore algorithm

4. Apostolico-giancarlo algorithm

5. Zhu-Takaoka algorithm

6. Berry-Ravindran algorithm

**Nhóm 3: Nhóm các thuật toán duyệt tại vị trí cụ thể**

1. Colussi algorithm

2. Skip search algorithm

**Nhóm 4: Nhóm các thuật toán duyệt tại vị trí bất kỳ**

1. Horspool algorithm

2. Quick Search algorithm

3. Smith algorithm

4. Raita algorithm

## Thuật toán tìm kiếm mẫu từ trái qua phải

## 2.1.1.Brute Force algorithm

* *Đặc điểm*
* Không có pha chuẩn bị
* Bộ nhớ cần dùng cố định
* Luôn luôn dịch 1 bước sang phải
* Việc so sánh có thể dùng trong các trường hợp
* Độ phức tạp pha thực thi là O( m x n)
* *Thuật toán*

Thuật toán Brute Force bao gồm kiểm tra tất cả các vị trí trong đoạn văn bản giữa 0 và n-m , không cần quan tâm liệu mẫu này có tồn tại ở vị trí đó hay không. Sau đó, sau mỗi lần kiểm tra mẫu sẽ dịch sang phải một vị trí.

* *Kiểm thử thuật toán*

X = “GCAGAGAG”

Y =”GCATCGCAGAGAGTATACAGTACG”

G C A T C G C A G A G A G T A T A C A G T A C G

1 2 3 4

G C A G A G A G

G C A T C G C A G A G A G T A T A C A G T A C G

1

G C A G A G A G

G C A T C G C A G A G A G T A T A C A G T A C G

1

G C A G A G A G

G C A T C G C A G A G A G T A T A C A G T A C G

1

G C A G A G A G

G C A T C G C A G A G A G T A T A C A G T A C G

1

G C A G A G A G

G C A T C G C A G A G A G T A T A C A G T A C G

1 2 3 4 5 6 7 8

G C A G A G A G

G C A T C G C A G A G A G T A T A C A G T A C G

1

G C A G A G A G

G C A T C G C A G A G A G T A T A C A G T A C G

1

G C A G A G A G

G C A T C G C A G A G A G T A T A C A G T A C G

1 2

G C A G A G A G

G C A T C G C A G A G A G T A T A C A G T A C G

1

G C A G A G A G

G C A T C G C A G A G A G T A T A C A G T A C G

1 2

G C A G A G A G

G C A T C G C A G A G A G T A T A C A G T A C G

1

G C A G A G A G

G C A T C G C A G A G A G T A T A C A G T A C G

1 2

G C A G A G A G

G C A T C G C A G A G A G T A T A C A G T A C G

1

G C A G A G A G

G C A T C G C A G A G A G T A T A C A G T A C G

1

G C A G A G A G

G C A T C G C A G A G A G T A T A C A G T A C G

1

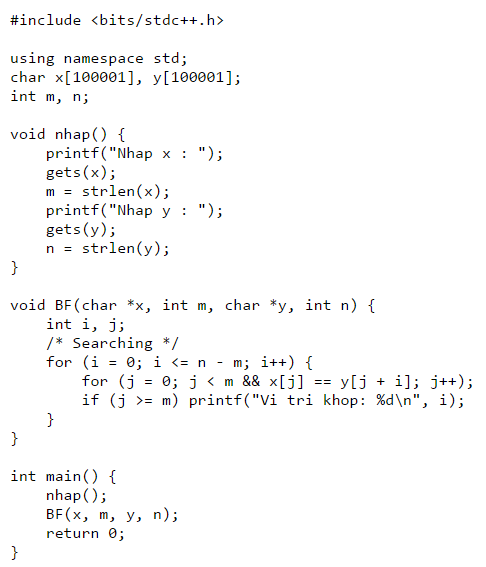
G C A G A G A G

G C A T C G C A G A G A G T A T A C A G T A C G

1

G C A G A G A G

* *Chương trình chạy*



## 2.1.2. Search with an automaton

* *Đặc điểm*
* Yêu cầu xây dựng automation đơn định (DFA)
* Pha xử lý có độ phức tạp tính toán là O(n∂)
* Quá trình tím kiếm có độ phức tạp là O(n) . trường hợp DFA đươc xây dựng bằng cây cân bằng thì độ phức tạp là O(n\*log(∂))
* *Trình bày thuật toán*

Trong thuật toán này, quá trình tìm kiếm được đưa về một quá trình biến đổi trạng thái automat. Hệ thống automat trong thuật toán DFA sẽ được xây dựng dựa trên xâu mẫu. Mỗi trạng thái (nút) của automat lúc sẽ đại diện cho số ký tự đang khớp của mẫu với văn bản. Các ký tự của văn bản sẽ làm thay đổi các trạng thái. Và khi đạt được trạng cuối cùng có nghĩa là đã tìm được một vị trí xuất hiện ở mẫu.

Thuật toán này có phần giống thuật toán Knuth-Morris-Pratt trong việc nhảy về trạng thái trước khi gặp một ký tự không khớp, nhưng thuật toán DFA có sự đánh giá chính xác hơn vì việc xác định vị trí nhảy về dựa trên ký tự không khớp của văn bản (trong khi thuật toán KMP lùi về chỉ dựa trên vị trí không khớp).

Việc xây dựng hệ automat khá đơn giản khi được cài đặt trên ma trận kề. Khi đó thuật toán có thời gian xử lý là O(n) và thời gian và bộ nhớ để tạo ra hệ automat là O(m\*d) (tùy cách cài đặt) . Nhưng ta nhận thấy rằng trong DFA chỉ có nhiều nhất m cung thuật và m cung nghịch, vì vậy việc lưu trữ các cung không cần thiết phải lưu trên ma trận kề mà có thể dùng cấu trúc danh sách kề Forward Star để lưu trữ. Như vậy thời gian chuẩn bị và lượng bộ nhớ chỉ là O(m). Tuy nhiên thời gian tìm kiếm có thể tăng lên một chút so với cách lưu ma trận kề.

* *Kiểm thử thuật toán*

X = “GCAGAGAG”

Y = “GCATCGCAGAGAGTATACAGTACG”

Pha tiền xử lý xây dựng DFA

Pha tìm kiếm: Curent state is: 0

**G** C A T C G C A G A G A G T A T A C A G T A C G

1

G **C** A T C G C A G A G A G T A T A C A G T A C G

2

G C **A** T C G C A G A G A G T A T A C A G T A C G

3

G C A **T** C G C A G A G A G T A T A C A G T A C G

0

G C A T **C** G C A G A G A G T A T A C A G T A C G

0

G C A T C **G** C A G A G A G T A T A C A G T A C G

1

G C A T C G **C** A G A G A G T A T A C A G T A C G

2

G C A T C G C **A** G A G A G T A T A C A G T A C G

3

G C A T C G C A **G** A G A G T A T A C A G T A C G

4

G C A T C G C A G **A** G A G T A T A C A G T A C G

5

G C A T C G C A G A **G** A G T A T A C A G T A C G

6

G C A T C G C A G A G **A** G T A T A C A G T A C G

7

G C A T C G C A G A G A **G** T A T A C A G T A C G

8

G C A T C G C A G A G A G **T** A T A C A G T A C G

0

G C A T C G C A G A G A G T **A** T A C A G T A C G

0

G C A T C G C A G A G A G T A **T** A C A G T A C G

0

G C A T C G C A G A G A G T A T **A** C A G T A C G

0

G C A T C G C A G A G A G T A T A **C** A G T A C G

0

G C A T C G C A G A G A G T A T A C **A** G T A C G

0

G C A T C G C A G A G A G T A T A C A **G** T A C G

1

G C A T C G C A G A G A G T A T A C A G **T** A C G

0

G C A T C G C A G A G A G T A T A C A G T **A** C G

0

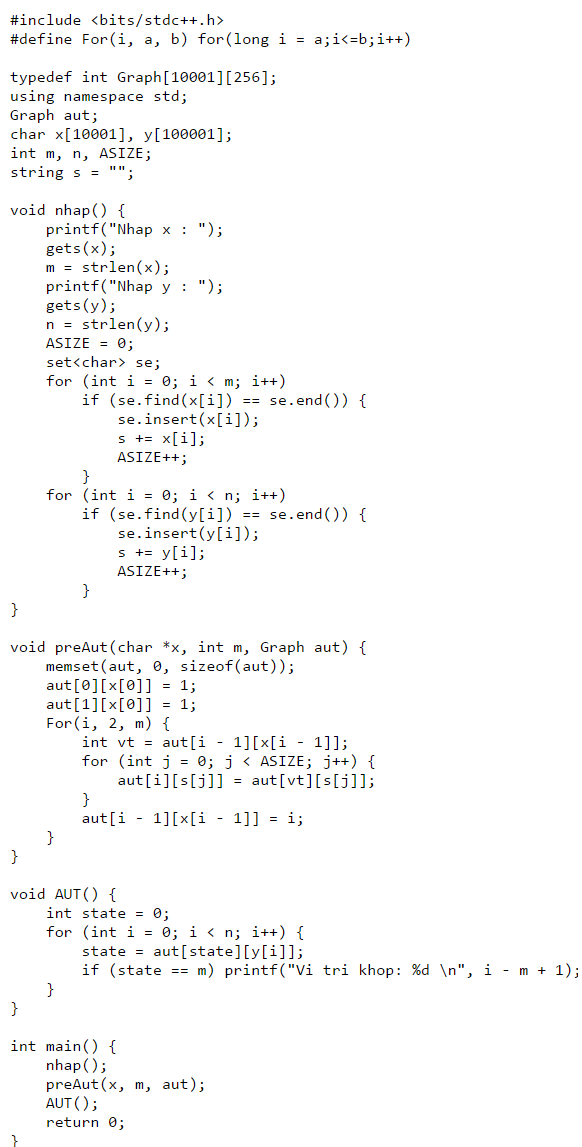
G C A T C G C A G A G A G T A T A C A G T A **C** G

0

G C A T C G C A G A G A G T A T A C A G T A C **G**

1

* *Chương trình chạy*



## 2.1.3.Karp-Rabin algorithm

* *Đặc điểm*
* Sử dụng hàm băm
* Pha chuẩn bị có độ phức tạp thuật toán và có không gian lưu trữ O(m)
* Chương trình chạy có độ phức tạp O(m+n)
* *Trình bày thuật toán*

Hàm băm cung cấp phương thức đơn giản để tránh những con số phức tạp trong việc so sánh những kí tự trong hầu hết các trường hợp thực tế. Thay cho việc kiểm tra từng vị trí trong văn bản nếu như có mẫu xuất hiện, nó chỉ phải kiểm tra những đoạn “gần giống” xâu mẫu. Để kiểm tra sự giống nhau giữa 2 từ sử dụng hàm băm. Giúp cho việc đối chiếu xâu, hàm băm hash:

* Có khả năng tính toán được
* Đánh giá xâu mức cao
* Hash(y[j+1…j+m]) được tính toán dễ hơn dựa trên hash(y[j…j+m-1]) và hash(y[j+m]):

hash(y[j+1 .. j+m])= rehash(y[j], y[j+m], hash(y[j .. j+m-1]).

Với từ w có độ dài m có hash(w) là:

hash(w[0 .. m-1])=(w[0]\*2^(m-1)+ w[1]\*2^(m-2)+···+ w[m-1]\*2^0) mod q

Với q là một số lớn.

Sau đó rehash(a,b,h)= ((h-a\*2^(m-1))\*2+b) mod q.

Pha chuẩn bị của Karp- Rabin có hàm hash(x) có thể tính toán được. Nó được dùng lại không gian nhớ và có độ phức tạp O(m). Trong quá trình thực thi nó so sánh hash(x) với hash([j..j+m-1]) với 0<= j<= n-m. Nếu so sánh đúng, nó phải kiểm tra lại xem các kí tự trong x và y có đúng bằng nhau hay không x=y[j…j+m-1].

* *Kiểm nghiêm thuật toán*

X = “GCAGAGAG”

Y =”GCATCGCAGAGAGTATACAGTACG”

Pha tiền xử lí: Tính được hash(X) = 17597

Pha tìm kiếm thực hiện như sau:

G C A T C G C A G A G A G T A T A C A G T A C G

G C A G A G A G

hash(y[0 .. 7]) = 17819

G C A T C G C A G A G A G T A T A C A G T A C G

G C A G A G A G

hash(y[1 .. 8]) = 17533

G C A T C G C A G A G A G T A T A C A G T A C G

G C A G A G A G

hash(y[2 .. 9]) = 17979

G C A T C G C A G A G A G T A T A C A G T A C G

G C A G A G A G

hash(y[3 .. 10]) = 19389

G C A T C G C A G A G A G T A T A C A G T A C G

G C A G A G A G

hash(y[4 .. 11]) = 17339

G C A T C G C A G A G A G T A T A C A G T A C G

1 2 3 4 5 6 7 8

G C A G A G A G

hash(y[5 .. 12]) = 17597

G C A T C G C A G A G A G T A T A C A G T A C G

G C A G A G A G

hash(y[6 .. 13]) = 17102

G C A T C G C A G A G A G T A T A C A G T A C G

G C A G A G A G

hash(y[7 .. 14]) = 17117

G C A T C G C A G A G A G T A T A C A G T A C G

G C A G A G A G

hash(y[8 .. 15]) = 17678

G C A T C G C A G A G A G T A T A C A G T A C G

G C A G A G A G

hash(y[9 .. 16]) = 17245

G C A T C G C A G A G A G T A T A C A G T A C G

G C A G A G A G

hash(y[10 .. 17]) = 17917

G C A T C G C A G A G A G T A T A C A G T A C G

G C A G A G A G

hash(y[11 .. 18]) = 17723

G C A T C G C A G A G A G T A T A C A G T A C G

G C A G A G A G

hash(y[12 .. 19]) = 18877

G C A T C G C A G A G A G T A T A C A G T A C G

G C A G A G A G

hash(y[13 .. 20]) = 19662

G C A T C G C A G A G A G T A T A C A G T A C G

G C A G A G A G

hash(y[14 .. 21]) = 17885

G C A T C G C A G A G A G T A T A C A G T A C G

G C A G A G A G

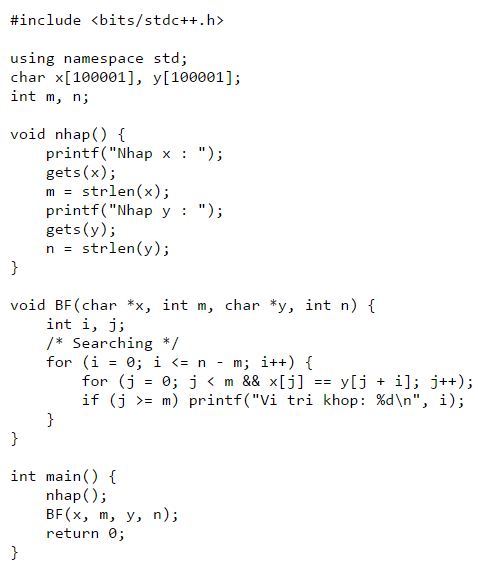
hash(y[15 .. 22]) = 19197

G C A T C G C A G A G A G T A T A C A G T A C G

G C A G A G A G

hash(y[16 .. 23]) = 16961

* *Chương trình chạy*



## 2.1.4.Shift Or Algorithm

* *Đặc điểm*
* Thuật toán sử dụng công nghệ bitwise
* Thuật toán đạt hiệu quả nếu độ dài mẫu không vượt quá kích thước bộ nhớ máy
* Pha xử lý có độ phức tạp O(m +∂)
* Pha tìm kiếm có độ phức tạp O(n). Nó còn phụ thuộc vào kích thước bảng chữ cái và độ dài mẫu
* *Trình bày thuật toán*

Thuật toán Shift Or sử dụng công nghệ bitwise. Ta có R là mảng bit độ dài là m. Vector Rj là giá trị của mảng R sau vị trí y[j] đã xử lý. Nó chứa các thông tin về tất cả các tiền tố x đã phù hợp tại vị trí j hay chưa.

for 0<i<= m-1:

Vector Rj+1 có thể xác định sau khi có Rj và nếu Rj+1[m-1]=0 có nghĩa là ta đã hoàn thành so sánh và tìm được vị trí x xuất hiện trong y. Với mỗi kí tự c ta có Sc là mảng bit độ dài m.

For 0<=i<=m-1, Sc[i]=0 khi và chỉ khi x[i]=c.

Mảng Sc xác định vị trí của kí tự c trong mẫu x. Mỗi Sc sẽ được chuẩn bị trước. Ta có thể xác định nhanh Rj+1:

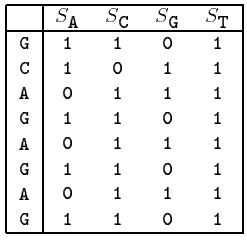
Rj+1=SHIFT(Rj) OR Sy[j+1]

* *Thử nghiệm thuật toán*

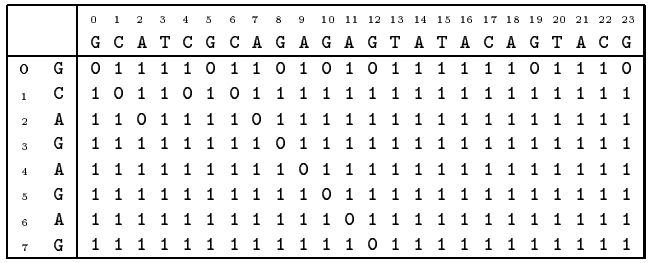
X = “GCAGAGAG”

Y =”GCATCGCAGAGAGTATACAGTACG”

Pha tiền xử lý xác định:

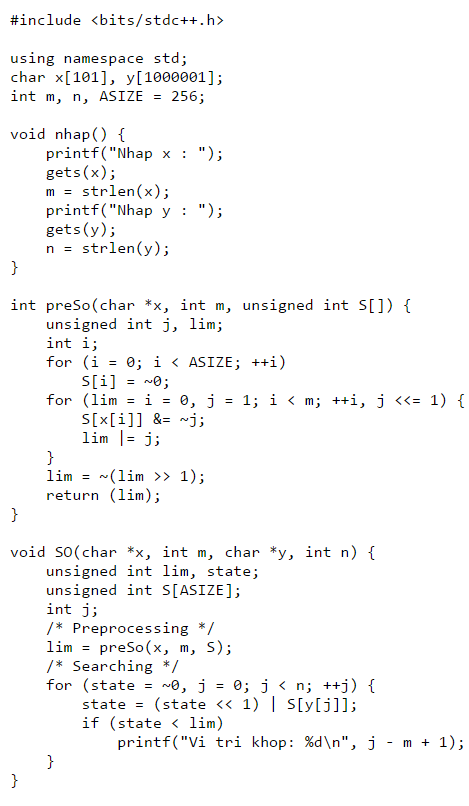


Pha tìm kiếm thực hiện qua bảng sau:



Tại vị trí 12 ta có R12 [7] = 0. Nghĩa là có x xuất hiện trong y tại vị trí 12-8+1 = 5;

* *Chương trình chạy*



## 2.1.5.Morris-Pratt algorithm

* *Đặc điểm*
* Thực hiện so sánh từ trái qua phải
* Pha tiền xử lý có độ phức tạp O(m)
* Pha thực thi có độ phức tạp O(m+n) phụ thuộc vào kích thước của mảng chữ cái
* Thực hiện nhiều nhất 2n-1 các so sánh ký tự trong pha thực thi
* *Thuật toán*

Thiết kế của thuật toán Morris Partt là thuật toán Brute Force với các ràng buộc chặt chẽ hơn, sử dụng những thông tin thu thập được trong quá trình quét văn bản.

Cùng nhìn lại thuật toán Brute Force, chúng có thể tăng độ dài bước dịch và đồng thời nhớ một vài phần của văn bản, phần mà nó chiếu với mẫu. Nó sẽ hạn chế số lần so sánh giữa các kí tự trong xâu mẫu và xâu văn bản và tăng tốc độ thực hiện.

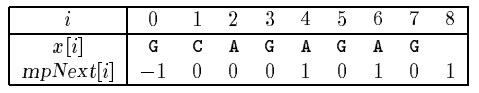
Quan tâm tới bên trái vị trí j trong y khi cửa sổ đang là vị trí của đoạn y[j…j+m]. Giả sử rằng vị trí đầu tiên mà chúng không giống nhau là x[i] và y[i+j] với 0<j<m và x[0…i-1]=y[j…..i+j-1]=u và a=x[i]y[i+j]=b. Khi dịch, nó là cơ sở để chắc chắn rằng tiền tố v của mẫu trùng với một hậu tố của u trong văn bản. Độ dài lớn nhất của v được gọi là biên của u, mảng mpNext[i] là độ dài lớn nhất của biên x[0..n-1] với 0<i<=m. Bên cạnh đó sau mỗi bước dịch ta tiếp tục so sánh kí tự c=x[mpNext[i]] và y[i+j]=b trong trường hợp không tồn tại kí tự nào của x trong y và tránh bị backtrack, giá trị của mpNext[0] được thiết lập là -1.

* *Thử nghiệm thuật toán*

X = “GCAGAGAG”

Y =”GCATCGCAGAGAGTATACAGTACG”

Pha tiền xử lí chuẩn bị bảng mpNext:



Pha tìm kiếm thực hiện:

G C A **T** C G C A G A G A G T A T A C A G T A C G

1 2 3 4

G C A **G** A G A G

Shift by: 3 (i-mpNext[i]=3-0)

G C A **T** C G C A G A G A G T A T A C A G T A C G

1

**G** C A G A G A G

Shift by: 1 (i-mpNext[i]=0- -1)

G C A T **C** G C A G A G A G T A T A C A G T A C G

1

**G** C A G A G A G

Shift by: 1 (i-mpNext[i]=0- -1)

G C A T C G C A G A G A G T A T A C A G T A C G

1 2 3 4 5 6 7 8

G C A G A G A G

Shift by: 7 (i-mpNext[i]=8-1)

G C A T C G C A G A G A G **T** A T A C A G T A C G

1

G **C** A G A G A G

Shift by: 1 (i-mpNext[i]=1-0)

G C A T C G C A G A G A G **T** A T A C A G T A C G

1

**G** C A G A G A G

Shift by: 1 (i-mpNext[i]=0- -1)

G C A T C G C A G A G A G T **A** T A C A G T A C G

1

**G** C A G A G A G

Shift by: 1 (i-mpNext[i]=0- -1)

G C A T C G C A G A G A G T A **T** A C A G T A C G

1

**G** C A G A G A G

Shift by: 1 (i-mpNext[i]=0- -1)

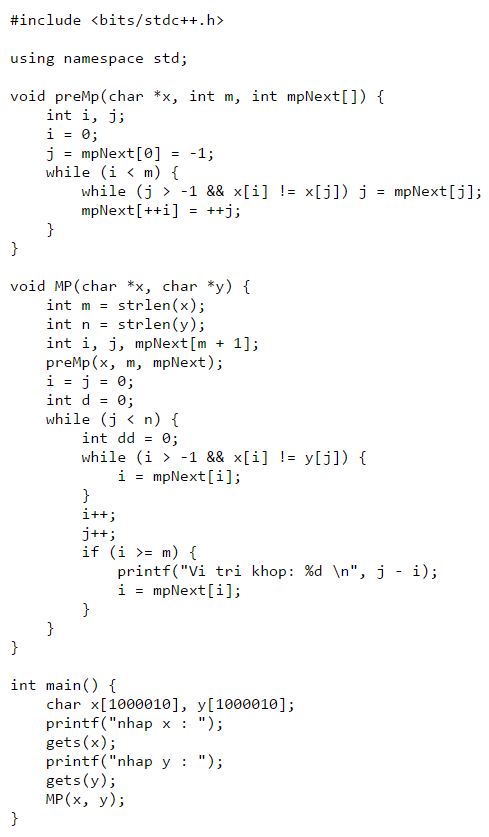
G C A T C G C A G A G A G T A T **A** C A G T A C G

1

**G** C A G A G A G

Shift by: 1 (i-mpNext[i]=0- -1)

* *Chạy chương trình*



## 2.1.6.Knuth-Morris-Pratt algorithm

* *Đặc điểm*
* Thực hiện so sánh từ trái qua phải
* Pha tiền xử lí có độ phức tạp thuật toán là O(m)
* Pha thực thi có độ phức tạp O(m+n) phụ thuộc vào kích thước của mảng chữ cái
* Thực hiện nhiều nhất 2n-1 các so sánh kí tự trong pha thực thi.
* Độ trễ biên được tính bằng logɸ(m) với ɸ được tính là:
* *Thuật toán*

Knuth Morris Partt là một phát triển chặt chẽ hơn của thuật toán Morris part. Hãy nhìn lại Morris Partt, nó có thể tăng được số bước dịch chuyển.

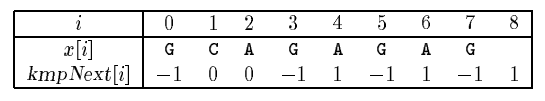
Quan tâm tới bên a trái vị trí j trong y khi cửa sổ đang là vị trí của đoạn y[j…j+m-1]. Giả sử rằng vi trí đầu tiên mà chúng không giống nhau là x[i] và y[i+j] với 0<j<m và x[0…i-1]=y[j…i+j-1]=u và a=x[i]#y[i+j]=b. Khi dịch, nó là cơ sở để chắc chắn rằng tiền tố v của mẫu trùng với hậu tố của u trong văn bản. Tuy nhiên, nếu chúng ta muốn tránh những mâu thuẫn không khớp ngay lập tức, kí tự ngay sau tiền tố v trong mẫu phải khác a. Độ dài lớn nhất của tiền tố v được gọi là nhãn biên của u (nó là kí tự xuất hiện ở 2 đầu của u theo sau bởi kí tự khác nhua trong x). kmpNext[i] là độ dài dài nhất của biên x[i…i-1] nếu theo sau là kí tự c khác với x[i] và bằng -1 nếu nhãn biên tồn tại với 0<i<=m. Bên cạnh đó sau mỗi bước dịch, ta tiếp tục so sánh kí tự c=x[mpNext[i]] và y[i+j]=b trong trường hợp không tồn tại kí tự nào của x trong y và tránh bị backtrack. Giá trị của kmpNext[0] được thiết lập bằng -1.

* *Thử nghiệm thuật toán*

X = “GCAGAGAG”

Y =”GCATCGCAGAGAGTATACAGTACG”

Pha tiền xử lý xác định bảng kmpNext:



Pha tìm kiếm thực hiện như sau:

G C A **T** C G C A G A G A G T A T A C A G T A C G

1 2 3 4

G C A **G** A G A G

Shift by: 4 (i-kmpNext[i]=3- -1)

G C A T **C** G C A G A G A G T A T A C A G T A C G

1

**G** C A G A G A G

Shift by: 1 (i-kmpNext[i]=0- -1)

G C A T C G C A G A G A G T A T A C A G T A C G

1 2 3 4 5 6 7 8

G C A G A G A G

Shift by: 7 (i-kmpNext[i]=8-1)

G C A T C G C A G A G A G **T** A T A C A G T A C G

2

G **C** A G A G A G

Shift by: 1 (i-kmpNext[i]=1-0)

G C A T C G C A G A G A G **T** A T A C A G T A C G

1

**G** C A G A G A G

Shift by: 1 (i-kmpNext[i]=0- -1)

G C A T C G C A G A G A G T **A** T A C A G T A C G

1

**G** C A G A G A G

Shift by: 1 (i-kmpNext[i]=0- -1)

G C A T C G C A G A G A G T A **T** A C A G T A C G

1

**G** C A G A G A G

Shift by: 1 (i-kmpNext[i]=0- -1)

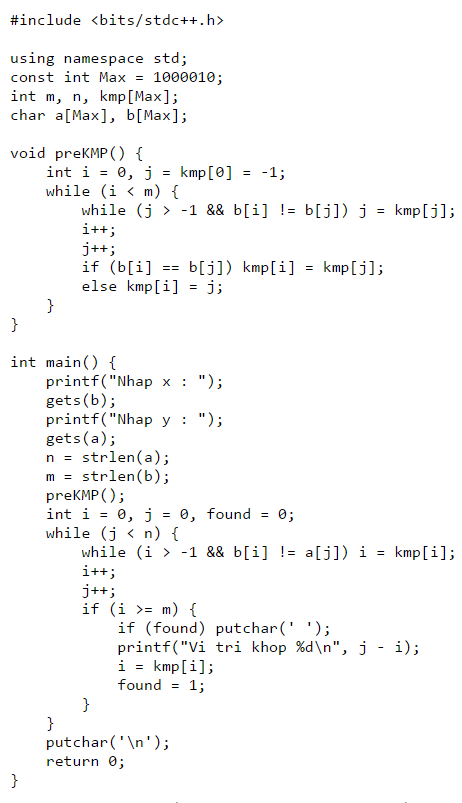
G C A T C G C A G A G A G T A T **A** C A G T A C G

1

**G** C A G A G A G

Shift by: 1 (i-kmpNext[i]=0- -1)

* *Chương trình chạy*

**

## 2.1.7.Apostolico-Crochemore algorithm

* *Đặc điểm*
* Pha tiền xử lý có độ phức tạp O(m)
* Pha tìm kiếm có độ phức tạp O(n)
* Trường hợp xấu nhất thuật toán thực hiện 3/2 n lần so sánh kí tự
* *Thuật toán*

Thuật toán Apostolico – Crochemor sử dụng bảng kmpNext từ thuật toán KMP. Định nghĩa l là số lượng kí tự giống nhau đầu tiên trong xâu x. Trường hợp x = c^m thì l = 0. Tổng quát x = a^l\*bu với a,b là các kí tự, u là một xâu con. a !=b trong mỗi một phép thứ ta sẽ so sánh mẫu lần lượt các vị trí : l, l+1,… ,m-2,m-1,0,1,..,l-1.

Trong pha tìm kiếm ta quan tâm tới bộ 3 số (i,j,k):

khi của sổ đặt ở vị trí khới với y[j… j+m-1]; 0 <=k <= l và x[0..k-1] = y[j.. j+k-1]; l <=i < m và x[l.. i-1] = y[j+l.. i+j-1]; trạng thái khới tạo bạn đâu với bộ (l,0,0). Chúng ta sẽ giải thích làm cách nào để xác định được bộ (i,j,k) tiếp theo.

Có 3 trường hợp tùy thuộc vào các giá trị của i như sau :

- Trường hợp i = l : nếu x[i] = y[i+j] thì bộ kế tiếp (i+1,j,k) nếu x[i] != y[i+j] thì bộ kế tiếp là (l,j+1,max{0,k-1}).

- Trường hớp l < i < m: nếu x[i] = y[i+j] thì bộ kế tiếp là (i+1,j,k) nếu x[i] != y[i+j] thì sẽ có 2 trường hợp xẩy ra phụ thuộc vào giá trị kmpNext[i]:

Nếu kmpNext[i] <= l thì bộ tiếp theo : ( , i+jkmpNext[i], max{0, kmpNext[i]})

Nếu kmpNext[i] > l thì bộ tiếp theo : (kmpNext[i], i+j-kmpNext[i], )

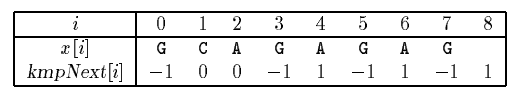
- Trường hợp l = m : nếu k < l và x[k]=y[j+k] thì bộ tiếp theo là (i, j, k+1).

* *Kiểm nghiệm thuật toán:*

X = “GCAGAGAG”

Y =”GCATCGCAGAGAGTATACAGTACG”

Pha tiền xử lý xác định :



Pha tìm kiếm thực hiện như sau :

G C A **T** C G C A G A G A G T A T A C A G T A C G

1 2 3

G C A **G** A G A G

Shift by: 4 (i-kmpNext[i]=3- -1)

G C A T C **G** C A G A G A G T A T A C A G T A C G

1

G **C** A G A G A G

Shift by: 1 (i-kmpNext[i]=1-0)

G C A T C G C A G A G A G T A T A C A G T A C G

8 1 2 3 4 5 6 7

G C A G A G A G

Shift by: 7 (i-kmpNext[i]=8-1)

G C A T C G C A G A G A G **T** A T A C A G T A C G

1

G **C** A G A G A G

Shift by: 1 (i-kmpNext[i]=1-0)

G C A T C G C A G A G A G T **A** T A C A G T A C G

1

G **C** A G A G A G

Shift by: 1 (i-kmpNext[i]=1-0)

G C A T C G C A G A G A G T A **T** A C A G T A C G

1

G **C** A G A G A G

Shift by: 1 (i-kmpNext[i]=1-0)

G C A T C G C A G A G A G T A T **A** C A G T A C G

1

G **C** A G A G A G

Shift by: 1 (i-kmpNext[i]=1-0)

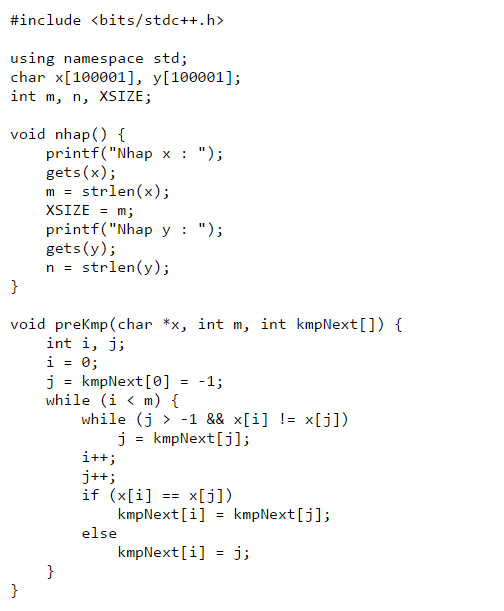
G C A T C G C A G A G A G T A T A C A G **T** A C G

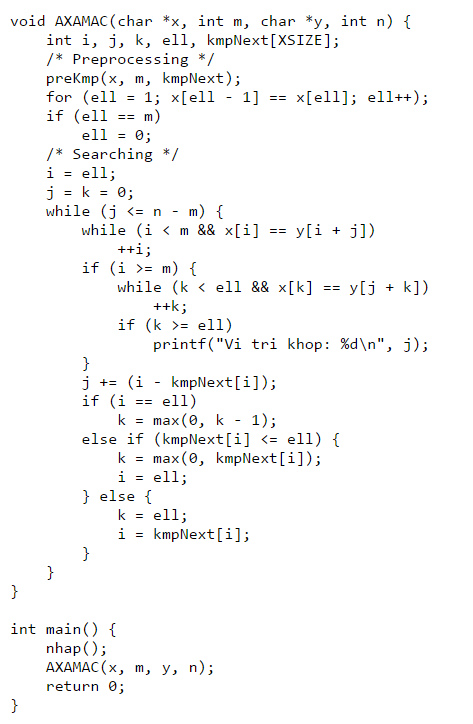
1 2 3 4

G C A G **A** G A G

Shift by: 3 (i-kmpNext[i]=4-1)

* *Chương trình chạy:*





## 2.1.8. Not so naive algorithm

* *Đặc điểm:*

- Pha chuẩn bị có độ phức tạp thuật toán không thay đổi.

- Cộng thêm phức tạp không gian lưu trữ.

- Độ phức tạp pha thực thi là O(mxn).

* *Trình bày thuật toán:*

Trong quá trình thực thi của Not So Naïve các kí tự so sánh với các vị trí của mẫu theo thứ tự 1,2,3…m-2, m-1, 0. Với mỗi trường hợp, trên cửa sổ dịch chuyển là các vị trí của xâu văn bản y[j…j+m-1]: nếu x[0] = x[1] và x[1] != y[j+1] hoặc nếu   
x[1] != x[0] và x[1] = y[j+1] thì số bước dịch là 2 so với vị trí trước đó. Và dịch 1 bước với các trường hợp còn lại. Do vậy, pha tiền xử lí có thể xác định được độ phức tạp thuật toán và không gian nhớ.

* *Kiểm nghiệm thuật toán:*

X = “GCAGAGAG”

Y =”GCATCGCAGAGAGTATACAGTACG”

Pha tiền xử lí:

k = 1; l = 2;

Pha tìm kiếm thực hiện như sau :

G C A **T** C G C A G A G A G T A T A C A G T A C G

1 2 3

G C A **G** A G A G

Shift by: 2 (l)

G C A **T** C G C A G A G A G T A T A C A G T A C G

1

G **C** A G A G A G

Shift by: 1 (k)

G C A T C **G** C A G A G A G T A T A C A G T A C G

1 2

G C **A** G A G A G

Shift by: 2 (l)

G C A T C G C A G A G A G T A T A C A G T A C G

8 1 2 7 3 6 4 5

G C A G A G A G

Shift by: 2 (l)

G C A T C G C A **G** A G A G T A T A C A G T A C G

1

G **C** A G A G A G

Shift by: 1 (k)

G C A T C G C A G **A** G A G T A T A C A G T A C G

1

G **C** A G A G A G

Shift by: 1 (k)

G C A T C G C A G A **G** A G T A T A C A G T A C G

1

G **C** A G A G A G

Shift by: 1 (k)

G C A T C G C A G A G **A** G T A T A C A G T A C G

1

G **C** A G A G A G

Shift by: 1 (k)

G C A T C G C A G A G A **G** T A T A C A G T A C G

1

G **C** A G A G A G

Shift by: 1 (k)

G C A T C G C A G A G A G **T** A T A C A G T A C G

1

G **C** A G A G A G

Shift by: 1 (k)

G C A T C G C A G A G A G T **A** T A C A G T A C G

1

G **C** A G A G A G

Shift by: 1 (k)

G C A T C G C A G A G A G T A **T** A C A G T A C G

1

G **C** A G A G A G

Shift by: 1 (k)

G C A T C G C A G A G A G T A T **A** C A G T A C G

1

G **C** A G A G A G

Shift by: 1 (k)

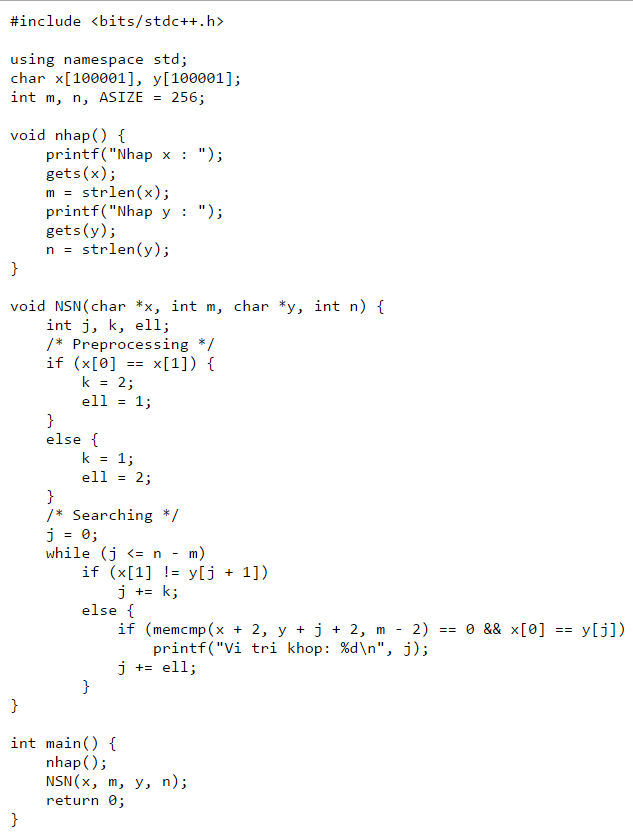
G C A T C G C A G A G A G T A T A C A G **T** A C G

1 2 3 4

G C A G **A** G A G

Shift by: 2 (l)

* *Chương trình chạy:*



# 2.2. Trình bày các thuật toán tìm kiếm mẫu từ phải qua trái

## 2.2.1. Boyer Moore algorithm

* *Đặc điểm:*

- Thuật toán thực thi việc so sánh từ phải sang trái.

- Pha cài đặt có độ phức tạp thuật toán và không gian nhớ là O(m+σ).

- Pha thực thi có độ phức tạp là O(mxn).

- 3n kí tự xâu văn bản được so sánh trong trường hợp xấu nhất khi thực thi với mẫu không tuần hoàn.

- Độ phức tạp O(n/m) khi thực thi tốt nhất.

* *Trình bày thuật toán:*

Boyer-Moore quan tâm tới thuật toán đối sánh xâu hiệu quả nhất thường thấy. Các biến thể của nó thường được dùng trong các bộ soạn thảo cho các lệnh như <<search>> và <<subtitute>>. Thuật toán sẽ quét các ký tự của mẫu (pattern) từ phải sang trái bắt đầu ở phần tử cuối cùng.

Trong trường hợp mis-match (hoặc là trường hợp đã tìm được 1 đoạn khớp với mẫu), nó sẽ dùng 2 hàm được tính toán trước để dịch cửa sổ sang bên phải. Hai hàm dịch chuyển này được gọi là good-suffix shift (còn được biết với cái tên phép dịch chuyển khớp) và bad-character shift (còn được biết với cái tên phép dịch chuyển xuất hiện).

Đối với mẫu x[0..m-1] ta dùng 1 biến chỉ số i chạy từ cuối về đầu, đối với chuỗi y[0..n-1] ta dùng 1 biến j để chốt ở phía đầu.

G/s rằng trong quá trình so sánh ta gặp 1 mis-match tai vị trí x[i]=a của mẫu và y[i+j]=b trong khi đang thử khớp tại vị trí j.

Khi đó, x[i+1..m-1]=y[j+i+1..j+m-1]=u và x[i]≠y[i+j] . Bây giờ ta đi xét xem đối với từng trường hợp, 2 hàm trên sẽ thực hiện việc dịch chuyển như thế nào:

Phép dịch chuyển good-suffix shift sẽ dịch cửa sổ sang bên phải cho đến khi gặp 1 ký tự khác với x[i] trong trường hợp đoạn u lại xuất hiện trong x.

Fig2: good-suffix shift, trường hợp u lại xuất hiện trong x

Nếu đoạn u không xuất hiện lại trong x, mà chỉ có 1 phần cuối( suffix) của u khớp với phần đầu (prefix) của x, thì ta sẽ dịch 1 đoạn sao cho phần suffix dài nhất vcủa y[j+i+1..j+m-1] khớp với prefix của x.

Fig3: good-suffix shift, trường hợp chỉ suffix của u xuất hiện trong x

Phép dịch chuyển bad-character shift sẽ khớp kí tự y[i+j] với 1 ký tự (bên phải nhất) trong đoạn x[0..m-2] (các bạn thử nghĩ xem tại sao không phải là m-1)

Fig4: bad-character shift

Nếu y[i+j] không xuất hiện trong x, ta thấy ngay rằng không có xuất hiện nào của x trong y mà lại chứa chấp y[i+j], do đó ta có thể đặt cửa sổ ngay sau y[i+j], tức là y[j+i+1] .

Thuật toán Boyer-Moore sẽ chọn đoạn dịch chuyển dài nhất trong 2 hàm dịch chuyển good-suffix shift và bad-character shift. Hai hàm này được định nghĩa như sau:

Hàm good-suffix shift được lưu trong bảng bmGs có kích thước m+1.

Ta định nghĩa 2 điều kiện sau:

1. Cs(i, s): với mỗi k mà i < k < m, s ≥ k hoặc x[k-s]=x[k] và

2. Co(i, s): nếu s <i thì x[i-s] ≠ x[i]

Khi đó, với 0≤ i <m: bmGs[i+1]=min{s>0 : Cs(i, s) and Co(i, s) hold} và chúng ta định nghĩa bmGs[0] là độ dài chu kỳ của x. Việc tính toán bảng bmGs sử dụng 1 bảng suff định nghĩa như sau: với 1 ≤ i < m, suff[i]=max{k : x[i-k+1 ..i]=x[m-k .. m-1]}

Hàm bad-character shift được lưu trong bảng bmBc có kích thước σ. Cho c trong Σ: bmBc[c] = min{i : 1≤ i <m-1 và x[m-1-i]=c} nếu c xuất hiện trong x, m ngược lại.

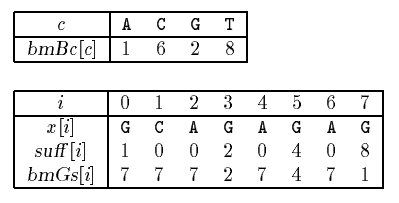
Bảng bmGs và bmBc được tính toán trong thời gian O(m+σ) trước khi thực hiện tìm kiếm và cần 1 không gian phụ là O(m+σ). Giai đoạn tìm kiếm có độ phức tạp thời gian bậc hai nhưng lại chỉ có 3n phép so sánh khi tìm kiếm 1 chuỗi không có chu kì. Đối với việc tìm kiếm trong 1 khối lượng lớn các chữ cái thuật toán thực hiện với tốc độ nhanh “khủng khiếp”. Khi tìm kiếm chuỗi a^(m-1)b trong a^nchuỗi thuật toán chỉ sử dụng O(n/m) phép so sánh, đây được coi là “cảnh giới” cho bất cứ một thuật toán tìm kiếm chuỗi nào mà mẫu đã được xử lý trước.

* *Kiểm nghiệm thuật toán:*

X = “GCAGAGAG”

Y =”GCATCGCAGAGAGTATACAGTACG”

Pha tiền xử lí:



Pha thực thi:

G C A T C G C **A** G A G A G T A T A C A G T A C G

1

G C A G A G A **G**

Shift by: 1 (bmGs[7]=bmBc[A]-8+8)

G C A T C G **C** A G A G A G T A T A C A G T A C G

3 2 1

G C A G A **G** A G

Shift by: 4 (bmGs[5]=bmBc[C]-8+6

G C A T C G C A G A G A G T A T A C A G T A C G

8 7 6 5 4 3 2 1

G C A G A G A G

Shift by: 7 (bmGs[0])

G C A T C G C A G A G A G T A T A **C** A G T A C G

3 2 1

G C A G A **G** A G

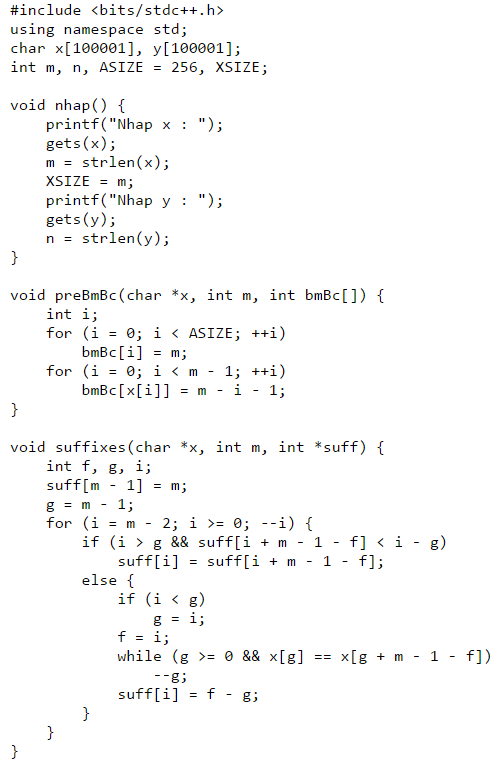
Shift by: 4 (bmGs[5]=bmBc[C]-8+6)

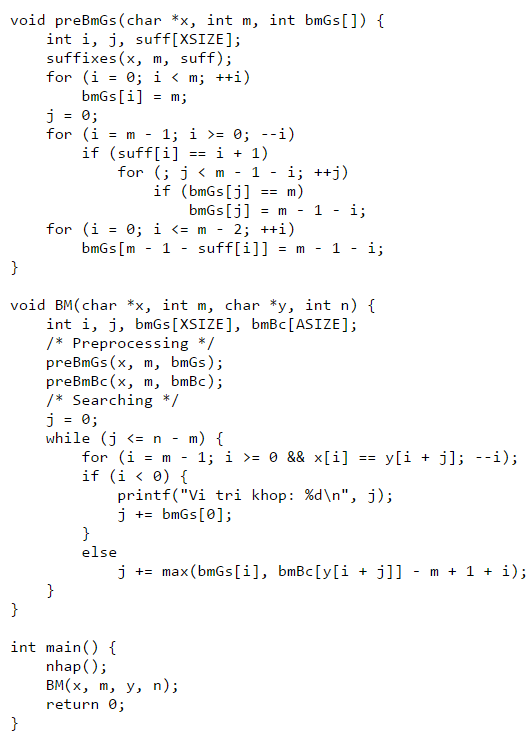
G C A T C G C A G A G A G T A T A C A G T A **C** G

2 1

G C A G A G **A** G

Shift by: 7 (bmGs[6])

* *Chương trình chạy:*
* 



## 2.2.2. Zhu Kataoka algorithm:

* *Đặc điểm:*

- Là một biến thể của Boyer Moore.

- Sử dụng 2 kí tự liên tiếp nhau để tính toán bước dịch bad charater.

- Pha cài đặt có độ phức tạp thuật toán và không gian nhớ là O(m+σ^2).

- Pha thực thi có độ phức tạp là O(mxn).

* *Trình bày thuật toán:*

Zhu và Takaoka thiết kế thuật toán mà chúng thực thi dựa trên bad charater. Trong quá trình tìm kiếm, việc so sánh được thực hiện từ phải qua trái, và khi cửa sổ đang ở vị trí y[j…j+m-1] và xuất hiện sự khác nhau giữa x[m- k] và y[j+m-k] trong khi x[m-k+1 .. m-1]=y[j+m-k+1 .. j+m-1]. Bước dịch good suffix cũng được sử dụng để tính toán bước dịch.

Pha tiền xử lí của thuật toán bao gồm việc tính toán mỗi cặp kí tự (a,b) với a,b là mút bên phải của đoạn x[0…m-2]   
  
Với a,b thuộc , ztBc[a, b]=k và k có các giá trị:

k <m-2 và x[m-k .. m-k+1]=ab và ab không xuất hiện trong đoạn x[m-k+2..m-2]

Hoặc

k=m-1 và x[0]=b và ab không xuất hiện trong đoạn x[0 .. m-2]

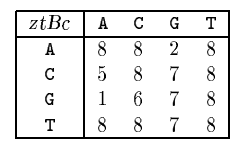
Hoặc

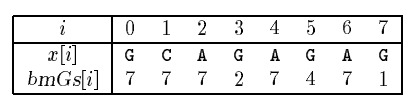
k=m and x[0] bvà ab không xuất hiện trong đoạn x[0 .. m-2]

* *Kiểm nghiệm thuật toán:*

X = “GCAGAGAG”

Y =”GCATCGCAGAGAGTATACAGTACG”

Pha tiền xử lí:

* 

Pha tìm kiếm thực hiện như sau :

G C A T C G C **A** G A G A G T A T A C A G T A C G

1

G C A G A G A **G**

Shift by: 5 (ztBc[C][A])

G C A T C G C A G A G A G T A T A C A G T A C G

8 7 6 5 4 3 2 1

G C A G A G A G

Shift by: 7 (bmGs[0])

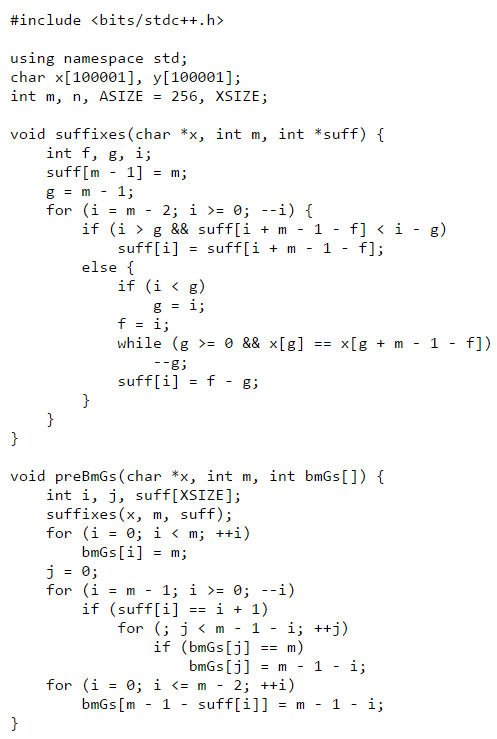
G C A T C G C A G A G A G T A T A **C** A G T A C G

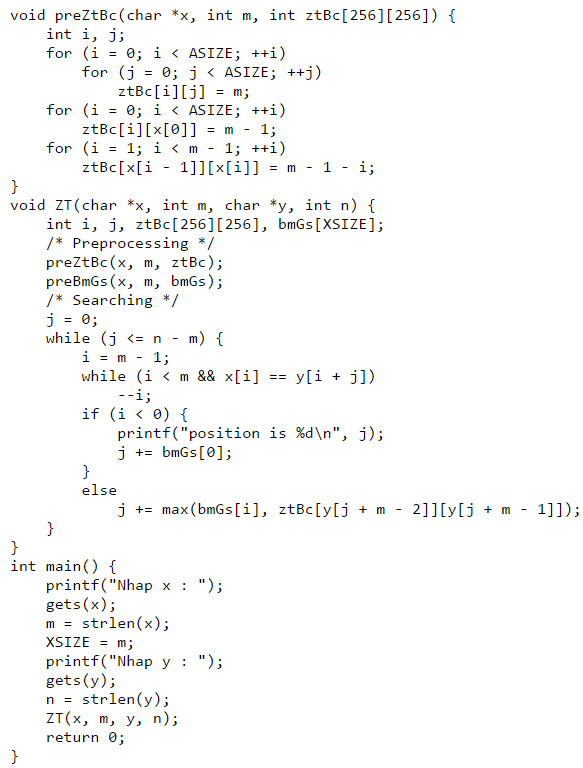
3 2 1

G C A G A **G** A G

Shift by: 7 (bmGs[6])

* *Chương trình chạy:*



* 

## 2.2.3. Berry-Ravindran algorithm:

* *Đặc điểm:*

- Đây là thuật toán kết hợp giữa thuật toán Quic search và Zhu and Takaoka.

- Pha chuẩn bị có độ phức tạp O(m +∂^2).

- Pha tìm kiếm có độ phức tạp O(mn).

* *Trình bày thuật toán:*

Berry và Ravindran đã thiết kế ra thuật toán thực hiện bước dịch dựa vào tư tưởng bad-character. Nhưng ở đây lấy 2 kí tự liên tiếp ngoài cùng bên phải của sổ để xác định bước dịch.

Quá trình chuẩn bị của thuật toán bao gồm việc xác định với mỗi cặp (a,b) vị trị ngoài cũng bên phải gần nhất bắt đầu xuất hiện ab trong x.

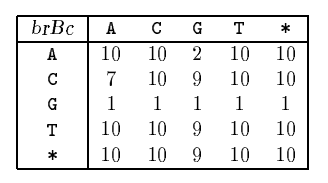
Sau mỗi lần thử mẫu khi cửa sổ đang ở vị trí tương ứng y[j .. j+m-1] bước dịch tiếp theo sẽ là brBc[y[j+m],y[j+m+1]]

* *Kiểm nghiệm thuật toán:*

X = “GCAGAGAG”

Y =”GCATCGCAGAGAGTATACAGTACG”

Pha tiền xử lý xác định:



Pha tìm kiếm thực hiện như sau:

G C A **T** C G C A G A G A G T A T A C A G T A C G

1 2 3 4

G C A **G** A G A G

Shift by: 1 (brBc[G][A])

G **C** A T C G C A G A G A G T A T A C A G T A C G

1

**G** C A G A G A G

Shift by: 2 (brBc[A][G])

G C A **T** C G C A G A G A G T A T A C A G T A C G

1

**G** C A G A G A G

Shift by: 2 (brBc[A][G])

G C A T C G C A G A G A G T A T A C A G T A C G

1 2 3 4 5 6 7 8

G C A G A G A G

Shift by: 10 (brBc[T][A])

G C A T C G C A G A G A G **T** A T A C A G T A C G

1

**G** C A G A G A G

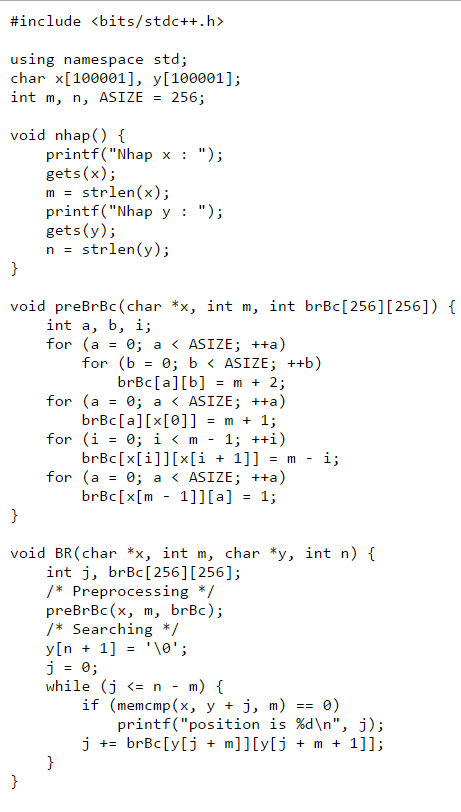
Shift by: 7 (brBc[G][0])

G C A T C G C A G A G A G T A T **A** C A G T A C G

1

**G** C A G A G A G

Shift by: 10 (brBc[0][0])

* *Chương trình chạy:*
* 

## 2.2.4. Turbo-BM algorithm

* *Đặc điểm:*

- Đây là thuật toán phát triến từ thuật toán boyer – moore.

- Pha chuẩn bị có độ phức tạp O(m + ∂).

- Pha tìm kiếm có độ phức tạp O(n).

- Trong trường hợp xấu nhất thuật toán cần so sánh 2n kí tự.

* *Trình bày thuật toán:*

Turbo- BM là thuật toán cải tiến từ thuật toán Boyer-moore. Nó yêu cầu mở rộng thêm quá trính tiền xử lý so với thuật toán boyer-moore cơ bản. Bao gồm việc nhớ một đoạn hậu tố của mẫu trùng với đoạn con của xâu y. Trong lần thử mẫu trước ( chỉ trường hợp good-suffix ta mới xác định được đoạn u này. Trường hợp khác u = 0 )

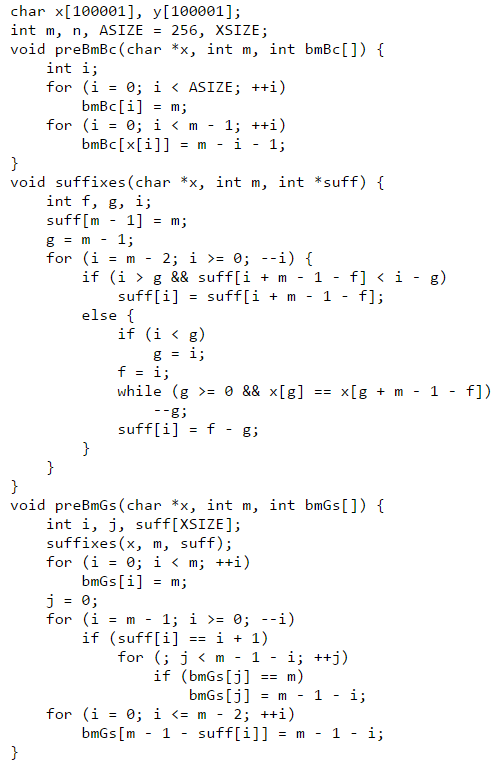
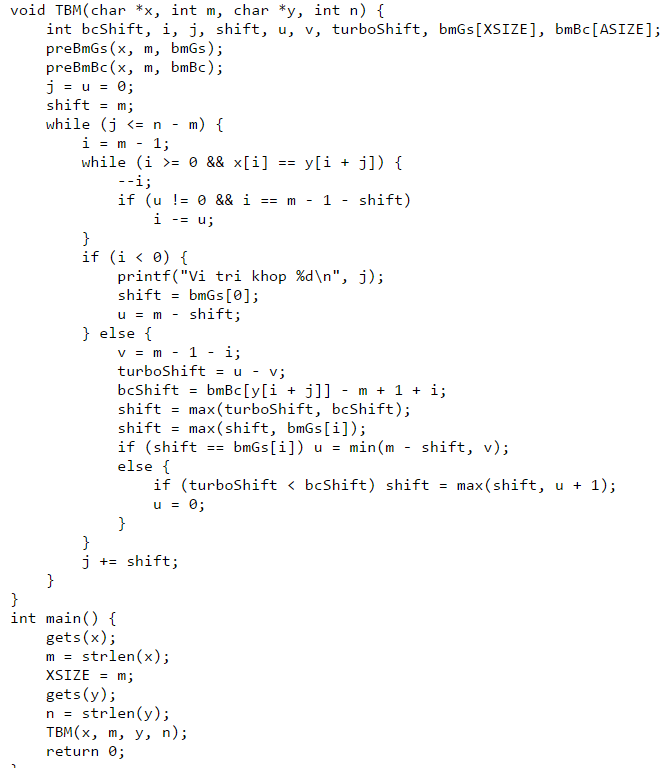
Kĩ thuật này có hai lợi thế sau :

- Nó có thể nhẩy qua đoạn xâu con đã biết trước.

- Nó có thể có một bước dịch khác gọi là turbo-shift.

Bước dịch turbo-shift tồn tại trong trường hợp tại vị trí thử hiện tại, hậu tố xâu mẫu đã matches nho hơn so với đoạn nhớ trước đó. Chúng ta có thể gọi u là đoạn nhớ trước đó và v là đoạn hậu tố của x đã match. Như vậy turbo-shift = |u| - |v|

Trường hợp |v| < |u| , độ dài bad-character shift lớn hơn độ dài goodsuffix shift và turbo-shift thì bước dịch thực sự sẽ lơn hơn |u|+1. Lý do là bởi 2 kí từ c và d là khác nhau. Với bước dịch < |u| + 1. Nếu c khớp với kí tự trong v. thì d sẽ không khớp. ta luôn thấy có 1 cập kí tự khác nhau trong mẫu x và xâu y.

* *Kiểm nghiệm thuật toán:*
* *Chương trình chạy:*
* 
* 

## 2.2.5. Tuned boyer-moore algorithm

* *Đặc điểm:*

- Đây là thuật toán đơn giản hóa từ thuật toán Boyer – Moore.

- Dễ cài đặt.

- Thuật toán có độ phức tạp bình phương trong trường hợp xấu nhất. Nhưng thường chạy rất nhanh trong thực tế.

* *Trình bày thuật toán:*

Tuned- Boyer – moore là cài đặt đơn giản của thuật toán boyer – moore. Chi phí cho thuật toán string- matching thường phần nhiều là việc kiểm tra xem xâu mẫu cố bằng xâu tại vị trí cửa số đang xét hay không.

Để tránh việc phải so sánh nhiều lần. chúng ta có thể thực hiện nhiều bước dịch hơn trước khi thực sự so sánh xâu.

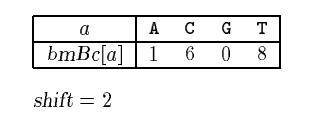
Thuật toán này sẽ sử dụng hàm bad-character xác định bước dịch. Và tìm x[m-1] trong y cho tới khi nào tìm được. Yêu cầu lưu giá trị bmBc[x[m-1]] vào biến shift và đặt lại giá trị bmBc[x[m-1]] = 0. Khi ta tìm được vị trí x[m-1] trong y, thì bước dịch tiếp theo sẽ là shift.

* *Kiểm nghiệm thuật toán:*

X = “GCAGAGAG”

Y =”GCATCGCAGAGAGTATACAGTACG”

Pha chuẩn bị xác định:



Pha tìm kiếm thực hiện như sau:

First attempt

G **C** A T C G C **A** **G** A G A G T A T A C A G T A C G

1

G C A G A G A G

2

G C A G A G A **G**

3

G C A G A G A **G**

4

G C A G A G A **G**

5

**G** C A G A G A G

Shift by: 2 (shift)

Second attempt

G C A **T** C G C A G A **G** A G T A T A C A G T A C G

1

G C A G A G A **G**

2

**G** C A G A G A G

Shift by: 2 (shift)

Third attempt

G C A T C **G C A G A G A G** T A T A C A G T A C G

1

G C A G A G A **G**

2 3 4 5 6 7 8

**G C A G A G A** G

Shift by: 2 (shift)

Fourth attempt

G C A T C G C A G A G A G T **A T A** C A G T A C **G**

1

G C A G A G A **G**

2

G C A G A G A **G**

3

G C A G A G A **G**

4

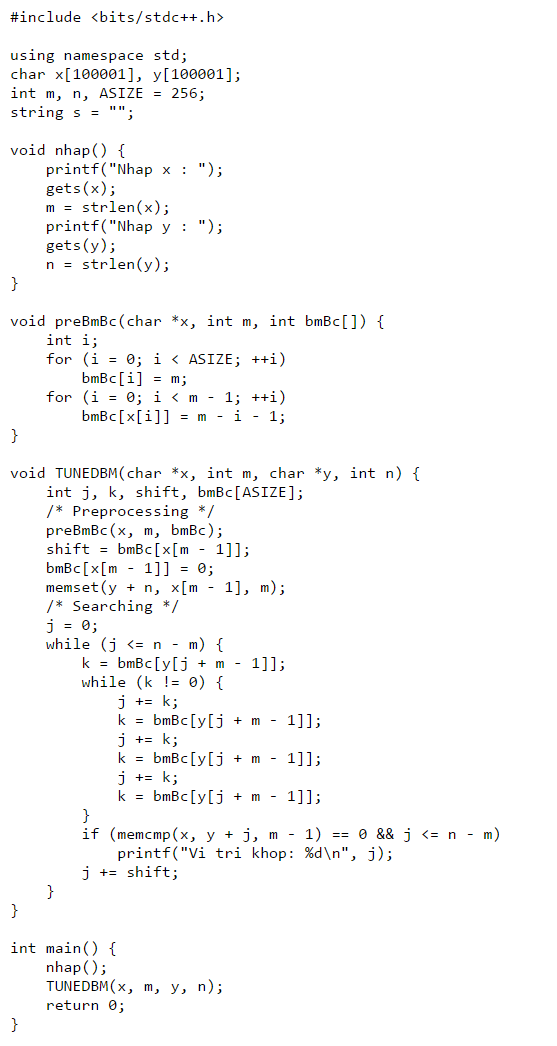
G C A G A G A **G**

5

**G** C A G A G A G

Shift by: 2 (shift)

* *Chương trình chạy:*

1. 

## 2.2.6. Apostolico-giancarlo algorithm:

* *Đặc điểm:*

- Đây là thuật toán cải tiến từ thuật toàn Boyer – moore.

- Pha xử lý có độ phực tạp O(m + ∂ ) không gian và thời gian.

- Pha tìm kiếm có độ phức tạp O(n).

- Trường hợp xấu nhất thuật toán cần 3/2n lần phép so sánh.

* *Trình bày thuật toán:*

Thuật toán Boyer – moore là thuật toán khó để phân tích bởi vì sau mỗi lần thử mẫu nó sẽ quen đi tất cả các kí tự đã đúng trước đó. Apostolico và Giancarlo đã thiết kế một thuật toán có thể nhớ được độ dài lớn nhất của hậu tố xâu mẫu x. Thông tin này sẽ được lưu vào một mảng skip.

Trong mỗi lần thứ tại vị trí j, nếu thuật toán đã so sánh đúng các kí tự y[i+j+1..j+m-1] .

Sẽ có trường hợp xảy ra :

- Trường hợp 1 : k > suff[i] && suff[i] = i+ 1 ; nghĩa là tồn tại x xuất hiện trong y tại vị trí j. bước dịch tiếp theo là per(x).

- Trường hợp 2 : k > suff[i] && suff[i] <= i ; nghĩa là xuất hiện vị trí sai giữa 2 kí tự x[i-suff[i]] and y[i+j-suff[i]] . bước dịch tiếp theo sẽ sử dụng bmBc[y[i+j-suff[i]]] and bmGs[i-suff[i]+1].

- Trường hợp 3 : k < suff[i] .nghĩa là vị trí sai xuật hiện tại x[i-k] and y[i+j-k] . bước dịch tiếp theo sẽ sử dụng bmBc[y[i+j-k] and bmGs[i-k+1].

- Trường hợp 4 : k = suff[i] ; nghĩa là ta đa có k phân tử khơp nhau. Ta sẽ nhẩy qua k phần tử này . tiếp tục so sánh ki tự tại y[i+j-k] and x[i-k]

Ta nhận thấy trong tất cả các trường hợp thông tin cân thiết duy nhất là độ dài hậu tố x xuất hiện tại vị trí i.

Aposltolico - Giancarlo sử dụng 2 cấu trúc dữ liệu :

\* Một mảng skip có thể cập nhập sau mỗi lần thử mẫu tại vị trí

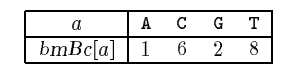
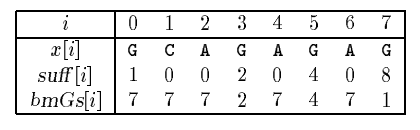
j:skip[j+m-1]=max{ k : x[m-k .. m-1]=y[j+m-k .. j+m-1]}

\* Một mảng suff thỏa mãn : for 1 i < msuff[i]=max{k : x[ik+1 .. i]=x[m-k .. m-1]}

* *Kiểm nghiệm thuật toán:*

X = “GCAGAGAG”

Y =”GCATCGCAGAGAGTATACAGTACG”

Pha tiên xử lý xác định :

Pha tìm kiếm thực hiện như sau :

G C A T C G C **A** G A G A G T A T A C A G T A C G

1

G C A G A G A **G**

Shift by: 1 (bmGs[7]=bmBc[a]-8+8)

G C A T C G **C** A G A G A G T A T A C A G T A C G

3 2 1

G C A G A **G** A G

Shift by: 4 (bmGs[5]=bmBc[c]-8+6)

G C A T C G C A G A G A G T A T A C A G T A C G

6 5 4 3 2 1

G C A G A G A G

Shift by: 7 (bmGs[0])

G C A T C G C A G A G A G T A T A **C** A G T A C G

3 2 1

G C A G A **G** A G

Shift by: 4 (bmGs[5]=bmBc[c]-8+6)

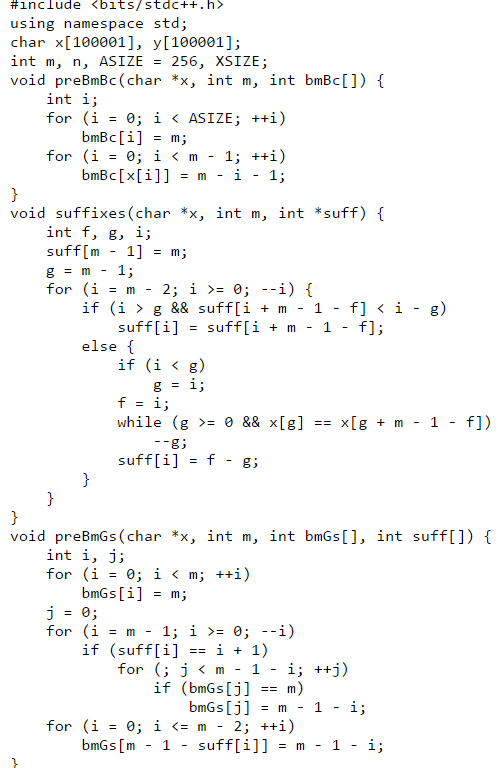
G C A T C G C A G A G A G T A T A C A G T A **C** G

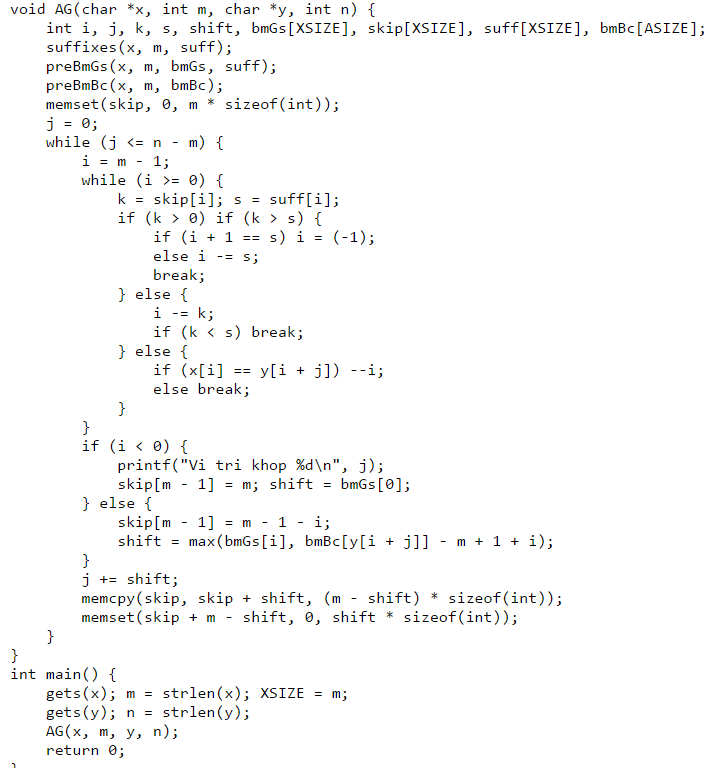
2 1

G C A G A G **A** G

Shift by: 7 (bmGs[6])

* *Chương trình chạy:*

1. 



# 2.3. Trình bày các thuật toán tìm kiếm mẫu từ vị trí cụ thể:

## 2.3.1. Colussi algorithm:

* *Đặc điểm:*

- Đây là thuật toán cái tiến từ thuật toán Knuth, Morris and Pratt.

- Mẫu tìm kiếm sẽ được chia thành 2 tập con. Tập thứ nhất sẽ được duyệt từ trái sang phải khi không gặp mismatch ta duyệt tập thứ hai từ trái qua phải.

- Pha xử lý có độ phức tạp là O(m).

- Pha tìm kiếm có độ phức tạp là O(n).

- Trong trường hợp xấu nhất quá trình thực thi sẽ mất 3/2 n lần so sánh kí tự.

* *Trình bày thuật toán:*

Thuật toán Colussi được thiết kế dựa theo những phân tích từ thuật toán knuth, Morris, và Pratt. các vị trí kí tự trong mẫu được chia thành 2 tập riêng biệt. mỗi phép thử sẽ bao gồm 2 pha :

- Trong pha thứ nhất thực hiên so sánh từ trái sang phải lần lượt các kí tự tại những vị trí kmpNext > -1. Những vị trí này gọi là noholes.

- Pha thứ hai so sánh kí tự từ phải sang trái( các kí tự còn lại ) nhưng kí tự này có kmpNext = -1 và gọi là holes. Chiến lược này có 2 lợi thế:

\* Khi gặp mismath trong pha thứ nhất, sau khi thực hiện shift bước dịch ta không cần phải so sánh lại các kí tự tại những vị trí noholes phía trước.

\* Khi gặp mismath trong pha thứ hai có nghĩa là đoạn hậu tố đã matches tất cả kí tự định nghĩa biến :

For 0 <= i <= m-1:

kmin[i] = d>0 iff x[0 .. i-1-d]=x[d .. i-1] và x[i-d] x[i],

kmin[i] = 0 trong các trường hợp còn lại.

For 0 <= i < m:

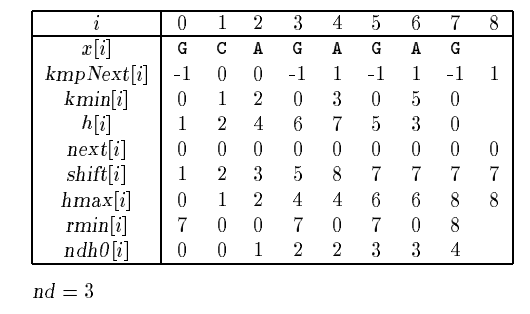
Nếu kmin[i] # 0 thì i là nohole ngược lại i là hole.

* *Kiểm nghiệm thuật toán:*

X = “GCAGAGAG”

Y =”GCATCGCAGAGAGTATACAGTACG”

Pha tiền xử lý xác định:



Pha tìm kiếm:

G C A T **C** G C A G A G A G T A T A C A G T A C G

1 2 3

G C A G **A** G A G

Shift by: 3 (shift[2])

G C A T C **G** C A G A G A G T A T A C A G T A C G

1 2

G C **A** G A G A G

Shift by: 2 (shift[1])

G C A T C G C A G A G A G T A T A C A G T A C G

8 1 2 7 3 6 4 5

G C A G A G A G

Shift by: 7 (shift[8])

G C A T C G C A G A G A G **T** A T A C A G T A C G

1

G **C** A G A G A G

Shift by: 1 (shift[0])

G C A T C G C A G A G A G T **A** T A C A G T A C G

1

G **C** A G A G A G

Shift by: 1 (shift[0])

G C A T C G C A G A G A G T A **T** A C A G T A C G

1

G **C** A G A G A G

Shift by: 1 (shift[0])

G C A T C G C A G A G A G T A T **A** C A G T A C G

1

G **C** A G A G A G

Shift by: 1 (shift[0])

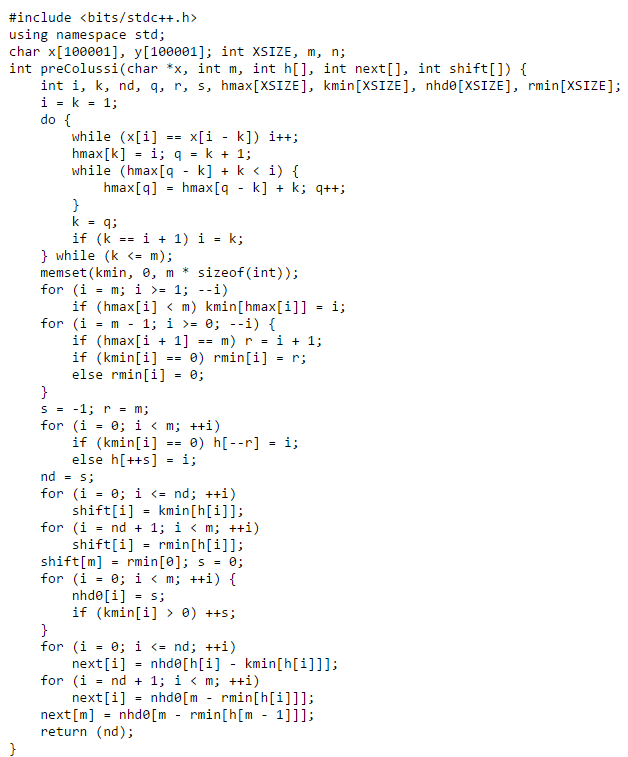
G C A T C G C A G A G A G T A T A C A G **T** A C G

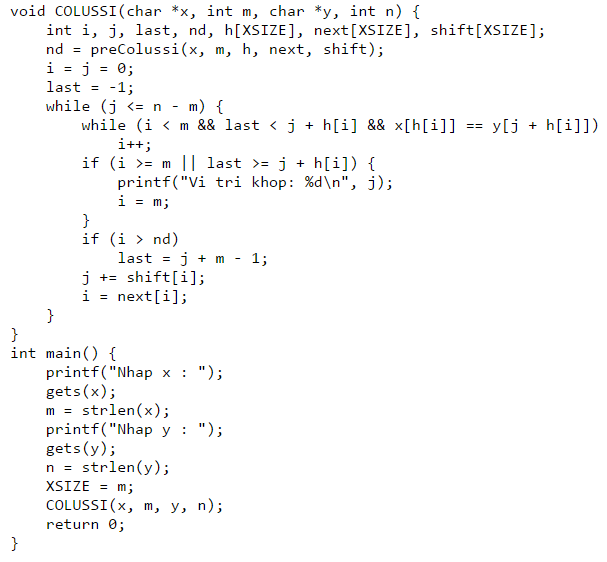
1 2 3

G C A G **A** G A G

Shift by: 3 (shift[2])

* *Chương trình chạy:*





## 2.3.2 Skip Search Algorithm

* *Đặc điểm:*

- Sử dụng thùng chứa (bucket) các vị trí xuất hiện của kí tự trong xâu mẫu.

- Pha xử lý có độ phức tạp thời gian và không gian chứa O(m+ ∂).

- Pha tìm kiếm có độ phức tạp thời gian O(mn).

* *Trình bày thuật toán:*

Với mỗi kí tự trong bảng chữ cái, một thùng chứa (bucket) sẽ chứa tất cả các vị trí xuất hiện của kí tự đó trong xâu mẫu x. khi một kí tự xuất hiện k lần trong mẫu. bucket sẽ lưu k vị trí của kí tự đó. Khi mà xâu y chứa it kí tự hơn trong bản chữ cái thì sẽ có nhiều bucket rỗng. Quá trình xử lý của thuật toán Skip Search bao gồm việc tính các buckets cho tất cả các kí tự trong bảng chữ cái: for c in z[c] = {i : 0<= i <= m-1 và x[i] = c}

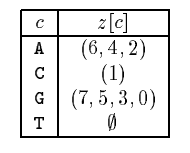
Thuật toán Skip Search có độ phức tạp bình phương trong trường hợp tồi nhất. Nhưng cũng có trường hợp là O(n).

* *Kiểm nghiệm thuật toán:*

X = “GCAGAGAG”

Y = ”GCATCGCAGAGAGTATACAGTACG”

Pha tiền xử lý xác định:



Pha tìm kiếm:

First attempt

G **C** A T C G C A G A G A G T A T A C A G T A C G

2 1

**G** C A G A G A G

G C A **T** C G C A G A G A G T A T A C A G T A C G

1

**G** C A G A G A G

G C A T C G C A G A G A G T A T A C A G T A C G

1 2 3 4 5 6 7 8

G C A G A G A G

Shift by: 8

Second attempt

G C A T C G C A G A G A G **T** A T A C A G T A C G

1

Shift by: 8

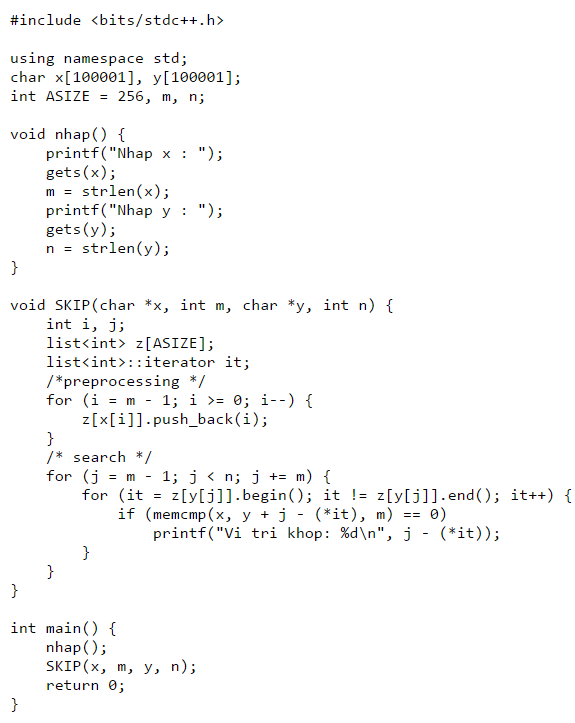
Third attempt

G C A T C G C A G A G A G T A T **A** C A G T A C G

2 1

**G** C A G A G A G

* *Chương trình chạy:*



# 2.4. Trình bày các thuật toán tìm kiếm mẫu từ vị trí bất kỳ:

## 2.4.1. Horspool algorithm:

* *Đặc điểm:*

- Là thuật toán đơn giản hơn của Boyer Moore.

- Sử dụng dụng bad – character.

- Dễ để cài đặt.

- Pha cài đặt có độ phức tạp thuật toán là O(m+σ) và độ phức tạp bộ nhớ là O(σ).

- Pha thực thi có độ phức tạp là O(mxn).

- Số lượng so sánh trung bình cho một văn bản là trong khoảng 1/ σ và 2/( σ +1).

* *Trình bày thuật toán:*

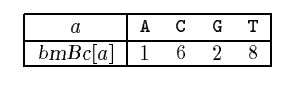
Quy tắc dịch bad- character được sử dụng trong Boyer Moore không có hiệu quả trong đoạn văn bản kí tự nhỏ, nhưng khi đoạn văn bản là lớn và phải so sánh với mẫu dài, nó thường trong trường hợp bảng bã ASCII và tìm kiếm thông thường trong soạn thảo văn bản, chúng rất h ữu dụng. Sử dụng nó như những kết quả riêng biệt lại rất hiệu quả.

* *Kiểm nghiệm thuật toán:*

X = “GCAGAGAG”

Y =”GCATCGCAGAGAGTATACAGTACG”

Pha tiền xử lí:



Pha tìm kiếm thực hiện :

G C A T C G C **A** G A G A G T A T A C A G T A C G

1

G C A G A G A **G**

Shift by: 1 (bmBc[A])

G **C** A T C G C A G A G A G T A T A C A G T A C G

2 1

**G** C A G A G A G

Shift by: 2 (bmBc[G])

G C A **T** C G C A G A G A G T A T A C A G T A C G

2 1

**G** C A G A G A G

Shift by: 2 (bmBc[G])

G C A T C G C A G A G A G T A T A C A G T A C G

2 3 4 5 6 7 8 1

G C A G A G A G

Shift by: 2 (bmBc[G])

G C A T C G C A G A G A G T **A** T A C A G T A C G

1

G C A G A G A **G**

Shift by: 1 (bmBc[A])

G C A T C G C A G A G A G T A **T** A C A G T A C G

1

G C A G A G A **G**

Shift by: 8 (bmBc[T])

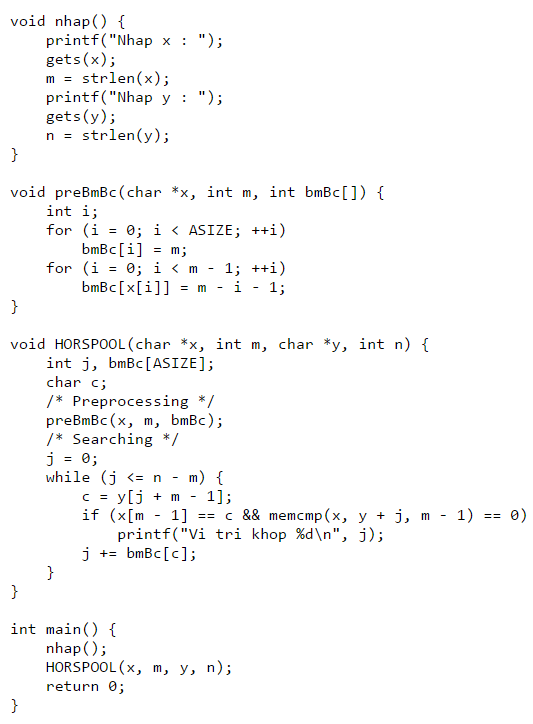
G C A T C G C A G A G A G T A T **A**  C A G T A C G

2 1

**G** C A G A G A G

Shift by: 2 (bmBc[G])

* *Chương trình chạy:*



## 2.4.2. Quick Search algorithm

* *Đặc điểm:*

- Là thuật toán đơn giản của Boyer Moore.

- Chỉ sử dụng dụng bad - character.

- Dễ để cài đặt.

- Pha cài đặt có độ phức tạp thuật toán là O(m+σ) và độ phức tạp bộ nhớ là O(σ).

- Pha thực thi có độ phức tạp là O(mxn).

- Rất nhanh trong thực tế với những mẫu ngắn và anphabets lớn.

* *Trình bày thuật toán:*

Thuật toán Quick Search chỉ sử dụng bảng bad charater. Giả sử trên cửa sổ dịch chuyển đang là đoạn văn bản y[j….j+m-1], độ dài bước dịch đồng đều là 1 bước. Vì vậy, vị trí tiếp theo có liên quan mật thiết tới số bước dịch chuyển. Quick Search sẽ quan tâm tới vị trí y[j+m].

qsBc[c] = min{i : 0 <= i <= m và x[m-i]=c} nếu c có trong x.

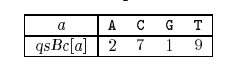
qsBc[c] = m+1 với trường hợp còn lại.

* *Kiểm nghiệm thuật toán:*

X = “GCAGAGAG”

Y =”GCATCGCAGAGAGTATACAGTACG”

Pha tiền xử lí:



Pha tìm kiếm thực hiện:

G C A **T** C G C A G A G A G T A T A C A G T A C G

1 2 3 4

G C A **G** A G A G

Shift by: 1 (qsBc[G])

G **C** A T C G C A G A G A G T A T A C A G T A C G

1

**G** C A G A G A G

Shift by: 2 (qsBc[A])

G C A **T** C G C A G A G A G T A T A C A G T A C G

1

**G** C A G A G A G

Shift by: 2 (qsBc[A])

G C A T C G C A G A G A G T A T A C A G T A C G

1 2 3 4 5 6 7 8

G C A G A G A G

Shift by: 9 (qsBc[T])

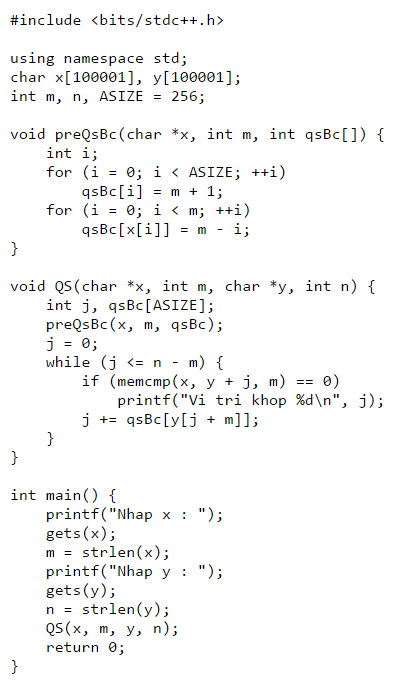
G C A T C G C A G A G A G T **A**  T A C A G T A C G

1

**G**  C A G A G A G

Shift by: 7 (qsBc[C])

* *Chương trình chạy:*



## 2.4.3. Smith algorithm

* *Đặc điểm:*

- Thuật toán Smith tính max của hàm dịch bad charater Horspool và hàm dịch bad charater Quick Search.

- Pha cài đặt có độ phức tạp thuật toán là O(m+σ) và độ phức tạp bộ nhớ là O(σ).

- Pha thực thi có độ phức tạp là O(mxn).

* *Trình bày thuật toán*

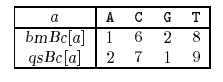
Smith nhận thấy rằng có thể tính toán được rằng số bước dịch với kí tự tiếp theo kí tự đầu mút bên phải của cửa sổ dịch chuyển đôi khi cho số bước dịch ngắn hơn việc sử dụng đúng kí tự đầu mút bên phải cửa sổ ấy. Ông ấy khuyên nên chọn max giữa 2 giá trị.Pha tiền xử lí của thuật toán Smith bao gồm tính toán của hàm dịch chuyển các bad- character: preBmBc(X,m) , và hàm dịch chuyển các badcharacter trong Quick Search: preQsBc(X,m) Pha tìm kiếm có độ phức tạp về thời gian trong trường hợp xấu nhất là một hàm bậc 2.

* *Kiểm nghiệm thuật toán*

X = “GCAGAGAG”

Y =”GCATCGCAGAGAGTATACAGTACG”

Pha tiền xử lí tính toán được:



G C A **T** C G C A G A G A G T A T A C A G T A C G

1 2 3 4

G C A **G** A G A G

Shift by: 1 (bmBc[A]=qsBc[G])

G **C** A T C G C A G A G A G T A T A C A G T A C G

1

**G** C A G A G A G

Shift by: 2 (bmBc[G]=qsBc[A])

G C A **T**  C G C A G A G A G T A T A C A G T A C G

1

**G** C A G A G A G

Shift by: 2 (bmBc[G]=qsBc[A])

G C A T C G C A G A G A G T A T A C A G T A C G

1 2 3 4 5 6 7 8

G C A G A G A G

Shift by: 9 (qsBc[T])

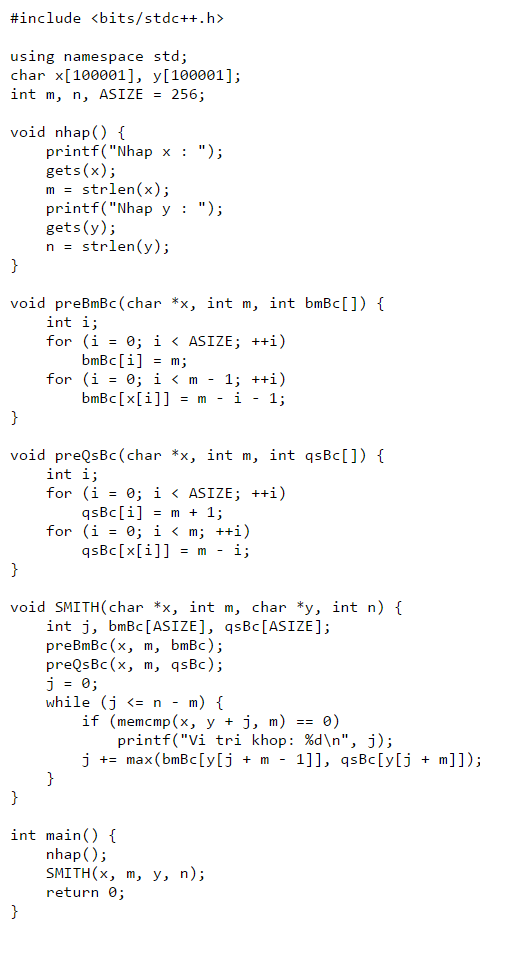
G C A T C G C A G A G A G T **A** T A C A G T A C G

1

**G** C A G A G A G

Shift by: 7 (qsBc[C])

* *Chương trình chạy:*

1. 

## 2.4.4. Raita algorithm

* *Đặc điểm:*

- Đầu tiên so sánh kí tự cuối cùng của xâu mẫu, sau đó là kí tự đầu tiên và kí tự giữa trước khi thực thi lệnh so sánh khác.

- Thực thi bước dịch giống thuật toán Horspool.

- Pha cài đặt có độ phức tạp thuật toán là O(m+σ) và độ phức tạp bộ nhớ là O(σ).

- Pha thực thi có độ phức tạp là O(mxn).

* *Trình bày thuật toán:*

Raita thiết kế một thuật toán mà với mỗi lần duyệt, đầu tiên đem so sánh kí tự cuối cùng của mẫu với kí tự đầu mút bên phải của cửa sổ, sau đó nếu chúng trùng nhau, so sánh kí tự đầu tiên của mẫu với kí tự đầu mút bên trái của cửa sổ. Nếu chúng trùng nhau, tiếp tục so sánh kí tự giữa của mẫu với kí tự giữa của xâu văn bản đang trong cửa sổ duyệt. Nếu chúng vẫn trùng nhau, sẽ đem so sánh các kí tự khác trong xâu mẫu bắt đầu vị trí thứ 2 tới vị trí cuối cùng(so sánh lại kí tự ở giữa một lần nữa).

Raita thấy rằng thuật toán này sử dụng tốt trong những trường hợp: khi cần thực thi với mẫu trong văn bản tiếng anh. Pha tìm kiếm có độ phức tạp về thời gian trong trường hợp xấu nhất là một hàm bậc 2

* *Kiểm nghiệm thuật toán:*

X = “GCAGAGAG”

Y =”GCATCGCAGAGAGTATACAGTACG”

Pha tiền xử lí:



Pha tìm kiếm thực hiện :

G C A T C G C **A** G A G A G T A T A C A G T A C G

1

G C A G A G A **G**

Shift by: 1 (bmBc[A])

G  **C** A T C G C A G A G A G T A T A C A G T A C G

2 1

**G** C A G A G A G

Shift by: 2 (bmBc[G])

G C A  **T** C G C A G A G A G T A T A C A G T A C G

2 1

**G** C A G A G A G

Shift by: 2 (bmBc[G])

G C A T C G C A G A G A G T A T A C A G T A C G

2 4 5 6 3 8 9 1

G C A G A G A G

Shift by: 2 (bmBc[G])

G C A T C G C A G A G A G T  **A** T A C A G T A C G

1

G C A G A G A **G**

Shift by: 1 (bmBc[A])

G C A T C G C A G A G A G T A **T** A C A G T A C G

1

G C A G A G A **G**

Shift by: 8 (bmBc[T])

G C A T C G C A G A G A G T A T **A** C A G T A C G

2 1

**G** C A G A G A G

Shift by: 2 (bmBc[G])

* *Chương trình chạy:*

