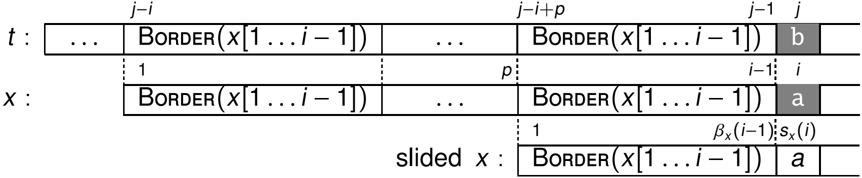
Knuth-Morris-Pratt algorithm

The algorithm we present now is an improvement due to Knuth of the Morris-Pratt algorithm, based on avoiding situations which lead to certain letter comparison failures.



The observation is that if $a = \mathrm{a}$ then the sliding would lead to a comparison failure. So let us enforce that a sliding to compare $x[s_x(i)]$ to $t[j]$ is done if and only if $x[s_x(i)] \neq x[i]$.

Knuth-Morris-Pratt/Maximum disjoint borders

If $x[s_x(i)] = x[i]$, what should we do?

Let us note $y = x[1...i-1]$. We should consider successively $\textsc{Border}^2(y) \cdot a$, $\textsc{Border}^3(y) \cdot a$ etc., until we find a k such that $\textsc{Border}^k(y) \cdot a \not\preccurlyeq x$ or $\textsc{Border}^k(y) = \varepsilon$. Such a border $\textsc{Border}^k(y)$ is called a **maximum disjoint border** of y in x.

If it is non-empty, we slide the word so that we compare $x[s_x^k(i)]$ and $t[j]$.

This disjointedness constraint can be precomputed on the pattern x *alone* before comparing it to the text t: all we have to do is to change the failure function s_x of the algorithm of Morris and Pratt and provide a new one. *The search algorithm itself does not change.*

Knuth-Morris-Pratt/Maximum disjoint borders (cont)

The disjointedness implicitly entails that it is a relative concept: we consider the disjoint border of a *proper prefix* of another word. Thus, the letter with follows the prefix is used as a constraint, i.e., it must not follow the disjoint border.

Let $ya \preccurlyeq x$. Let us note $\overline{\text{Border}}_x(y)$ the maximum disjoint border of y *in* x. In this case, the letter a is used to constrain the definition of the disjoint border, as we must have $\overline{\text{Border}}_x(y) \cdot a \not\preccurlyeq y$.

What if $y = x$, then? In this case, there is no right-context, like the letter a above, to constrain the maximum disjoint border. In this case, we still can take the maximum border, i.e., $\overline{\text{Border}}_y(y) = \text{Border}(y)$.

Knuth-Morris-Pratt/Maximum disjoint borders (cont)

More precisely, if the maximum border of y is not followed by a, then it is also the maximum disjoint border we are looking for. In other words:

$$\overline{\textsc{Border}}_x(y) = \textsc{Border}(y) \qquad \text{if } ya \preccurlyeq x \text{ and } \textsc{Border}(y) \cdot a \not\preccurlyeq y$$

If the maximum border of y is followed by a we must take the maximum disjoint border of the maximum border:

$$\overline{\textsc{Border}}_x(y) = \overline{\textsc{Border}}_x(\textsc{Border}(y)) \qquad \text{if } ya \preccurlyeq x \text{ and } \textsc{Border}(y) \cdot a \preccurlyeq y$$

By extension, if $x = y$, we take the maximum border:

$$\overline{\textsc{Border}}_y(y) = \textsc{Border}(y)$$

Knuth-Morris-Pratt/Maximum disjoint borders (cont)

In summary, assuming $ya \preccurlyeq x$ **and** $y \neq \varepsilon$

$$\overline{\textsc{Border}}_y(y) = \textsc{Border}(y)$$

$$\overline{\textsc{Border}}_x(y) = \begin{cases} \overline{\textsc{Border}}_x(\textsc{Border}(y)) & \text{if } \textsc{Border}(y) \cdot a \preccurlyeq y \\ \textsc{Border}(y) & \text{otherwise} \end{cases}$$

Knuth-Morris-Pratt/Maximum disjoint borders (cont)

By taking the length of each side of the equations,

$$|\overline{\textsc{Border}}_x(y)| = \begin{cases} |\textsc{Border}(y)| & \text{if } x = y \text{ or } \textsc{Border}(y) \cdot a \not\preccurlyeq y \\ |\overline{\textsc{Border}}_x(\textsc{Border}(y))| & \text{otherwise} \end{cases}$$

Let $\gamma_x(i)$ be the length of the disjoint maximum border of $x[1 \ldots i]$:

$$\gamma_x(i) = |\overline{\textsc{Border}}_x(x[1...i])| \qquad 1 \leqslant i \leqslant |x|$$

or, equivalently,

$$\gamma_x(|y|) = |\overline{\textsc{Border}}_x(y)| \qquad y \preccurlyeq x$$

Knuth-Morris-Pratt/Maximum disjoint borders (cont)

Therefore

$$\gamma_x(|y|) = \begin{cases} \beta_x(|y|) & \text{if } x = y \text{ or } \textsc{Border}(y) \cdot a \not\preceq y \\ \gamma_x(|\textsc{Border}(y)|) & \text{otherwise} \end{cases}$$

that is to say

$$\gamma_x(|y|) = \begin{cases} \beta_x(|y|) & \text{if } x = y \text{ or } \textsc{Border}(y) \cdot a \not\preceq y \\ \gamma_x(\beta_x(|y|)) & \text{otherwise} \end{cases}$$

If $|y| = i$, we can write instead

$$\gamma_x(i) = \begin{cases} \beta_x(i) & \text{if } x = y \text{ or } \textsc{Border}(y) \cdot a \not\preceq y \\ \gamma_x(\beta_x(i)) & \text{otherwise} \end{cases}$$

Knuth-Morris-Pratt/Maximum disjoint borders (cont)

The condition can also be rewritten in terms of β_x and i, just as we did with the Morris-Pratt algorithm, pages 35 and 19. Let $|y| = i$, then

$$\textsc{Border}(y) \cdot a \not\leqslant y \iff x[\beta_x(i) + 1] \neq x[i + 1]$$
$$x = y \iff |x| = i$$

So, finally, for $1 \leqslant i \leqslant |x|$,

$$\gamma_x(i) = \begin{cases} \beta_x(i) & \text{if } i = |x| \text{ or } x[\beta_x(i) + 1] \neq x[i + 1] \\ \gamma_x(\beta_x(i)) & \text{otherwise} \end{cases}$$

which allows us to naturally extends γ_x on 0: $\gamma_x(0) = \beta_x(0) = -1$.

Knuth-Morris-Pratt/Maximum disjoint borders (cont)

Before going further, let us check by hand the following values of $\gamma_x(i)$ and compare them to $\beta_x(i)$: we must always have $\gamma_x(i) \leqslant \beta_x(i)$, since we plan an optimisation.

x		a	b	c	a	b	a	b	c	a	c
i	0	1	2	3	4	5	6	7	8	9	10
$\beta_x(i)$	−1	0	0	0	1	2	1	2	3	4	0
$\gamma_x(i)$	−1	0	0	−1	0	2	0	0	−1	4	0

One difference between β_x and γ_x is that, in the worst case, there is always an empty maximum border ε, i.e., $\beta_x(i) = 0$, whereas there may be no maximum disjoint border at all, i.e., $\gamma_x(i) = -1$. For instance, the prefix abc of x has an empty maximum border, i.e., $\beta_x(3) = 0$, but has no maximum disjoint border, i.e., $\gamma_x(3) = -1$, since $x[1] = x[4]$.

Knuth-Morris-Pratt/Maximum disjoint borders (cont)

This definition of γ_x relies on β_x, more precisely, the computation of $\gamma_x(i)$ requires $\beta_x^p(i)$, i.e., values $\beta_x(j)$ with $j < i$, since

$$\beta_x(0) = -1 \qquad \beta_x(i) = 1 + \beta_x^k(i-1) \qquad 1 \leqslant i$$

where k is the smallest non-zero integer such that

- either $1 + \beta_x^k(i-1) = 0$ or $x[1 + \beta_x^k(i-1)] = x[i]$

or, equivalently,

$$\textsc{Border}(ya) = \begin{cases} \textsc{Border}^k(y) \cdot a & \text{if } \textsc{Border}^k(y) \cdot a \preccurlyeq y \\ \varepsilon & \text{if } \textsc{Border}^{k-1}(y) = \varepsilon \end{cases}$$

where k is the smallest non-zero integer such that $\textsc{Border}^k(y) \cdot a \preccurlyeq y$ or $\textsc{Border}^{k-1}(y) = \varepsilon$.

Knuth-Morris-Pratt/Maximum disjoint borders (cont)

Therefore, let us find another definition of $\textsc{Border}(ya)$ which relies on $\textsc{Border}(y)$ but **not** on $\textsc{Border}^q(y)$ with $2 \leqslant q$. This can be achieved by considering again the figure

$y \cdot a$:	$\textsc{Border}(y)$	b		$\textsc{Border}(y)$	a

Indeed, if $a = b$ then $\textsc{Border}(ya) = \textsc{Border}(y)$. Else, $a \neq b$, and thus the maximum border can be found among the *disjoint* borders of $\textsc{Border}(y)$, otherwise $\textsc{Border}(ya) = \varepsilon$.

$$\textsc{Border}(ya) = \begin{cases} \overline{\textsc{Border}}_x^{\,q}(\textsc{Border}(y)) \cdot a & \text{if } \overline{\textsc{Border}}_x^{\,q}(\textsc{Border}(y)) \cdot a \leqslant y \\ \varepsilon & \text{if } y = \varepsilon \text{ or } \overline{\textsc{Border}}_x^{\,q-1}(\textsc{Border}(y)) = \varepsilon \end{cases}$$

where $ya \leqslant x$ and q is the smallest integer such that $\overline{\textsc{Border}}_x^{\,q}(\textsc{Border}(y)) \cdot a \leqslant y$ or $\overline{\textsc{Border}}_x^{\,q-1}(\textsc{Border}(y)) = \varepsilon$.

Knuth-Morris-Pratt/Maximum disjoint borders (cont)

By taking the lengths and letting $0 \leqslant i \leqslant |x| - 1$ and $|y| = i$,

$$\beta_x(0) = -1$$

$$\beta_x(i+1) = \begin{cases} \gamma_x^q(\beta_x(i)) + 1 & \text{if } x[\gamma_x^q(\beta_x(i)) + 1] = x[i+1] \\ 0 & \text{if } i = 0 \text{ or } \gamma_x^{q-1}(\beta_x(i)) = 0 \end{cases}$$

where q is the smallest integer such that

- either $x[\gamma_x^q(\beta_x(i)) + 1] = x[i+1]$
- or $\gamma_x^{q-1}(\beta_x(i)) = 0$

Knuth-Morris-Pratt/Maximum disjoint borders (cont)

Since $\gamma_x(i) = -1$ if and only if $i = 0$, we have

$$\gamma_x^{q-1}(\beta_x(i)) = 0 \iff \gamma_x(\gamma_x^{q-1}(\beta_x(i))) = \gamma_x(0) \iff \gamma_x^{q}(\beta_x(i)) = -1$$
$$\iff \gamma_x^{q}(\beta_x(i)) + 1 = 0$$

Therefore we can simplify the new definition of β_x as

$$\begin{aligned} \beta_x(0) &= -1 \\ \beta_x(1) &= 0 \\ \beta_x(i+1) &= \gamma_x^{q}(\beta_x(i)) + 1 \qquad 1 \leqslant i \leqslant |x| - 1 \end{aligned}$$

where q is the smallest integer such that

- either $x[\gamma_x^{q}(\beta_x(i)) + 1] = x[i+1]$
- or $\gamma_x^{q}(\beta_x(i)) + 1 = 0$

Knuth-Morris-Pratt/Maximum disjoint borders/Imperative

GAMMA(x)

$\gamma_x[0] \leftarrow -1$; $offset \leftarrow -1$ $\quad \triangleright\ offset = \beta_x(0)$

for $i \leftarrow 1$ **to** $|x| - 1$

  do repeat

      $offset \leftarrow \gamma_x[offset]$

    until $offset = -1$ **or** $x[offset+1] = x[i]$

   $offset \leftarrow offset + 1$ $\quad \triangleright\ offset = \gamma_x^q(\beta_x(i-1)) + 1 = \beta_x(i)$

   if $i = |x|$ **or** $x[offset+1] = x[i+1]$

    then $\gamma_x[i] \leftarrow offset$ $\quad \triangleright\ \gamma_x(i) = \beta_x(i)$

    else $\gamma_x[i] \leftarrow \gamma_x[offset]$ $\quad \triangleright\ \gamma_x(i) = \gamma_x(\beta_x(i))$

GAMMA $\leftarrow \gamma_x$

Knuth-Morris-Pratt/Better failure function

With the algorithm of Morris and Pratt, it was convenient to use a failure function $s(i) = 1 + \beta(i - 1)$. Similarly, we need here another failure function, noted r, such that $r(i) = 1 + \gamma(i - 1)$, for $1 \leqslant i$.

In order to finish, we need to give the algorithm for the failure function r.

We could then keep Beta and create a separate function KMP-Fail simply by sticking to the above definition:

KMP-Fail(γ_x, i)
KMP-Fail $\leftarrow 1 + \gamma_x[i - 1]$ $\qquad \triangleright\ r_x[i] \leftarrow 1 + \gamma_x[i - 1]$

But this is not efficient: it would be better to precompute and store all the values of KMP-Fail in an array.

Knuth-Morris-Pratt/Better failure function (cont)

Let us recall the definition of γ_x. Let $1 \leqslant i \leqslant |x|$ and

$$\gamma_x(0) = -1$$
$$\gamma_x(i) = \begin{cases} \beta_x(i) & \text{if } i = |x| \text{ or } x[\beta_x(i)+1] \neq x[i+1] \\ \gamma_x(\beta_x(i)) & \text{otherwise} \end{cases}$$

Then, for $1 \leqslant i \leqslant |x| - 1$

$$r_x(1) = 1 + \gamma_x(0) = 0$$
$$r_x(1+i) = 1 + \gamma_x(i) = \begin{cases} 1 + \beta_x(i) & \text{if } x[\beta_x(i)+1] \neq x[i+1] \\ 1 + \gamma_x(\beta_x(i)) & \text{otherwise} \end{cases}$$

Knuth-Morris-Pratt/Better failure function (cont)

This can be slightly simplified into

$$
\begin{aligned}
r_x(1) &= 0 \\
r_x(1+i) &= \begin{cases} r_x(1+\beta_x(i)) & \text{if } x[1+\beta_x(i)] = x[1+i] \\ 1+\beta_x(i) & \text{otherwise} \end{cases}
\end{aligned}
$$

where for $1 \leqslant i \leqslant |x| - 1$.

Now, we need to express β_x in terms of r_x instead of γ_x.

Knuth-Morris-Pratt/Better failure function (cont)

First, let us prove by induction on $0 \leqslant p$ that

$$r_x^p(1+i) = 1 + \gamma_x^p(i) \qquad 0 \leqslant i \leqslant |x| - 1$$

Because

$$r_x^0(1+i) = 1 + \gamma_x^0(i) \Longleftrightarrow 1 + i = 1 + i$$

and, assuming it is true up to p, we have

$$r_x^{p+1}(1+i) = r_x(r_x^p(1+i)) = r_x(1+\gamma_x^p(i)) = 1 + \gamma_x(\gamma_x^p(i)) = 1 + \gamma_x^{p+1}(i)$$

which is the property at rank $p+1$.

Knuth-Morris-Pratt/Better failure function (cont)

So, the definition of β_x

$$\beta_x(0) = -1 \qquad \beta_x(1) = 0 \qquad \beta_x(1+i) = 1+\gamma_x^q(\beta_x(i)) \qquad 1 \leqslant i \leqslant |x|-1$$

where q is the smallest integer such that either $x[\gamma_x^q(1+\beta_x(i))] = x[1+i]$ or $1+\gamma_x^q(\beta_x(i)) = 0$, is equivalent to

$$\beta_x(0) = -1 \qquad \beta_x(1) = 0 \qquad \beta_x(1+i) = r_x^q(1+\beta_x(i)) \qquad 1 \leqslant i \leqslant |x|-1$$

where q is the smallest integer such that

- either $x[r_x^q(1+\beta_x(i))] = x[1+i]$
- or $r_x^q(1+\beta_x(i)) = 0$

Knuth-Morris-Pratt/Better failure function (cont)

Or, by changing $i+1$ into i,

$$\beta_x(0) = -1 \qquad \beta_x(1) = 0 \qquad \beta_x(i) = r_x^q(1+\beta_x(i-1)) \qquad 2 \leqslant i \leqslant |x|$$

where q is the smallest integer such that

- either $x[r_x^q(1+\beta_x(i-1))] = x[i]$
- or $r_x^q(1+\beta_x(i-1)) = 0$

Knuth-Morris-Pratt/Better failure function/Imperative

KMP-FAIL(x)

$r_x[1] \leftarrow 0$; $\mathit{offset} \leftarrow 0$ $\quad \triangleright\ \mathit{offset} = 1 + \beta_x(1-1)$
for $i \leftarrow 1$ **to** $|x| - 1$
  do repeat
      $\mathit{offset} \leftarrow r_x[\mathit{offset}]$
    until $\mathit{offset} = 0$ **or** $x[\mathit{offset}] = x[i]$
   $\mathit{offset} \leftarrow \mathit{offset} + 1$ $\quad \triangleright\ \mathit{offset} = 1 + \beta_x(i)$
   $\triangleright\ \mathit{offset} = 1 + r_x^q(1 + \beta_x(i-1))$
   if $x[\mathit{offset}] = x[1+i]$
    then $r_x[i] \leftarrow r_x[\mathit{offset}]$ $\quad \triangleright\ r_x(i) = r_x(1 + \beta_x(i))$
    else $r_x[i] \leftarrow \mathit{offset}$ $\quad \triangleright\ r_x(i) = 1 + \beta_x(i)$
KMP-FAIL $\leftarrow r_x$

Knuth-Morris-Pratt/The code

The Knuth-Morris-Pratt algorithm is exactly the Morris-Pratt algorithm, except that we use a better failure function r instead of s:

```
KMP(x, t)
r ← KMP-FAIL(x)
i ← 1; j ← 1
while i ⩽ |x| and j ⩽ |t|
     do if i = 0 or x[i] = t[j]
           then i ← i + 1; j ← j + 1
           else i ← r[i]
if |x| < i
   then . . .                ▷ Occurrence of x in t at position j – |x|.
   else . . .                                         ▷ No occurrences.
```

Knuth-Morris-Pratt/The code (cont)

If we allow arguments to be functions, we can elegantly factorise the two text search algorithms into one:

```
SEARCH(x, t, f)
i ← 1; j ← 1
while i ⩽ |x| and j ⩽ |t|
      do if i = 0 or x[i] = t[j]
            then i ← i + 1; j ← j + 1
            else  i ← f[i]
if |x| < i
   then ...                ▷ Occurrence of x in t at position j – m.
   else ...                                        ▷ No occurrences.
```

Knuth-Morris-Pratt/The code (cont)

Then

$$\mathrm{MP}(x,t) = \textsc{Search}(x,t,\mathrm{MP}\text{-}\textsc{Fail})$$
$$\mathrm{KMP}(x,t) = \textsc{Search}(x,t,\mathrm{KMP}\text{-}\textsc{Fail})$$

Knuth-Morris-Pratt/Maximum disjoint borders (cont)

Example. Consider the following table for the search of the word $x = \mathtt{abacabac}$ in the text $t = \mathtt{babacacabacaab}$.

j	1	2	3	4	5	6	7	8	9	10	11	12	13	14
	b	a	b	a	c	a	c	a	b	a	c	a	a	b
i	1	1	2	3	4	5	6	1	2	3	4	5	6	2
	0						1						1	
							0							

Knuth-Morris-Pratt/Example

	1	2	3	4	5	6	7	8	9	10	11	12	13	14						
t :	b	a	b	a	c	a	c	a	b	a	c	a	a	b						
x :	a	b	a	c	a	b	a	c												
		a	b	a	c	a	b	a	c											
							a	b	a	c	a	b	a	c						
								a	b	a	c	a	b	a	c					
													a	b	a	c	a	b	a	c

Comparison of Morris-Pratt and Knuth-Morris-Pratts

For example, here is the table comparing the values of the two failure functions for the same pattern:

x	a	b	a	c	a	b	a	c
i	1	2	3	4	5	6	7	8
$s_x(i)$	0	1	1	2	1	2	3	4
$r_x(i)$	0	1	0	2	0	1	0	2