
Leidsche Flesch Lunchlezing
KMP
patroonherkenning
Hendrik Jan Hoogeboom
LIACS
26 september 2012

Enterobacteria phage MS2, complete genome (GenBank: EF204940.1)

```
gggtgggacc cctttcgggg tcctgctcaa cttcctgtcg agctaatgcc atttttaatg tctttagcga gacgctacca tggctatcgc
tgtaggtagc cggaattcca ttcctaggag gtttgacctg tgcgagcttt tagtaccctt gatagggaga acgagacctt cgtcccctcc
gttcgcgttt acgcggacgg tgagactgaa gataactcat tctctttaaa atatcgttcg aactggactc ccggtcgttt taactcgact
ggggccaaaa cgaaacagtg gcactacccc tctccgtatt cacggggggc gttaagtgtc acatcgatag atcaaggtgc ctacaagcga
agtgggtcat cgtggggtcg cccgtacgag gagaaagccg gtttcggctt ctccctcgac gcacgctcct gctacagcct cttccctgta
agccagaact tgacttacat cgaagtgccg cagaacgttg cgaaccgggc gtcgaccgaa gtcctgcaaa aggtcaccca gggtaatttt
aaccttggtg ttgctttagc agaggccagg tcgacagcct cacaactcgc gacgcaaacc attgcgctcg tgaaggcgta cactgccgct
cgtcgcggta attggcgcca ggcgctccgc taccttgccc taaacgaaga tcgaaagttt cgatcaaaac acgtggccgg caggtggttg
gagttgcagt tcggttggtt accactaatg agtgatatcc agggtgcata tgagatgctt acgaaggttc accttcaaga gtttcttcct
atgagagccg tacgtcaggt cggtactaac atcaagttaa atggccgtct gtcgtatcca gctgcaaact tccagacaac gtgcaacata
tcgcgacgta tcgtgatatg gttttacata aacgatgcac gtttggcatg gttgtcgtct ctaggtatct tgaacccact aggtatagtg
tgggaaaagg tgcctttctc attcgttgtc gactggctcc tacctgtagg taacatgctc gagggcctta cggcccccgt gggatgctcc
tacatgtcag gaacagttac tgacgtaata acgggtgagt ccatcataag cgttgacgct ccctacgggt ggactgtgga gagacagggc
actgctaagg cccaaatctc agccatgcat cgaggggtac aatccgtatg gccaacaact ggcgcgtacg taaagtctcc tttctcgatg
gtccatacct tagatgcgtt agcattaatc aggcaacggc tctctagata gagccctcaa ccggagtttg aagcatggct tctaacttta
ctcagttcgt tctcgtcgac aatggcggaa ctggcgacgt gactgtcgcc ccaagcaact tcgctaacgg ggtcgctgaa tggatcagct
ctaactcgcg ttcacaggct tacaaagtaa cctgtagcgt tcgtcagagc tctgcgcaga atcgcaaata caccatcaaa gtcgaggtgc
ctaaagtggc aacccagact gttggtggtg tagagcttcc tgtagccgca tggcgttcgt acttaaatat ggaactaacc attccaattt
tcgctacgaa ttccgactgc gagcttattg ttaaggcaat gcaaggtctc ctaaaagatg gaaacccgat tccctcagca atcgcagcaa
actccggcat ctactaatag acgccggcca ttcaaacatg aggattaccc atgtcgaaga caacaaagaa gttcaactct ttatgtattg
atcttcctcg cgatctttct ctcgaaattt accaatcaat tgcttctgtc gctactggaa gcggtgatcc gcacagtgac gactttacag
caattgctta cttaagggac gaattgctca caaagcatcc gaccttaggt tctggtaatg acgaggcgac ccgtcgtacc ttagctatcg
ctaagctacg ggaggcgaat gatcggtgcg gtcagataaa tagagaaggt ttcttacatg acaaatcctt gtcatgggat ccggatgttt
tacaaaccag catccgtagc cttattggca acctcctctc tggctaccga tcgtcgttgt ttgggcaatg cacgttctcc aacggtgcct
ctatggggca caagttgcag gatgcagcgc cttacaagaa gttcgctgaa caagcaaccg ttacccccgg cgctctgaga gcggctctat
tggtccgaga ccaatgtgcg ccgtggatca gacacgcggt ccgctataac gagtcatatg aatttaggct cgttgtaggg aacggagtgt
ttacagttcc gaagaataat aaaatagatc gggctgcctg taaggagcct gatatgaata tgtacctcca gaaaggggtc ggtgccttta
tcagacgccg gctcaaatcc gttggtatag atctgaatga tcaatcgatc aaccagcttc tggctcagca gggcagcgta gatggttcgc
ttgcgacgat agacttatcg tctgcatccg attccatctc cgatcgcctg gtgtggagtt ttctcccacc tgagctatat tcatatctcg
atcgtatccg ctcacactac ggaatcgtag atggcgagac gatacgatgg gaactatttt ccacaatggg aaatgggttc acgtttgagc
tagagtccat gatattctgg gcaatagtca aagcgaccca aatccatttt ggtaacgccg gaaccatagg catctacggg gacgatatta
tatgtcccag tgagattgca ccccgtgtgc tggaggcact tgcctactac ggtttcaaac cgaatctccg taaaacgttc gtgtccgggc
tctttcgcga gagctgcggc gcgcactttt accgtggtgt cgatgtcaaa ccgttttaca tcaagaaacc tgttgacaat ctcttcgccc
ttatgctgat attgaatcgg ctacggggtt ggggagttgt cggaggtatg tcagatccac gcctttacaa ggtgtgggta cgactctcct
cccaggtgcc ttcgatgttt ttcggtggga cggacctcgc tgccgactac tacgtagtca gcccgcccac ggcagtctcg gtatatacca
agactccgta tgggcggcta ctcgcggata cccgtacctc gggtttccgt cttgctcgta tcgctcgaga acgcaagttc ttcagcgaaa
agcatgacag tggccgctac atagcgtggt tccatactgg aggtgaagtc accgacagta tgaagtccgc cggcgtgcgt attatgcgca
cttcggagtg gctaacgccg gttcccacat tccctcagga gtgtgggcca gcgagctctc ctcggtagct gaccgaggga cccccgtaaa
cggggtgggt gtgctcgaaa gagcacgggt ccgcgaaagc ggtggctcca ccgaaaggtg ggcgggcttc ggcccaggga cctccccttg
aagagagggc ccgggattct cccgatttgg taactagctg cttggctagt taccaccca
```

SIAM J. COMPUT.
Vol. 6, No. 2, June 1977

FAST PATTERN MATCHING IN STRINGS*

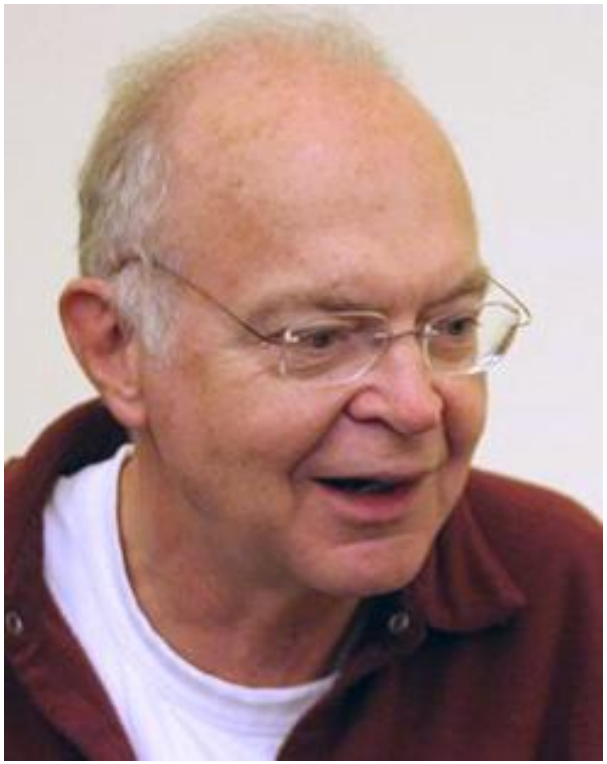
DONALD E. KNUTH†, JAMES H. MORRIS, JR.‡ AND VAUGHAN R. PRATT¶

Abstract. An algorithm is presented which finds all occurrences of one given string within another, in running time proportional to the sum of the lengths of the strings. The constant of proportionality is low enough to make this algorithm of practical use, and the procedure can also be extended to deal with some more general pattern-matching problems. A theoretical application of the

Donald Knuth, James H. Morris, Jr & Vaughan Pratt . Fast pattern matching in strings. SIAM Journal on Computing 6 (1977) 323–350. doi:10.1137/0206024

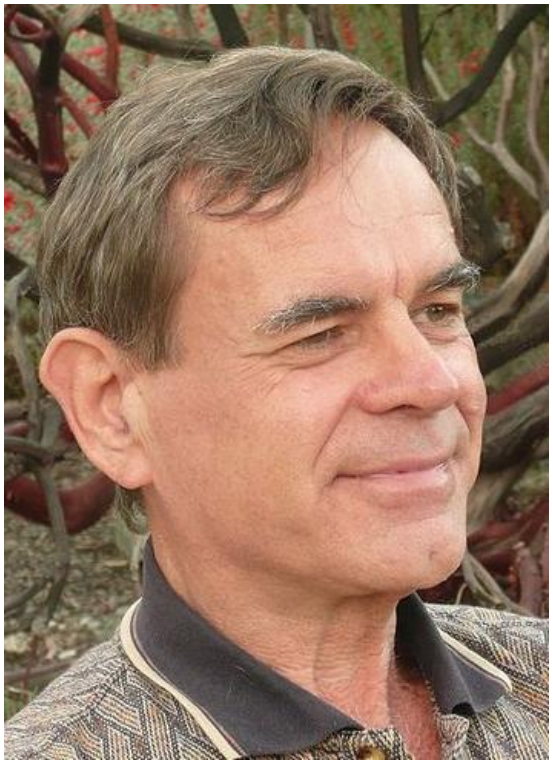
krasse knarren

Donald Knuth (1938-)



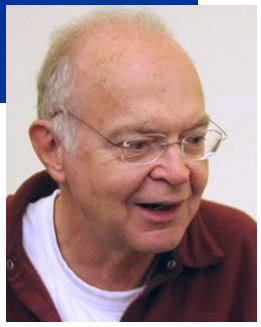
James Morris (1941-)

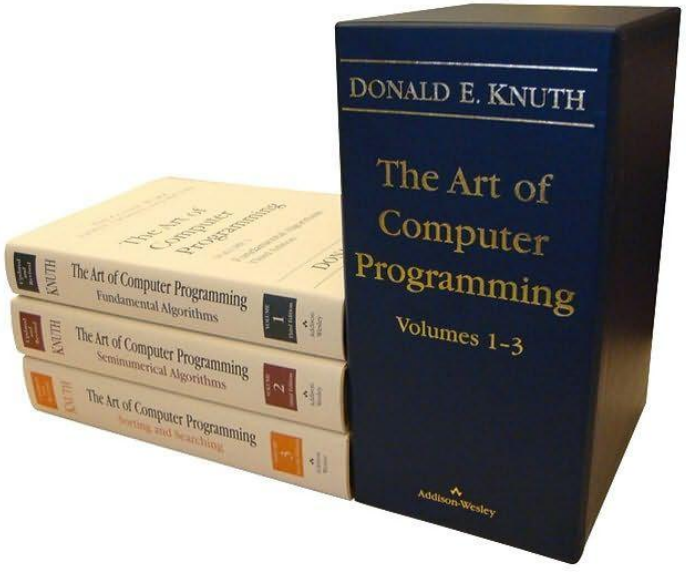
Vaughan Pratt (1944-)



wikipedia (Knut, Pratt) eigen website (Morris)

Donald Knuth





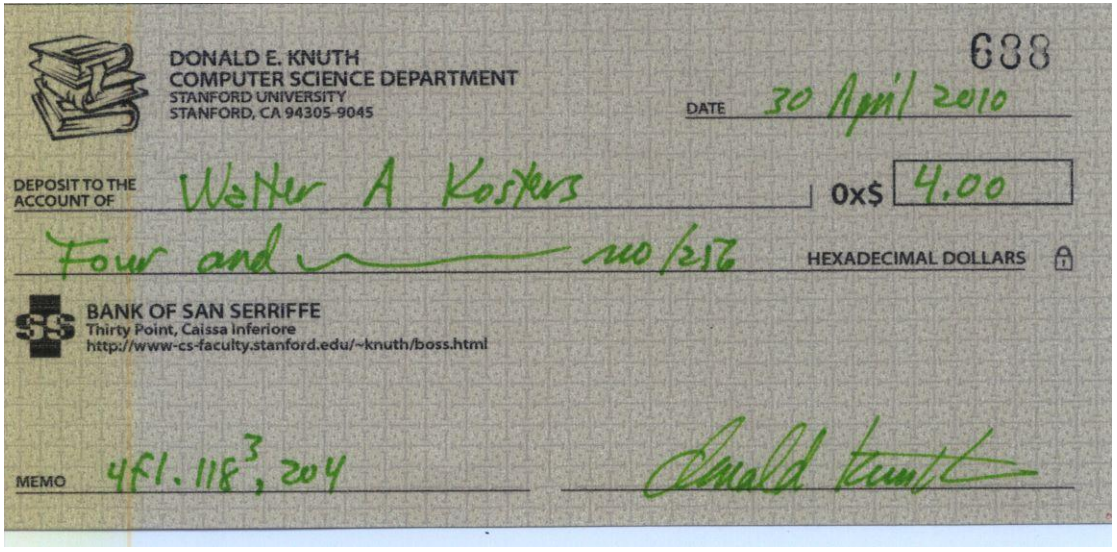
Stanford University

The Art of Computer Programming

géén email

$2.56

TEX



James Morris

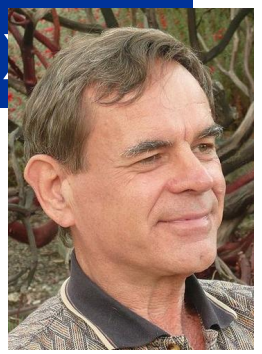
University of California, Berkeley

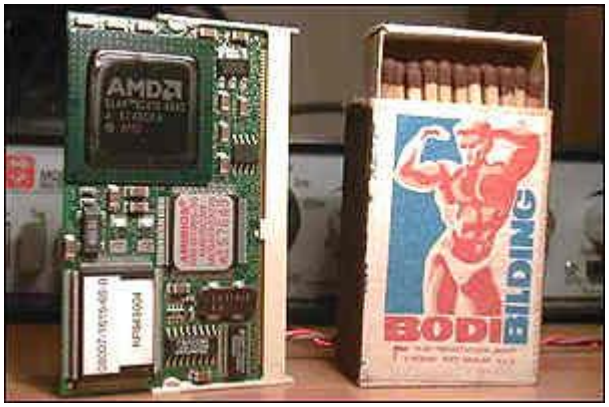
Xerox Alto
ongeveer 40 jaar oud
PC vs. mainframe

muis

“The base machine and one disk were housed in a cabinet about the size of a small refrigerator” (256kB / 2.5MB)

Vaughan Pratt





This computer runs a web site (1999)

Stanford University

PRIMES ∈ NP

SUN

pattern matching

Patroon P =

a b c a b c a c a b

Tekst T =

b a b c b a b c a b c a a b c a b c a b c a c a b

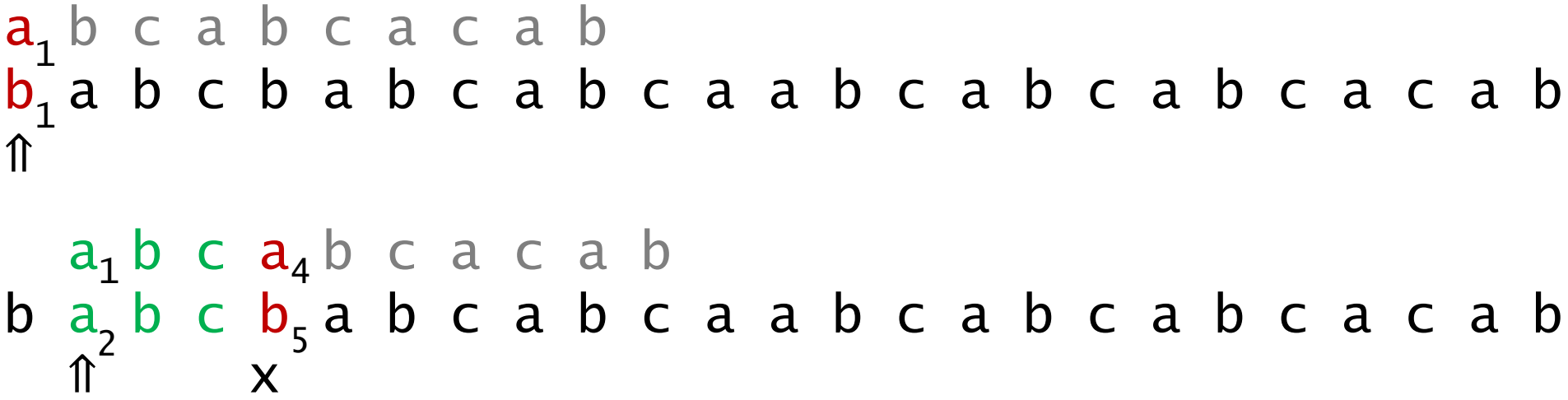
naïeve algoritme

a_1 b c a b c a c a b

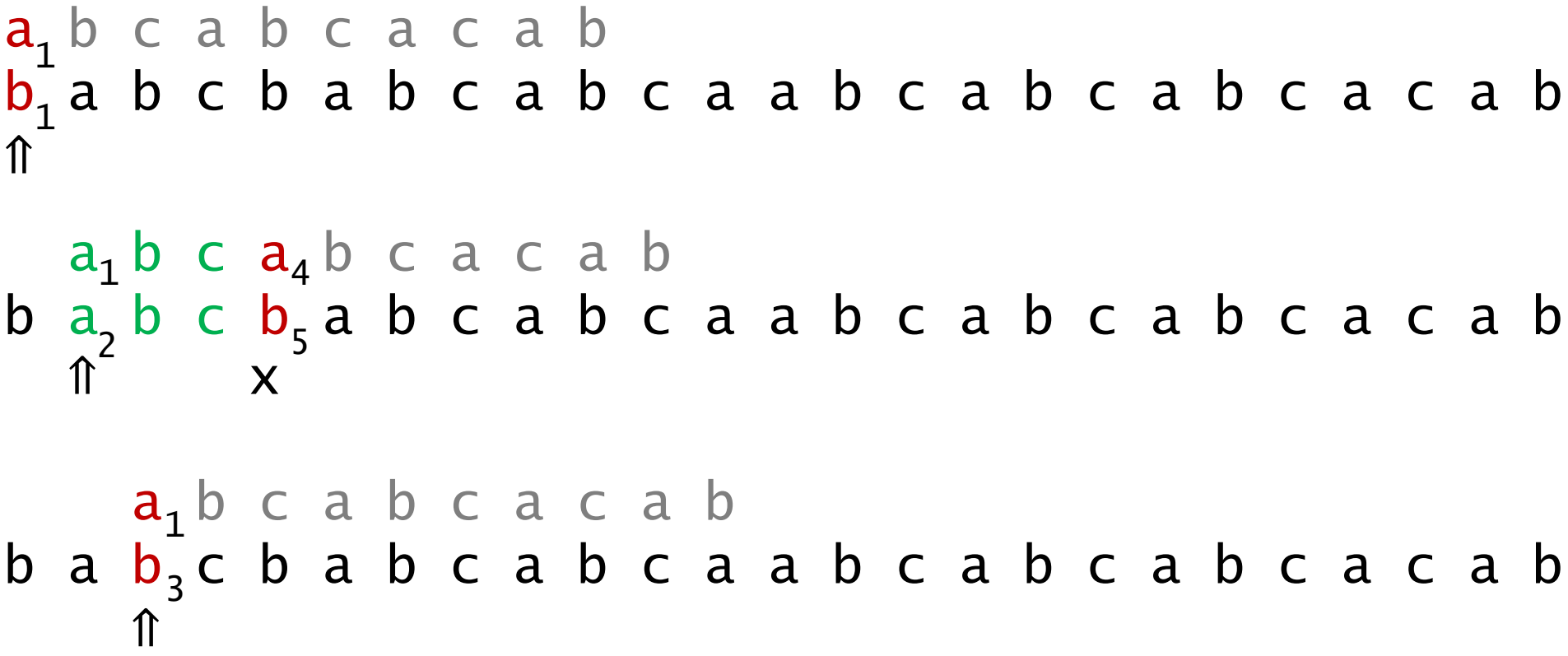
b_1 a b c b a b c a b c a a b c a b c a b c a c a b

⇑

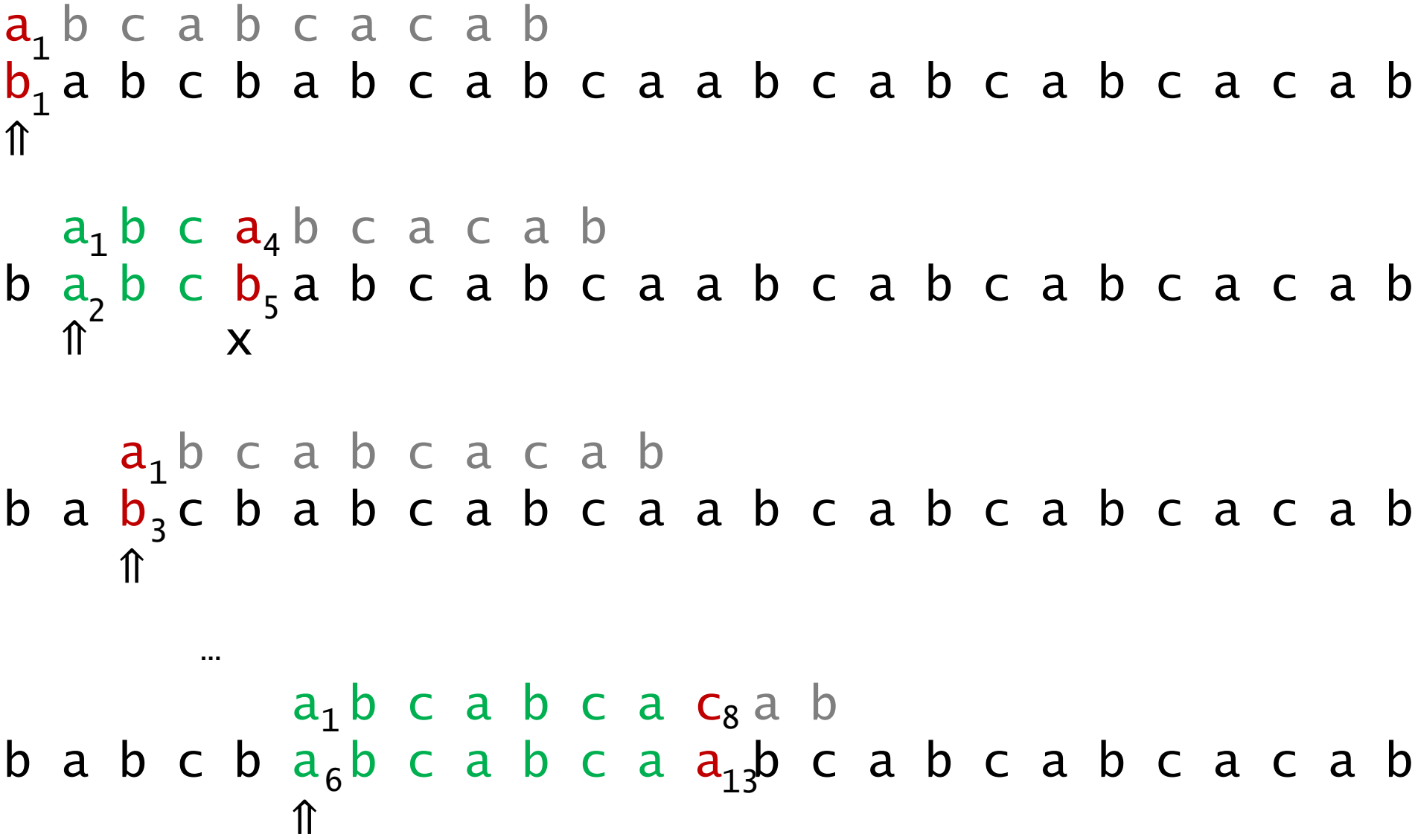
naïeve algoritme



naïeve algoritme



naïeve algoritme



naïeve algoritme

a a a a a a a a b
a a a a a a a a a a a a a a a a a a a a a b
⇑

kwadratische tijd

gebruik informatie!

```
          a1 b  c  a  b  c  a  c8 a  b
b  a  b  c  b a6 b  c  a  b  c  a  a13 b  c  a  b  c  a  b  c  a  c  a  b
          ⇑                       x
```

hoever moeten/kunnen we P doorschuiven?

```
T  =  …  a  b  c  a  b  c  a  x  …  …
P  =     a1 b  c  a  b  c  a  c8 a  b
P′ =        a  b  c  a  b  c  a  c  a  b
               a  b  c  a  b  c  a  c  a  b
                  a1 b  c  a  b5 c  a  c  a  b
```

dat weten we zonder kennis van T !

preprocessing

we gebruiken x niet
(die is ‘onbekend’)

```
     … a b c a b c a x …
P =    a1 b c a b c a c8 a b
P'=          a1 b c a b5 c a c a b
```

fout
op positie

j	=	1	2	3	4	5	6	7	8	9	10
P[j]	=	a	b	c	a	b	c	a	c	a	b
next[j]		0	1	1	0	1	1	0	5	0	1

vervolg
op positie

nul: geen match
schuif helemaal voorbij
huidige positie

positie j: verschuif over j – next[j]

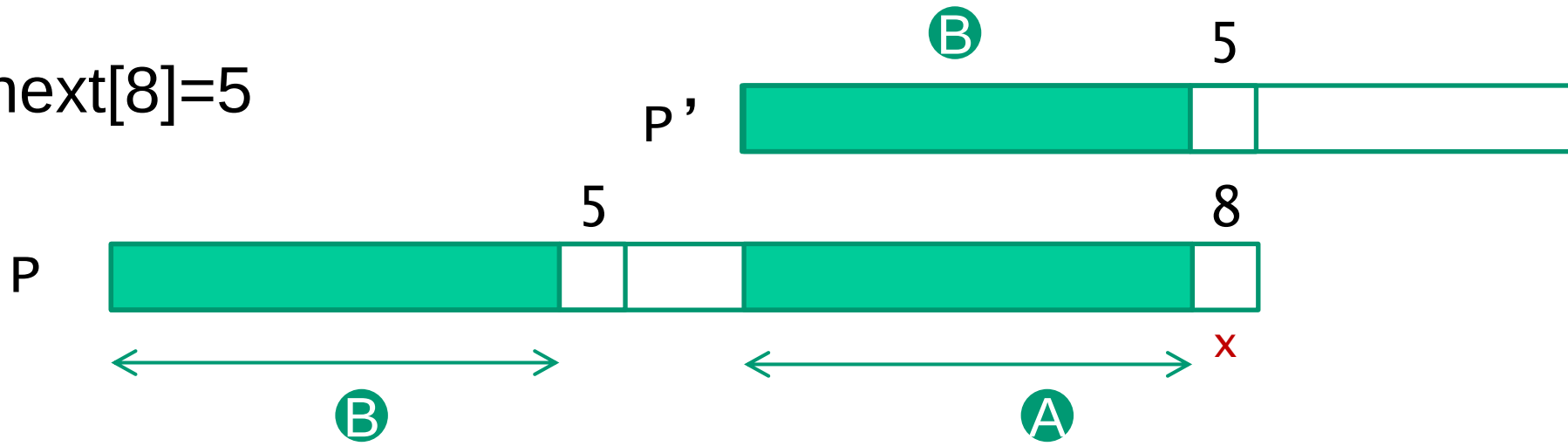
KMP algoritme

```
place pattern at left;
while pattern not fully matched
      and text not exhausted do
      begin
            while pattern character differs from
                  current text character
                  do shift pattern appropriately;
            advance to next character of text;
       end;
```

```
j := k := 1;
while j ≦ m and k ≦ n do
      begin
            while j > 0 and text[k] ≠ pattern[j]
                  do j := next[j];
            k := k + 1; j := j + 1;
      end;
```

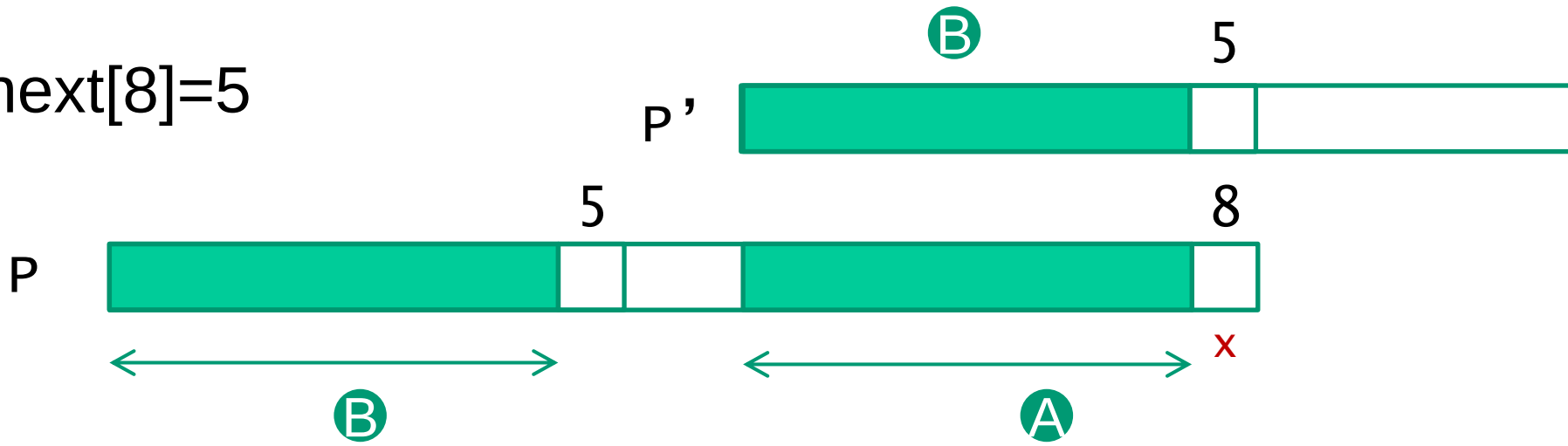
next[k] bepalen? overlap

next[8]=5

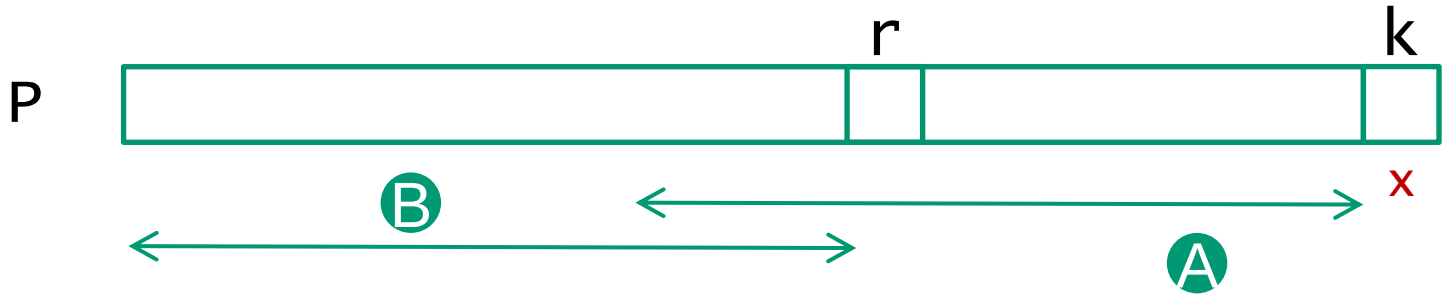


next[k] bepalen? overlap

next[8]=5



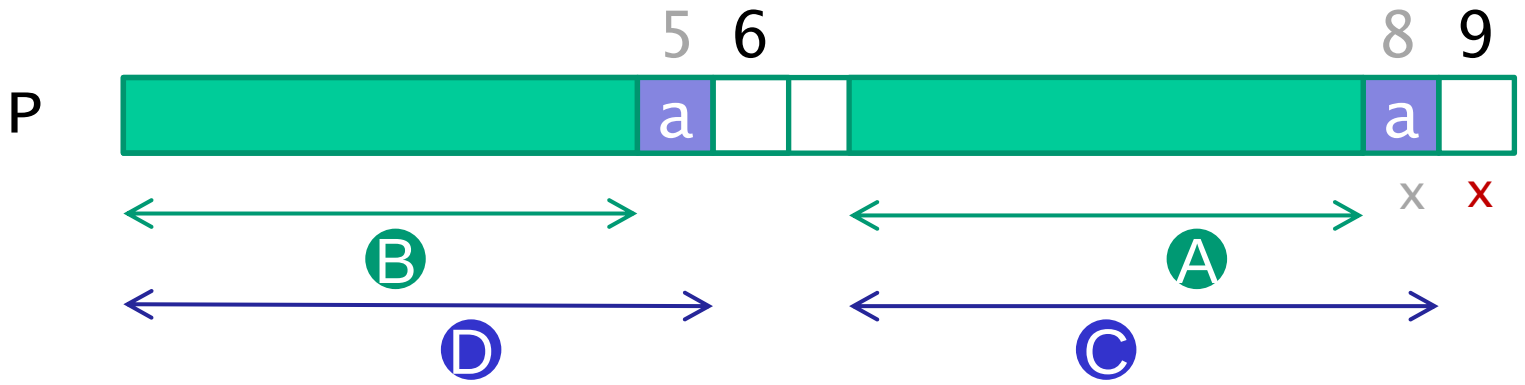
next[k]=r



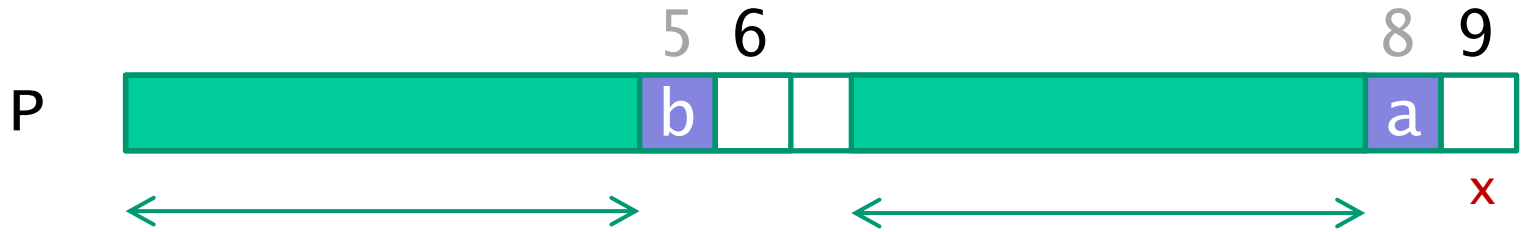
volgende overlap

next[8]=5

next[8+1]=5+1 mits P[8] = P[5]

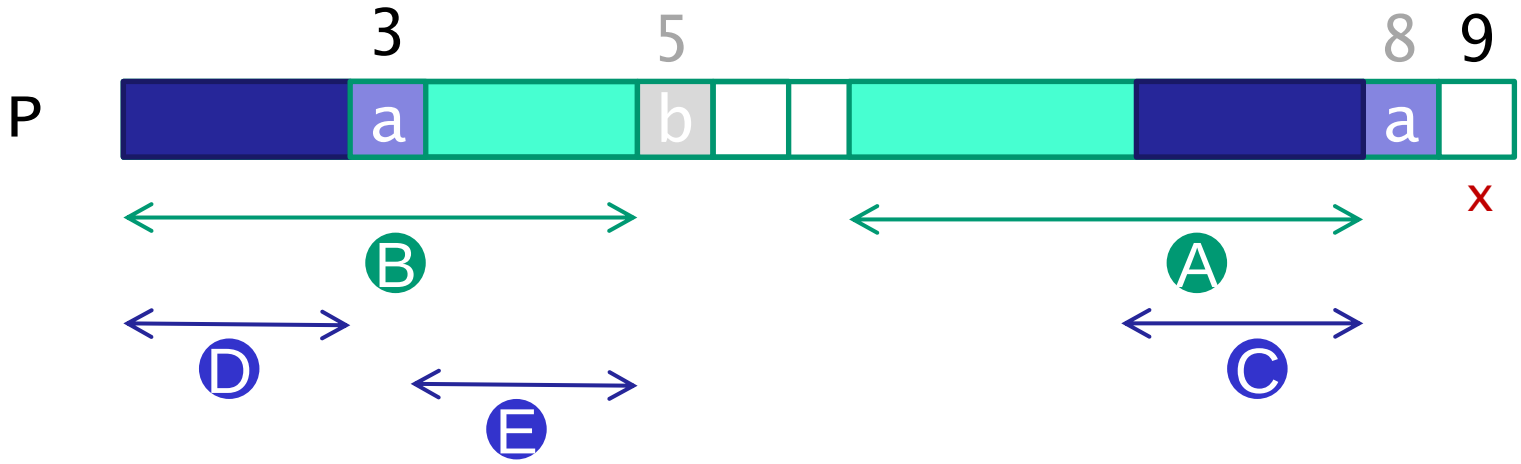


als P[8] ≠ P[5]



overlap

next[8+1]=3+1 mits P[8] = P[3]



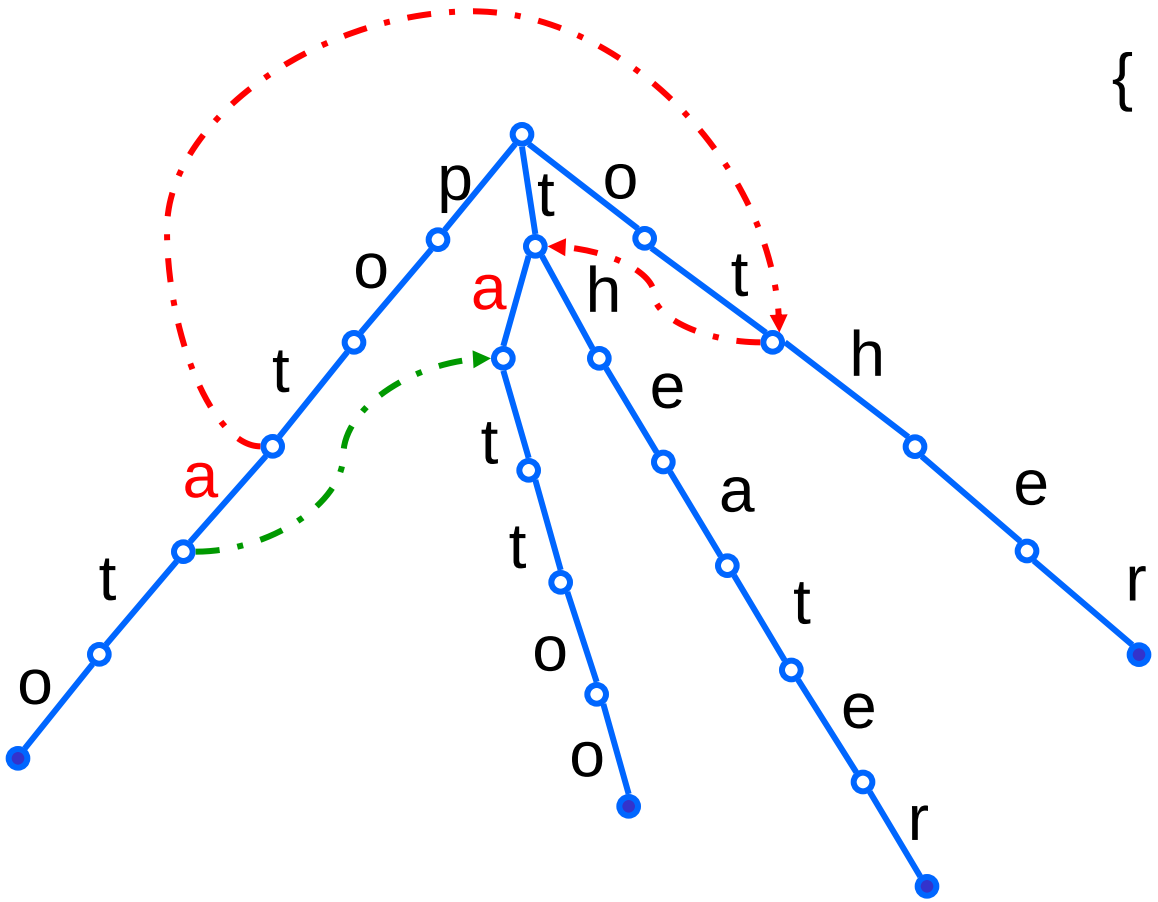
3 = next[5]

patroonherkenning op zichzelf toegepast!

andere methoden

Boyer & Moore (1977)

meerdere woorden: boom



{ potato,
tattoo,
theater,
other }

breadth first
(level-by-level)

potato
other
theater

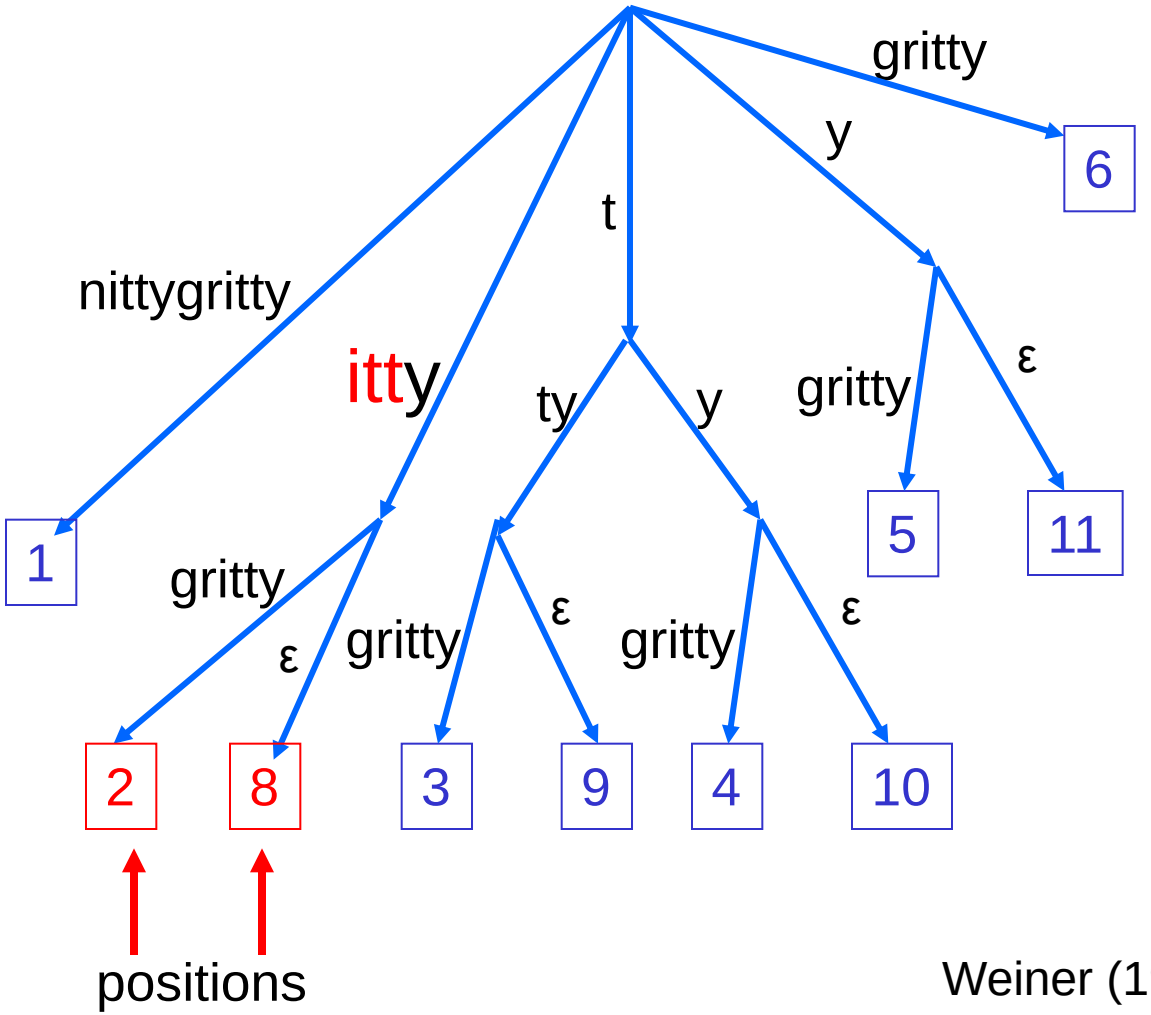
potato
ota ☹
tattoo

Aho & Corasick (1975)

suffix-tree ‘nittygritty’

“Algorithm of the Year 1973”

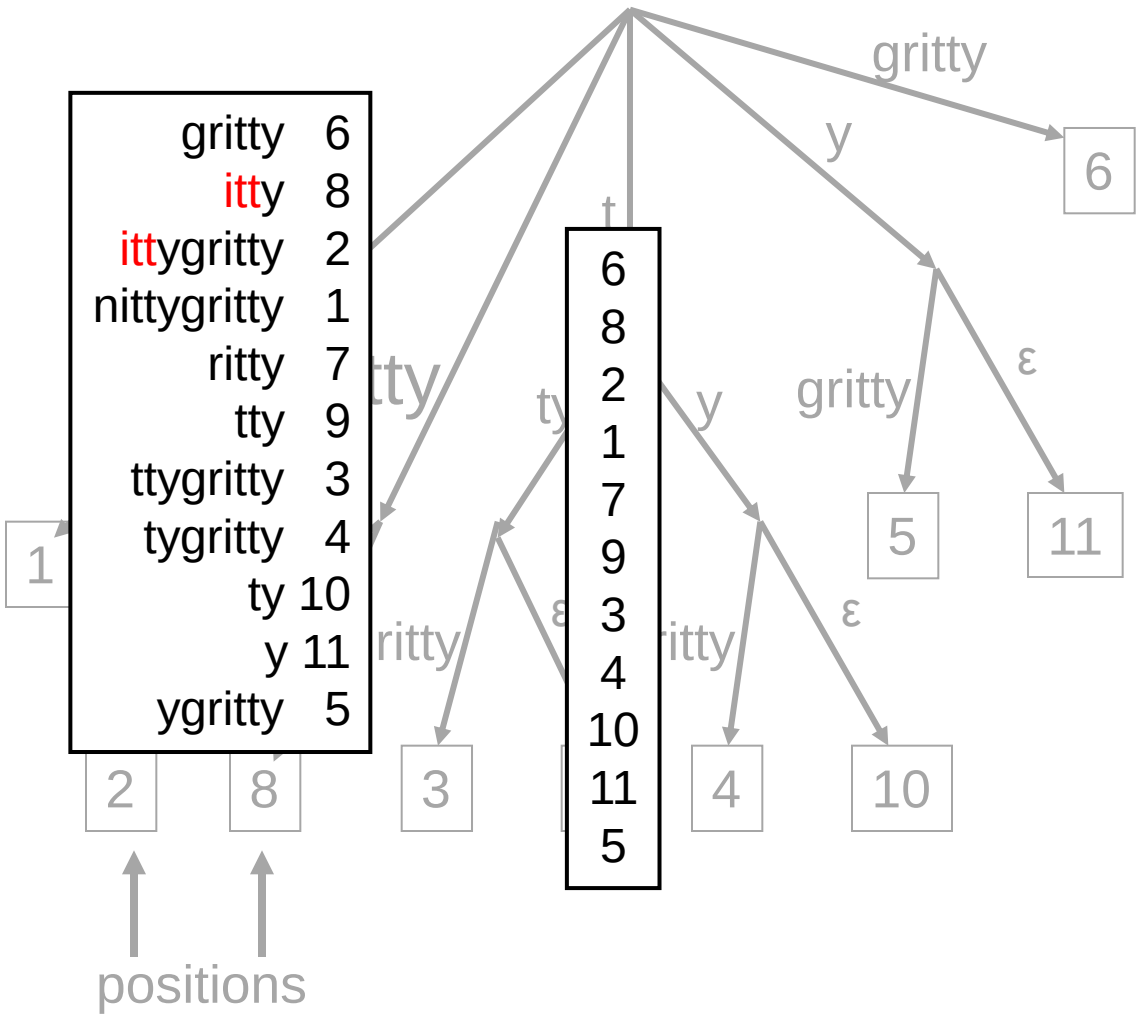
nittygritty	1
ittygritty	2
ttygritty	3
tygritty	4
ygritty	5
gritty	6
ritty	7
itty	8
tty	9
ty	10
y	11



Weiner (1973),
McCreight (1976), Ukkonen (1995)

suffix-array ‘nittygritty’

nittygritty	1
ittygritty	2
ttygritty	3
tygritty	4
ygritty	5
gritty	6
ritty	7
itty	8
tty	9
ty	10
y	11



Manber & Myers (1990)

KMP - Historical remarks

Morris text-editor (1969)

… was too complicated for other implementors of the system to understand, and he discovered several months later that gratuitous "fixes" had turned his routine into a shambles.

Knuth (1970)

- Cook two-way deterministic push-down automata
- Chester strings starting with palindromes

v v w v # u v w

This was the first time in Knuth's experience that automata theory had taught him how to solve a real programming problem better than he could solve it before

dankuwel …

klaar