

Linear-Time Computation of Local Periods

Thierry Lecroq

ABISS

University of Rouen - France

Thierry.Lecroq@univ-rouen.fr

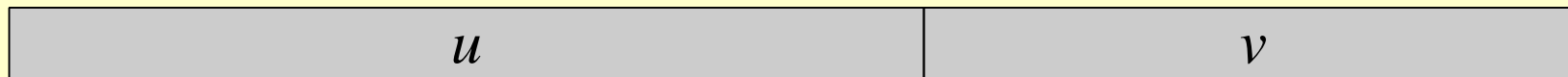
<http://www-igm.univ-mlv.fr/~lecroq>

joint work with Jean-Pierre Duval (Rouen), Roman Kolpakov (Moscow), Gregory Kucherov (Nancy) and Arnaud Lefebvre (Rouen)

Local Period

Definition 1: Let $w = uv$, and $|u| = i$. We say that a non-empty square tt is centered at position i of w (or matches w at central position i) iff the following conditions hold:

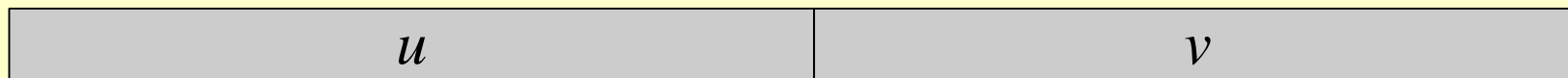
- (i) t is a suffix of u , or u is a suffix of t ,
- (ii) t is a prefix of v , or v is a prefix of t .



internal square



right external square



right and left external square

Local Period

Definition 2: The smallest square centered at a position i of w is called the minimal local square centered at i . The local period at position i of w , denoted $LP_w(i)$, is the period of the minimal square centered at i .

Local Period

Definition 2: The smallest square centered at a position i of w is called the minimal local square centered at i . The local period at position i of w , denoted $LP_w(i)$, is the period of the minimal square centered at i .

Note that $1 \leq LP_w(i) \leq |w|$.

Critical Factorization Theorem

Theorem: For each word w , there exists a position i (and the corresponding factorization $w = uv$ with $|u| = i$) such that $LP_w(i) = per(w)$. Moreover, such a position exists among any $per(w)$ consecutive positions of w .

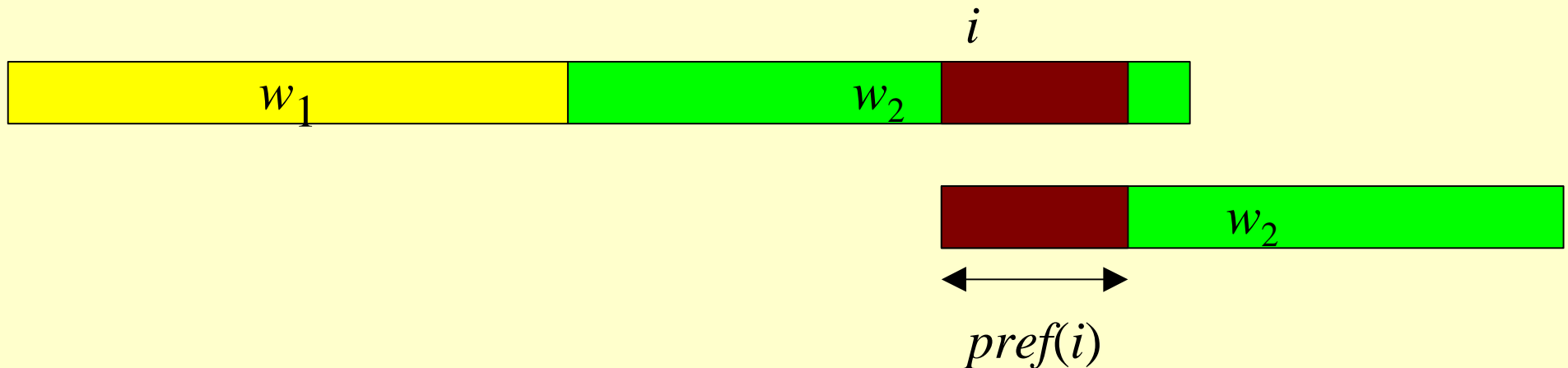
s -factorization

Definition 3: The s -factorization of w without copy overlap is the factorization $w = f_1 f_2 \dots f_m$, where f_i 's are defined inductively as follows:

- (i) $f_1 = w[1]$,
- (ii) assume we have computed $f_1 f_2 \dots f_{i-1}$ ($i \geq 2$), and let $w[b_i]$ be the letter immediately following $f_1 f_2 \dots f_{i-1}$ (i.e. $b_i = |f_1 f_2 \dots f_{i-1}| + 1$). If $w[b_i]$ does not occur in $f_1 f_2 \dots f_{i-1}$, then $f_i = w[b_i]$, otherwise f_i is the longest subword starting at position b_i , which has another occurrence in $f_1 f_2 \dots f_{i-1}$.

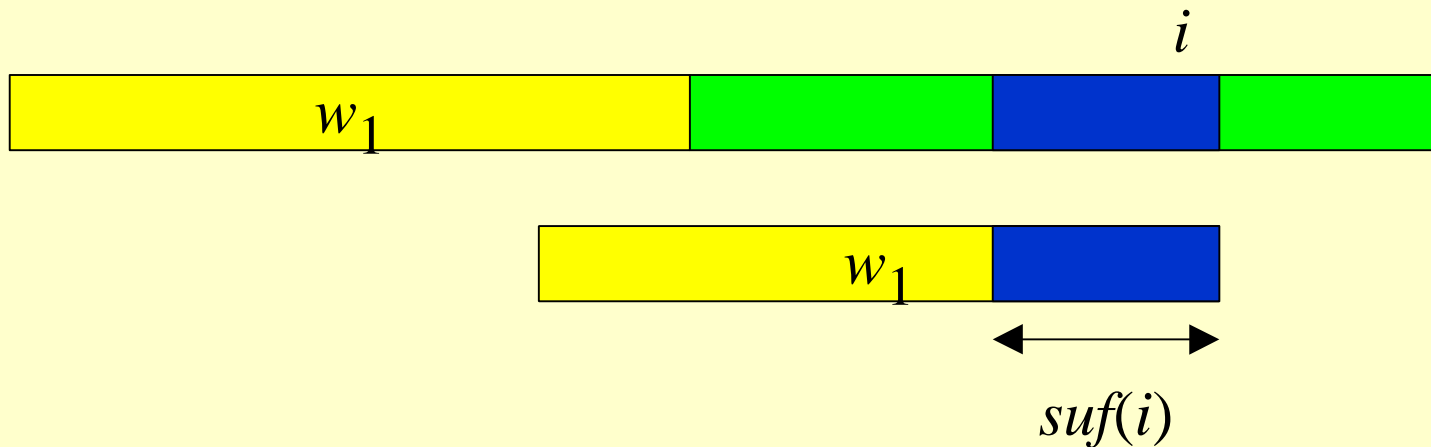
Extension functions

- $w = w_1[1..m]w_2[1..n]$
- $pref(i) = \max \{ j \mid w_2[1..j] = w_2[i..i+j-1] \}$ for $2 \leq i \leq n$
and $pref(n+1) = 0$



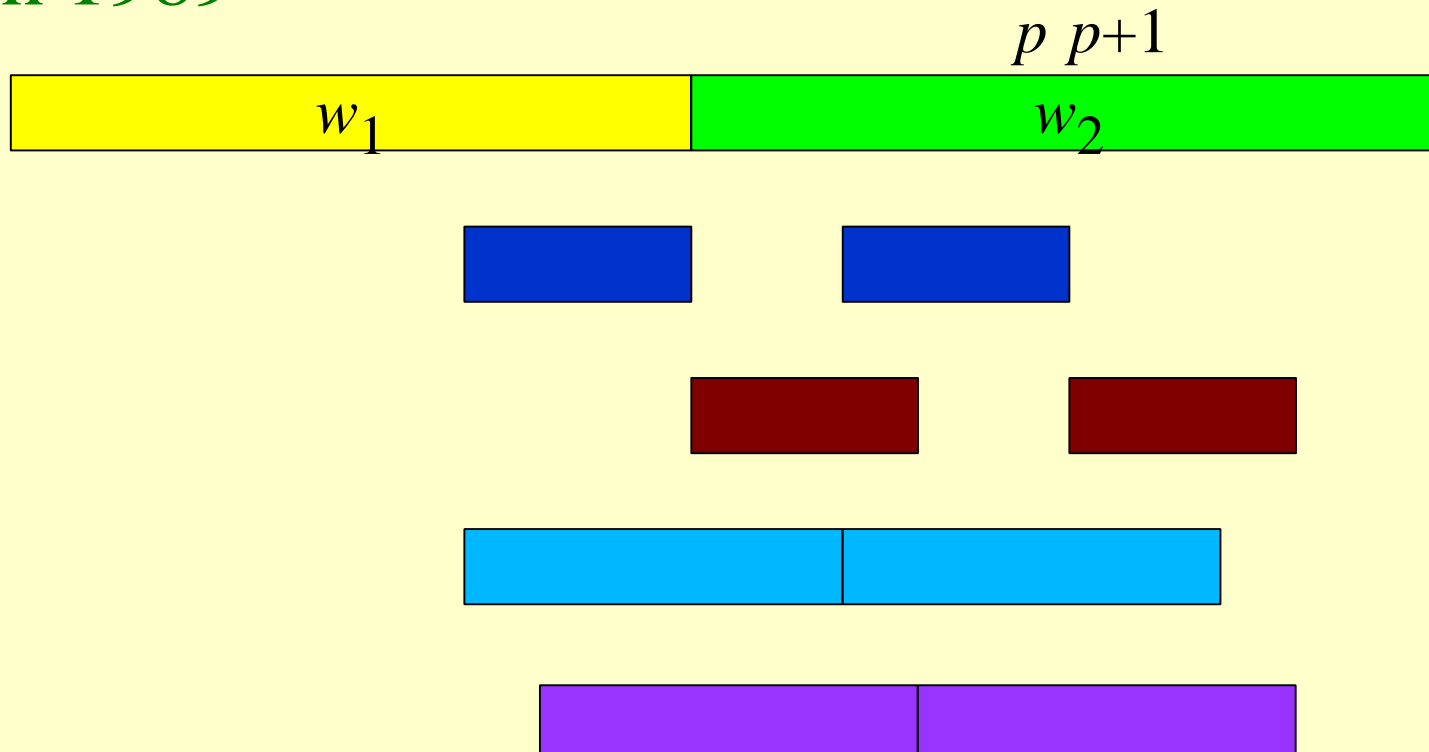
Extension functions

- " $w = w_1[1..m]w_2[1..n]$
- " $suf(i) = \max \{ j \mid w_1[m-j+1..m] = w_2[i-j+1..i] \}$ for $1 \leq i \leq n$



- Then there exists a square with period p iff
- $suf(p) + pref(p+1) \geq p$

Main 1989



If at each position p

$$\text{suf}(p) + \text{pref}(p+1) \geq p$$

is verified

then there is a run of squares centered at each position in the interval

$$[m - \text{suf}(p) + p, m + \text{pref}(p+1)] .$$

This run is a maximal repetition in w [Kolpakov & Kucherov 1999].

General idea for computing all the local periods

Two steps

- computation of the internal minimal squares
- computation of the external minimal squares

General idea for computing all the minimal internal squares

- compute the s -factorization and process factors one-by-one from left to right;
- for each factor f_r we consider separately the squares:
 - which occur completely inside f_r
 - which end in f_r and cross the boundary with f_{r-1}

- The squares of the first type are computed using the fact that f_r has a copy on the left $\Rightarrow O(|f_r|)$.
- The squares of the second type are computed using the extension functions and a lemma stating that squares cannot extend to the left by more than $|f_r| + 2|f_{r-1}|$ letters [Main 1989] $\Rightarrow O(|f_{r-1}| + |f_r|)$.
- Altogether, finding all internal minimal squares in a word of length n can be done in time $O(n)$.

- First we compute the s -factorization of w without copy overlap and we keep for each factor f_r a reference to its non overlapping left copy.
- The algorithm process all factors f_r from left to right and computes for each factor f_r all minimal squares **ending** in f_r .
- For each internal minimal square found centered at position i , $LP_w(i)$ is set.

After the whole word has been processed, positions i whose values $LP_w(i)$ have not been assigned are those for which no internal square centered at i exists and $LP_w(i)$ is computed with another technique.

- Let $f_r = w[m + 1 .. m + k]$ be the current factor and $w[j + 1 .. j + k]$ be its left factor ($j + k \leq m$)
- If for some position $m + i$ ($1 \leq i < k$) the minimal square centered at $m + i$ occurs entirely inside the factor f_r (i.e. $LP_w(m + i) \leq \min \{ i, k - i \}$)
- then

$$LP_w(m + i) = LP_w(j + i)$$

- $LP_w(j + i)$ has already been computed thus we can compute all values $LP_w(m + i) \leq \min \{ i, k - i \}$ in time $O(|f_r|)$.

$j+1$

$j+k$

$m+1$

$m+k$

f_r



$j+1$

$j+k$

$m+1$

$m+i$

$m+k$



$j+1$

$j+i$

$j+k$

$m+1$

$m+i$

$m+k$



It remains to find the values $LP_w(m + i)$ which corresponds to minimal squares that end in f_r and extend to the left beyond the border between f_r and f_{r-1} .

- It remains to find the values $LP_w(m + i)$ which corresponds to minimal squares that end in f_r and extend to the left beyond the border between f_r and f_{r-1} .
- We divide those squares into 2 categories:
 - those centered in f_r
 - those centered to the left of f_r

- We concentrate on squares centered at positions in $[m, m + k - 1]$ starting at positions $\leq m$ and ending inside f_r .
- We compute all such squares in increasing order of periods using extension functions.
- For each $p \in [1, k - 1]$ we compute the run of all squares of period p centered at positions in $[m, m + k - 1]$ starting at positions $\leq m$ and ending inside f_r .

Assume that :

- we have computed a run of such squares of period p
- $q < p$ is the maximal period for which squares have been previously found

$$p \geq 2q$$

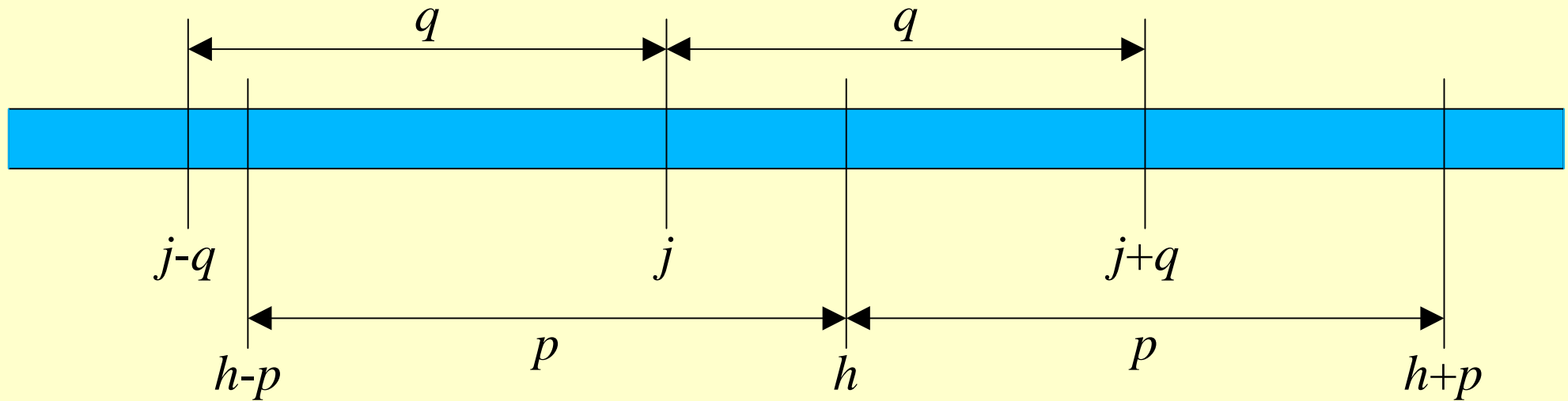
- If $p \geq 2q$ then we check each square of the run whether it is minimal or not by checking the value $LP_w(i)$.
- If this square is not minimal, then its center i has already been assigned a value $LP_w(i)$.
- If no value has previously been assigned then we have found the minimal square centered at i .

- There are at most p such squares of period p (their centers belong to $[m, m + p - 1]$)
- Checking all of them takes at most $2(p-q)$ individual checks (since $q \leq p/2$ and $p-q \geq p/2$)

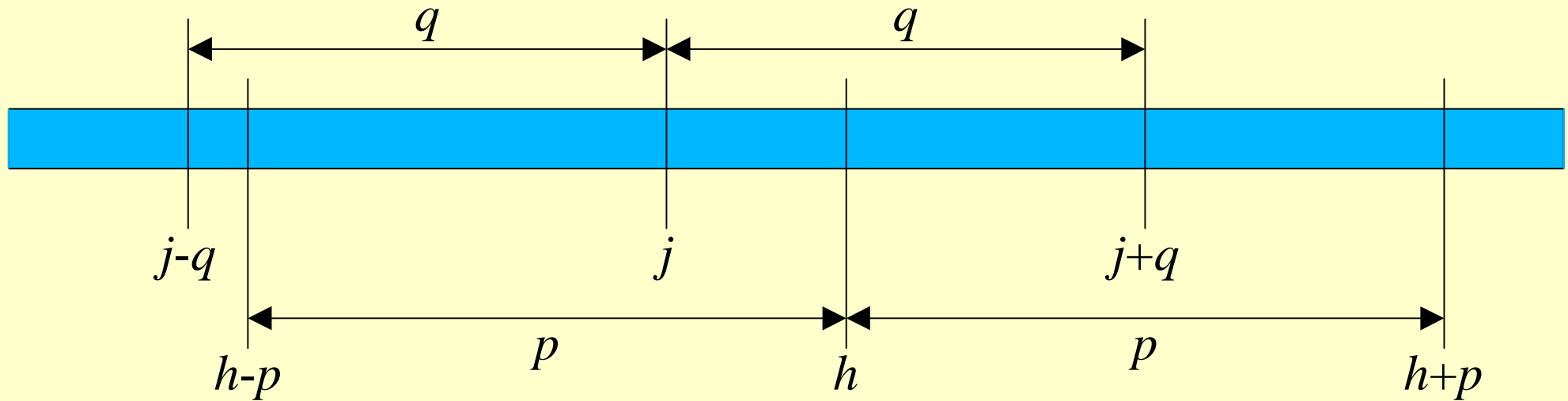
$$p < 2q$$

- consider a square $s_q = w[j - q + 1 .. j + q]$ of period q and center j
 - We claim that we need to check for minimality only the squares s_p of period p which have their center h verifying one of the following inequalities:
 - $|h - j| \leq p - q$ or
 - $h \geq j + q$
- h is located either within distance $p - q$ from j or beyond the end of square s_q .

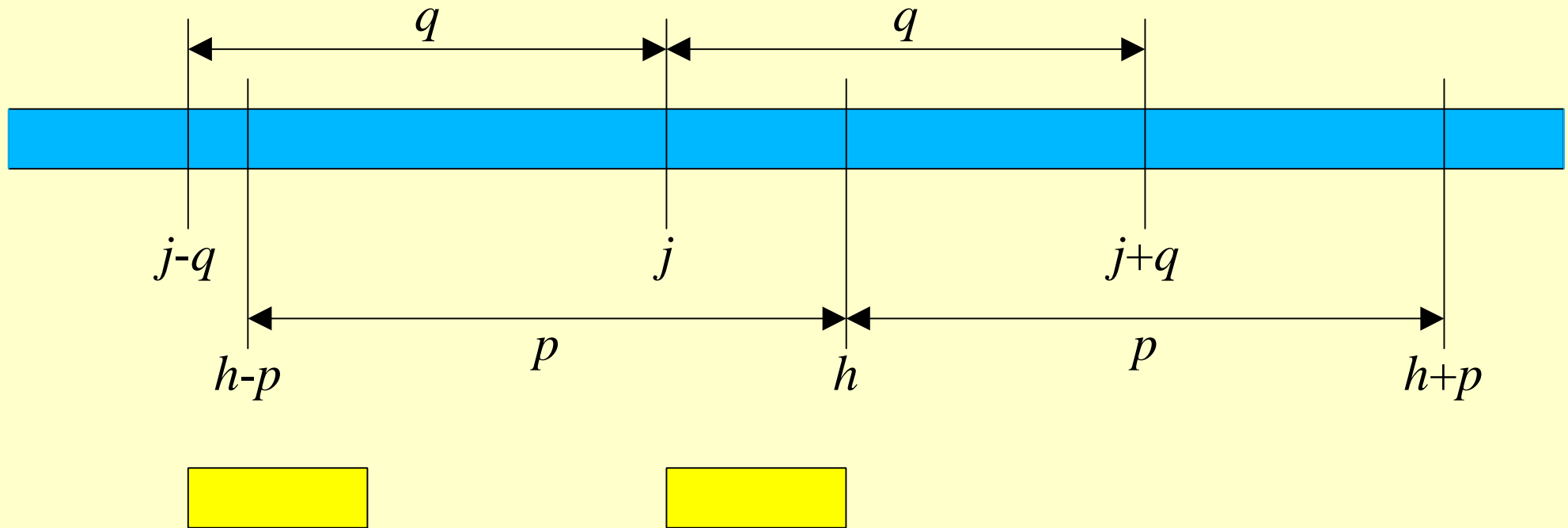
- Proof by contradiction: assume that $|h - j| > p - q$ and $h < j + q$
- 2 symmetrical cases: $h > j$ or $h < j$
- $h > j$



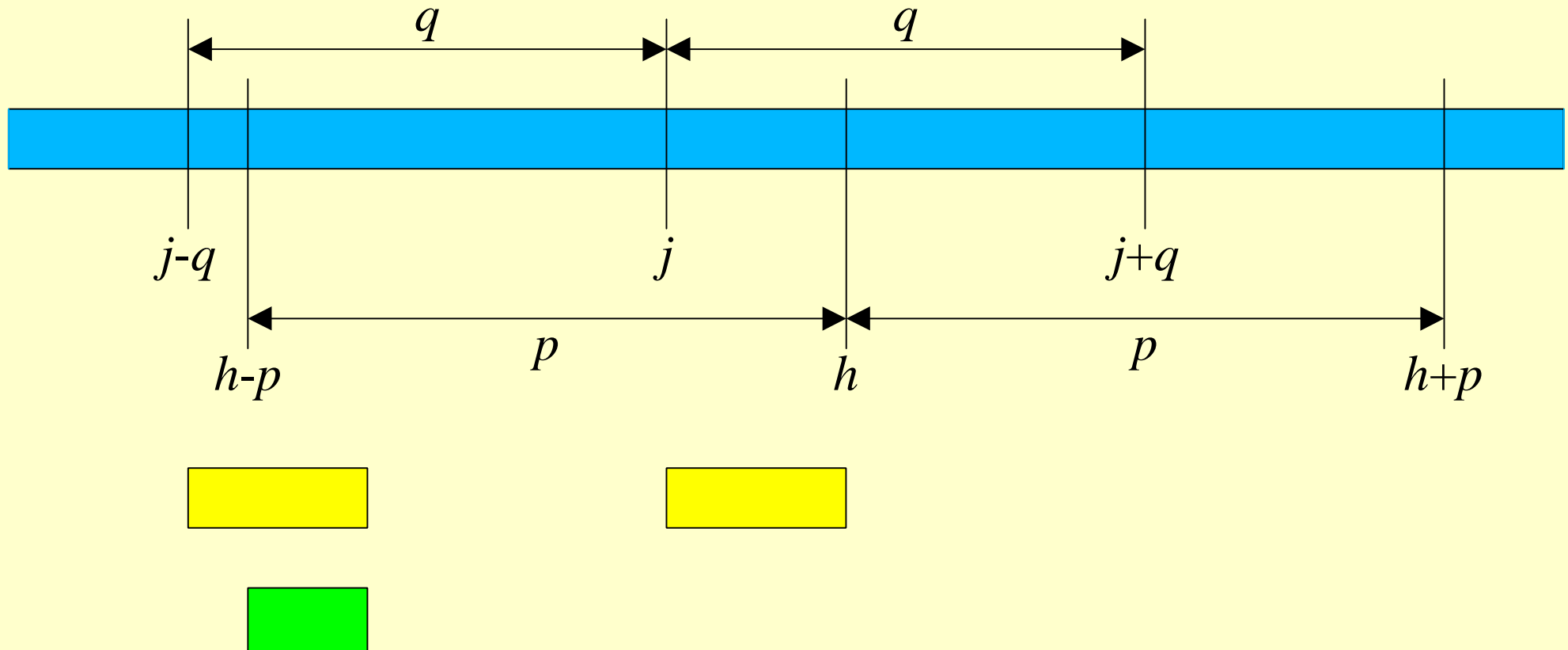
- Proof by contradiction: assume that $|h - j| > p - q$ and $h < j + q$
- 2 symmetrical cases: $h > j$ or $h < j$
- $h > j$



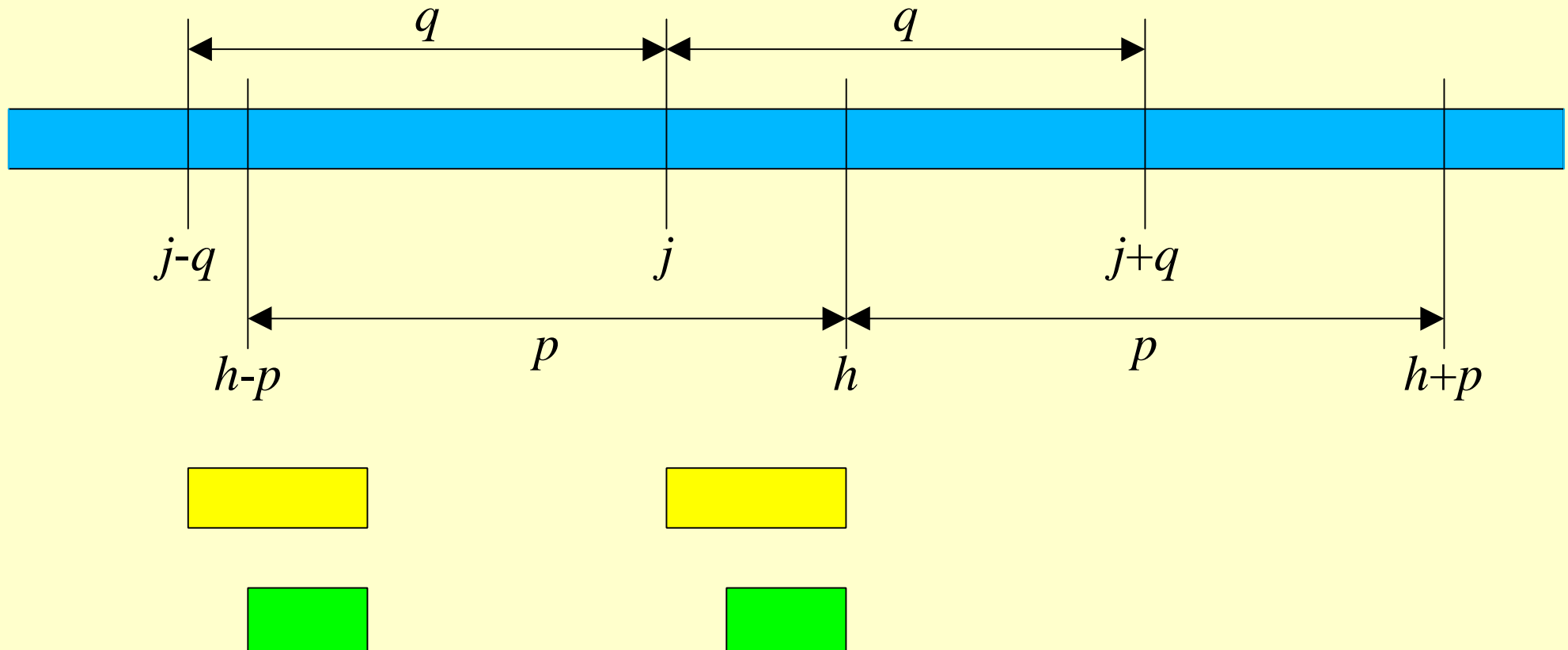
- Proof by contradiction: assume that $|h - j| > p - q$ and $h < j + q$
- 2 symmetrical cases: $h > j$ or $h < j$
- $h > j$



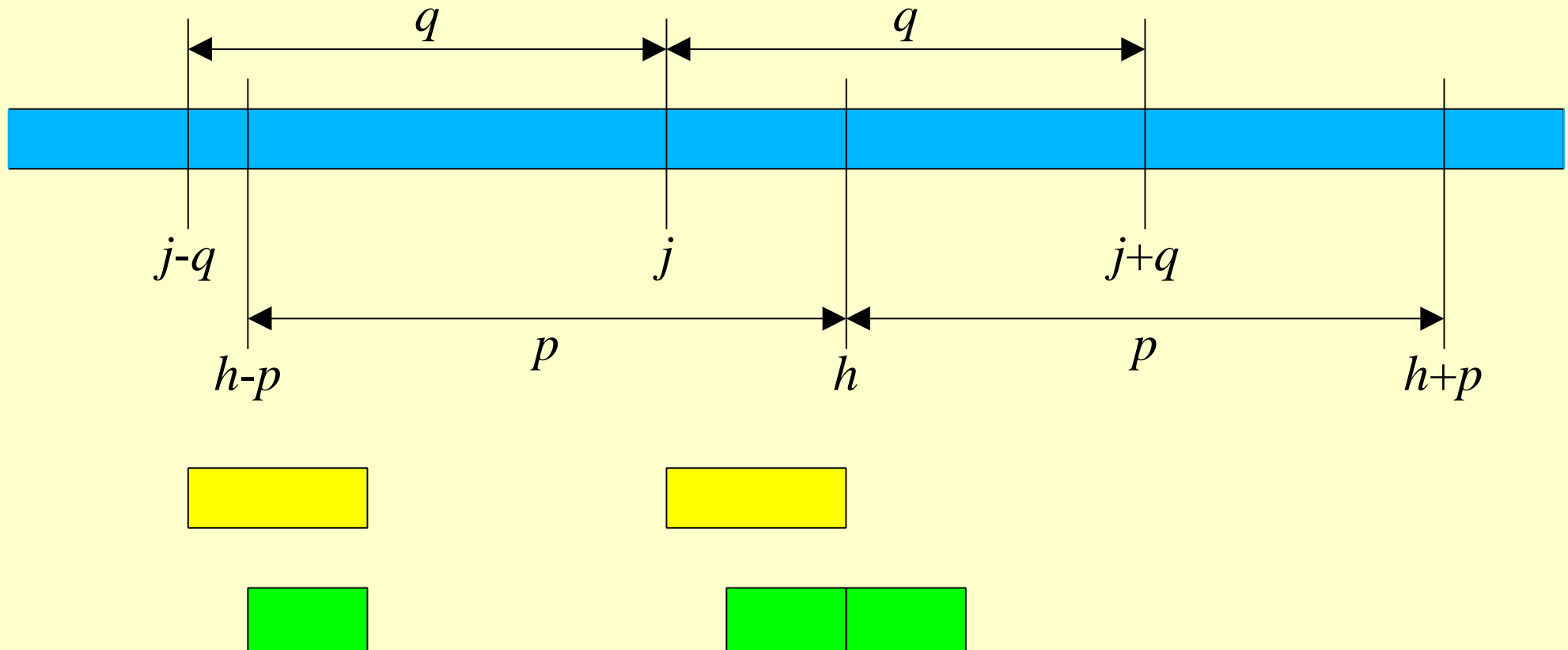
- Proof by contradiction: assume that $|h - j| > p - q$ and $h < j + q$
- 2 symmetrical cases: $h > j$ or $h < j$
- $h > j$



- Proof by contradiction: assume that $|h - j| > p - q$ and $h < j + q$
- 2 symmetrical cases: $h > j$ or $h < j$
- $h > j$



- Proof by contradiction: assume that $|h - j| > p - q$ and $h < j + q$
- 2 symmetrical cases: $h > j$ or $h < j$
- $h > j$



- there are at most $2(p-q)$ squares s_p verifying $|h - j| \leq p - q$
- there are at most $p - q$ squares s_p verifying $h \geq j + q$ because s_p must start before m ($h \leq m + p$)
- there are at most $3(p - q)$ squares of period p to check for minimality
- there are at most $O(|f_r|)$ squares to check for the current factor

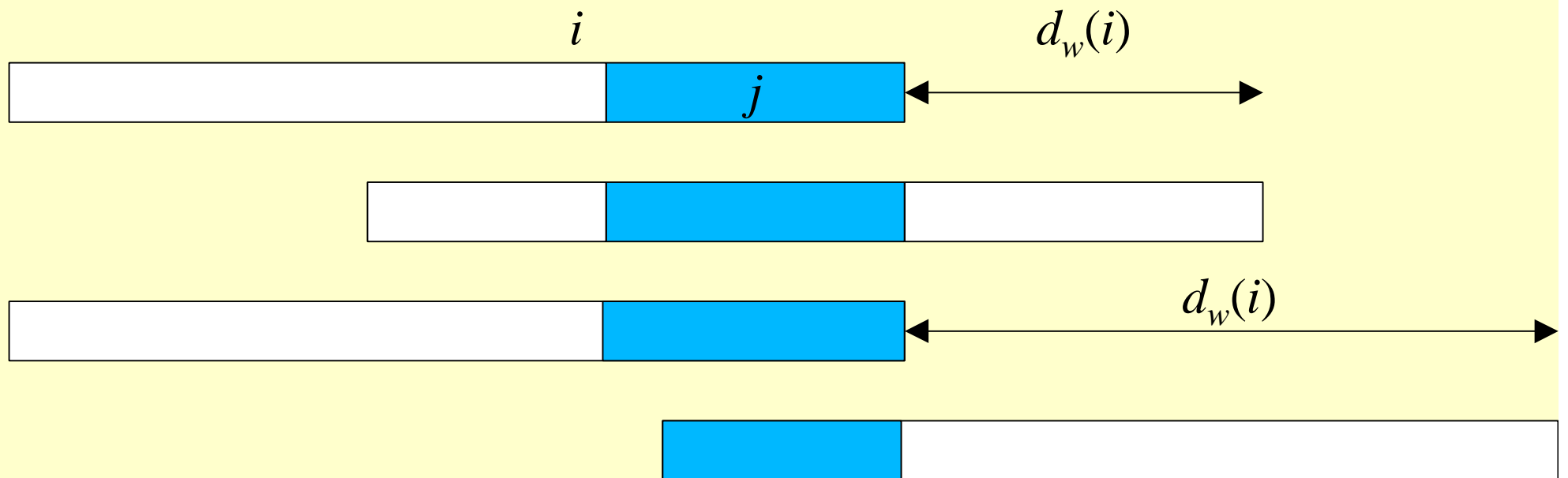
- A similar argument applies to the square centered on the left of f_r
- $O(|f_{r-1}| + |f_r|)$ checks for squares crossing the border between f_{r-1} and f_r
- $O(|f_r|)$ for squares inside f_r

Theorem 2: In a word of length n , all internal minimal squares can be computed in time $O(n)$.

Simplified Boyer-Moore shift function

Definition 4: For a word w of length n the simplified Boyer-Moore shift function is defined as follows:

$$d_w(i) = \min \{ k \mid k \geq 1 \text{ and } \forall j, i < j \leq n, k \geq j \text{ or } w[j] = w[j-k] \}$$

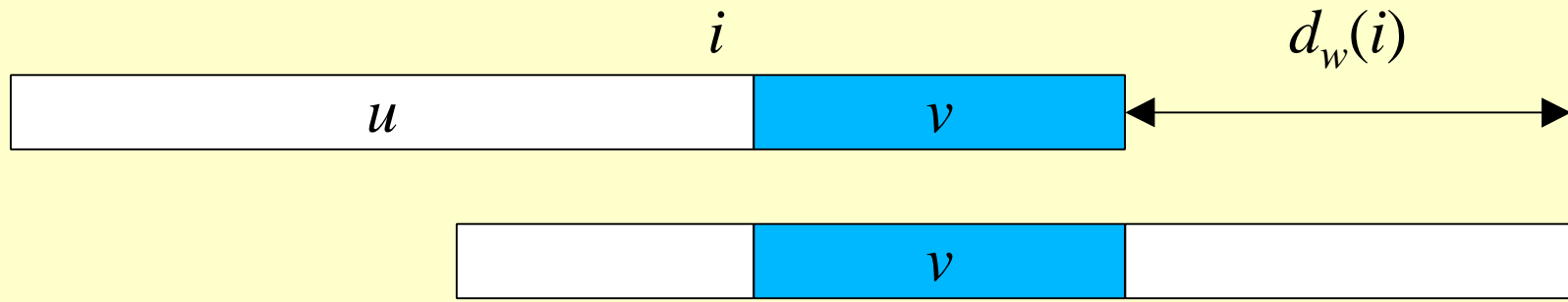


Lemma 1: Let $w = uv$ with $|u| \geq |v|$. If there is no internal square centered at $i = |u|$, then the minimal right external square has period $d_w(i)$.

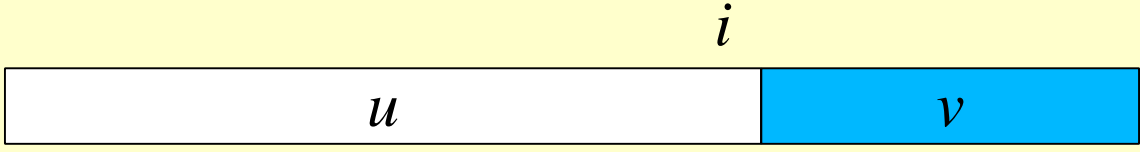
Proof:

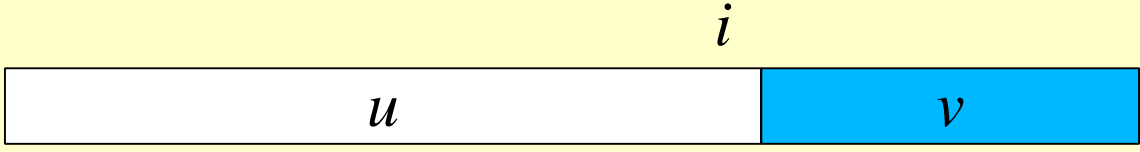
2 cases

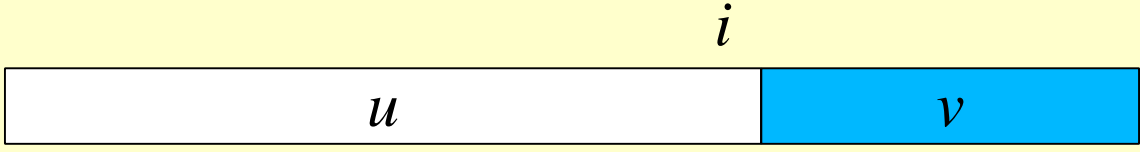
- $d_w(i) \leq |u|$
- $d_w(i) > |u|$

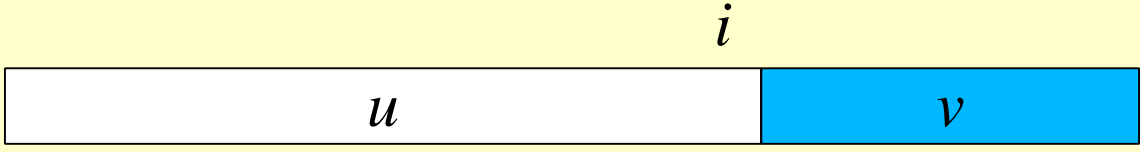


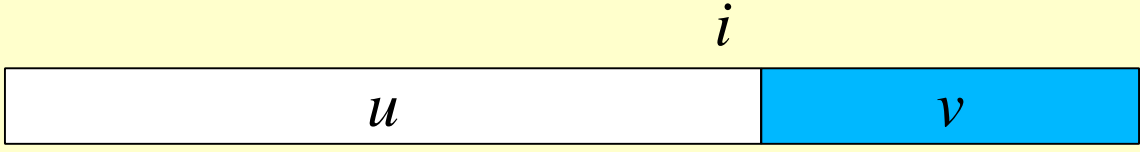
$$d_w(i) \leq |u|$$

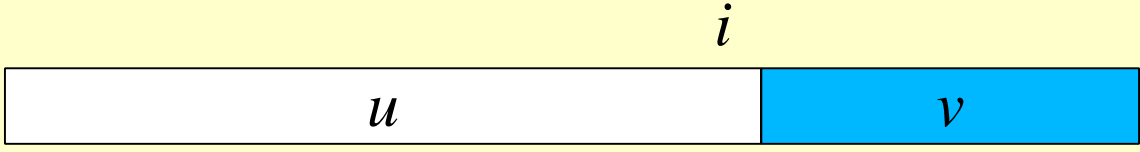


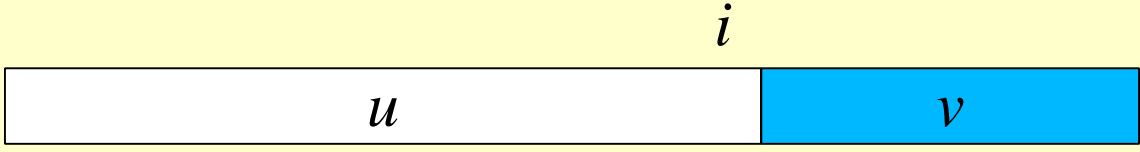


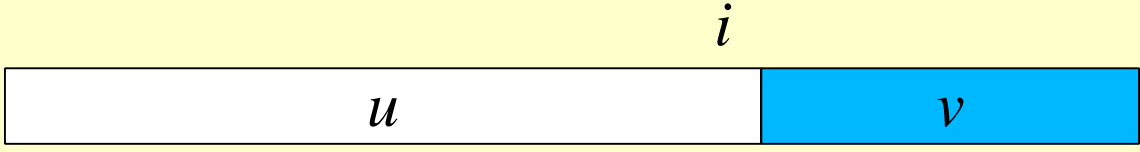


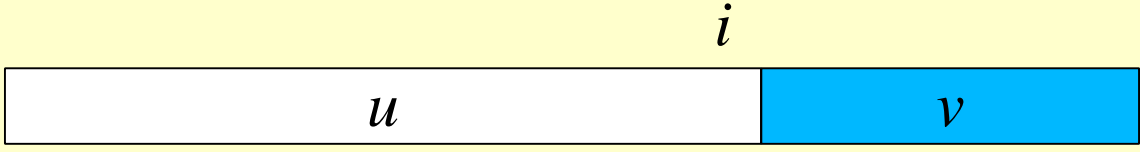


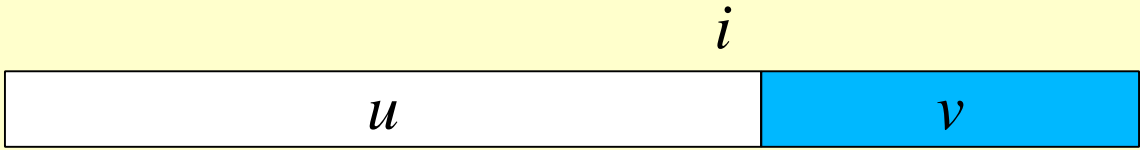


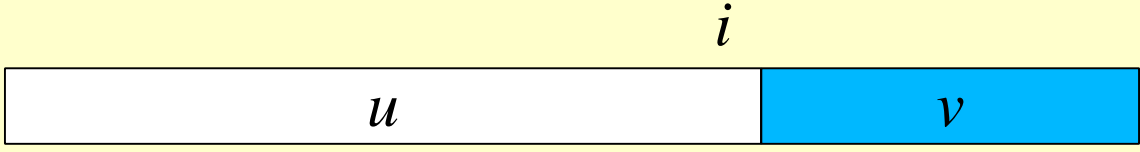


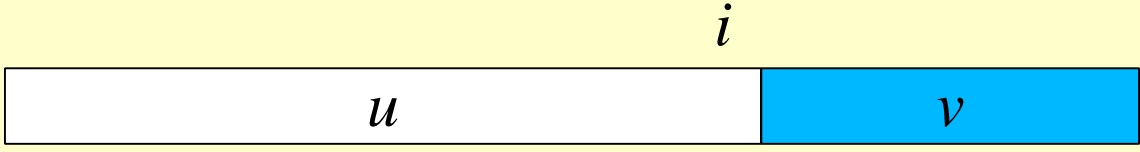


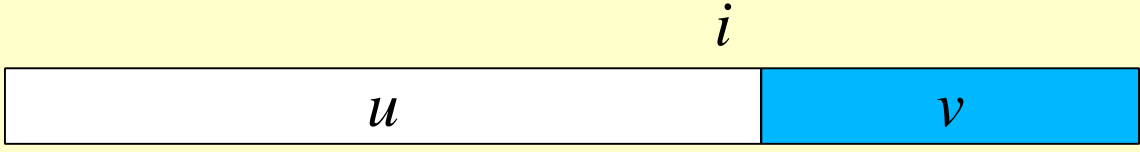


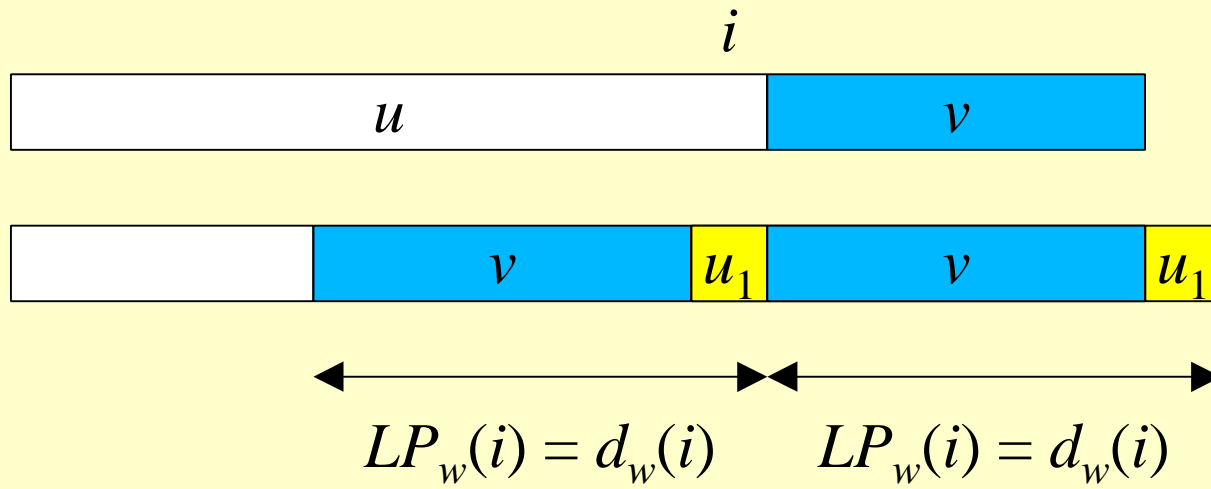


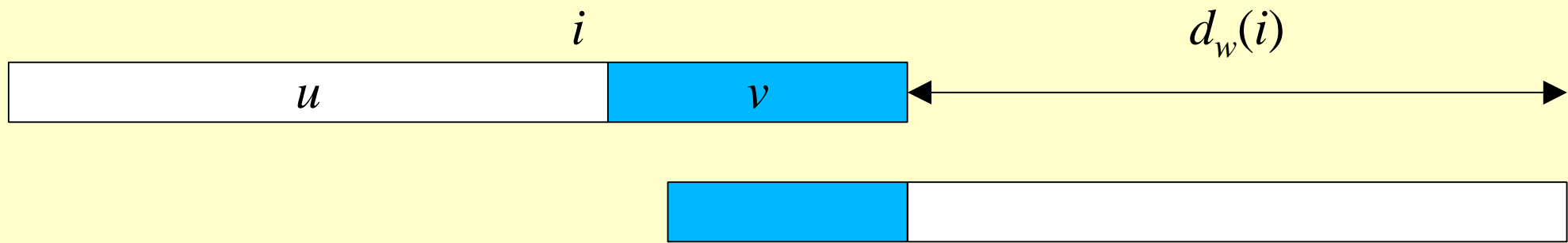




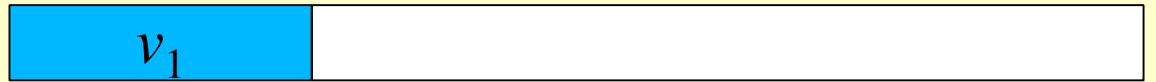
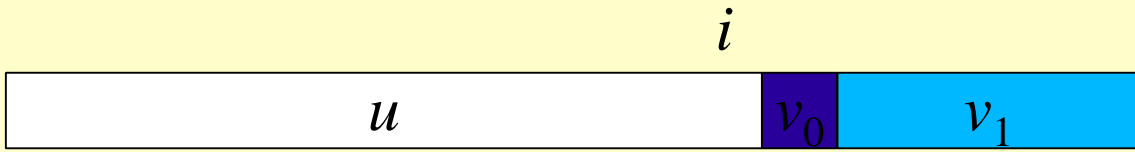


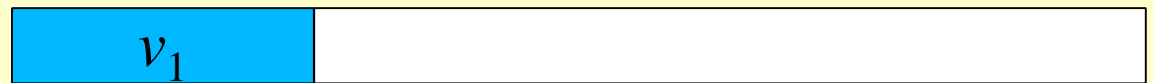


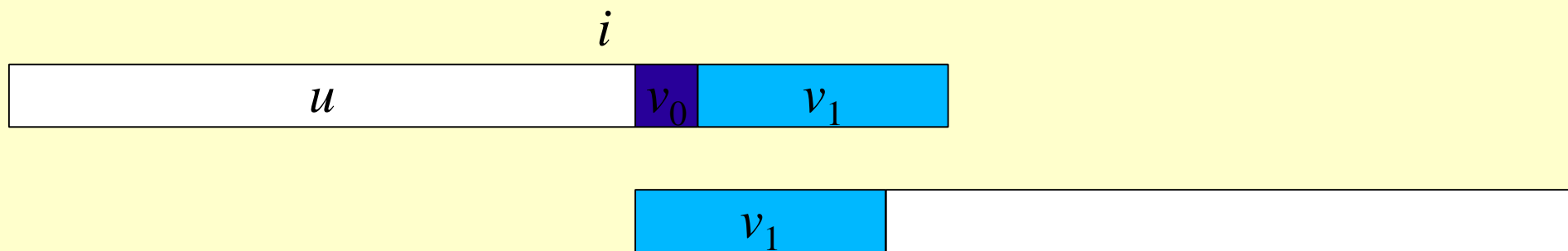


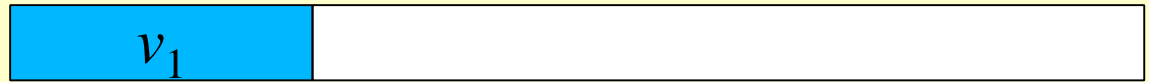
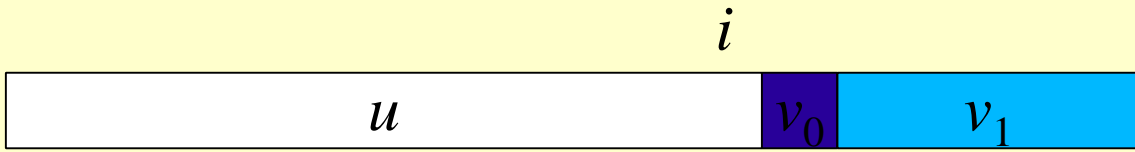


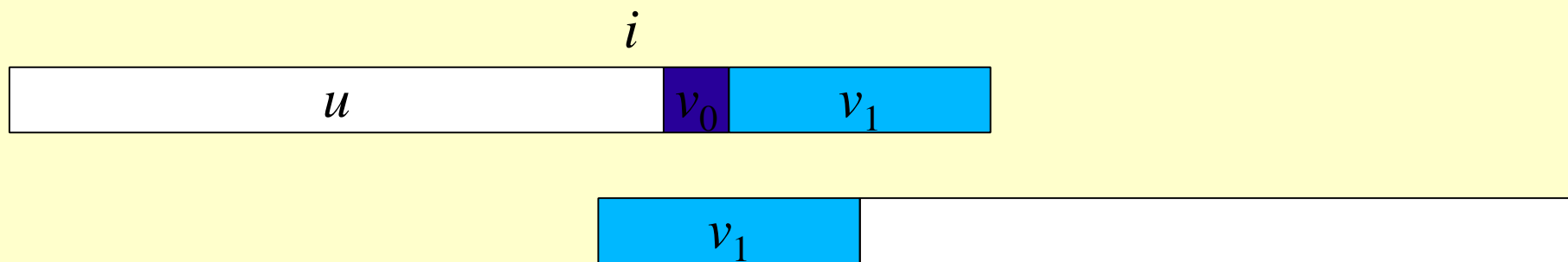
$$d_w(i) > |u|$$

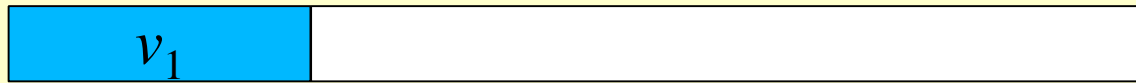
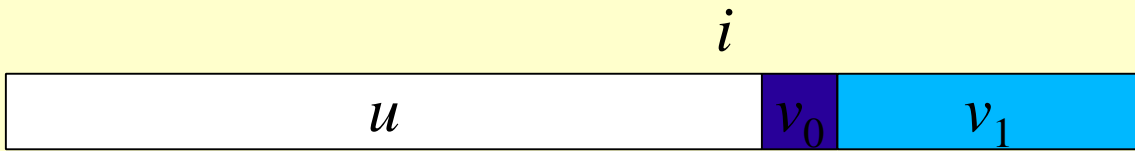


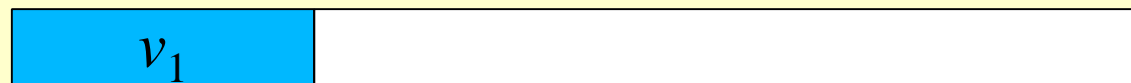
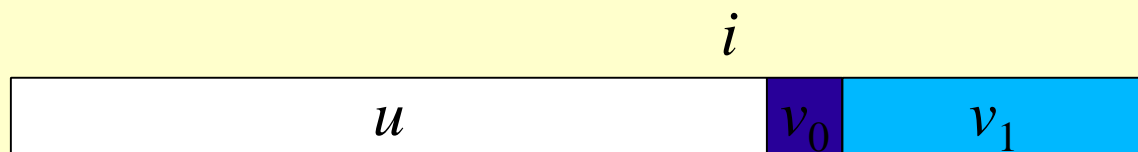


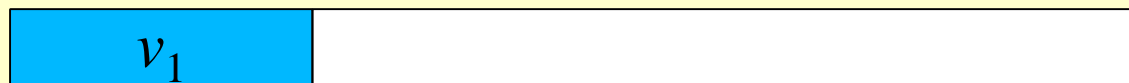
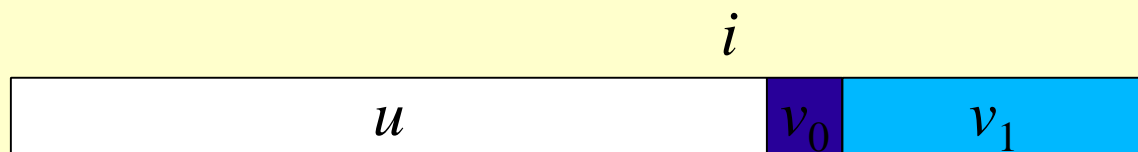


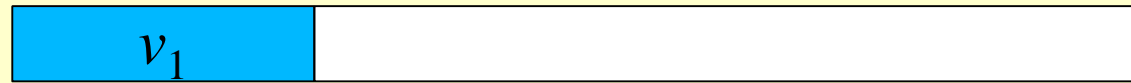
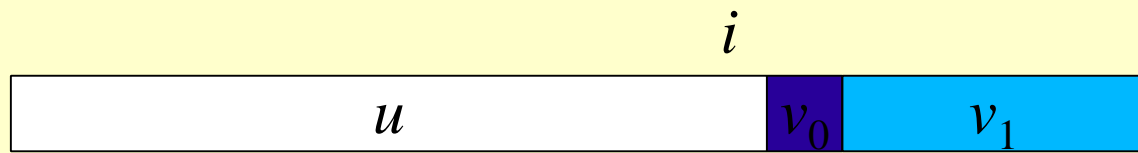


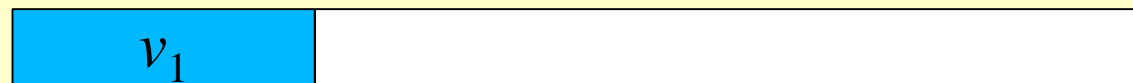
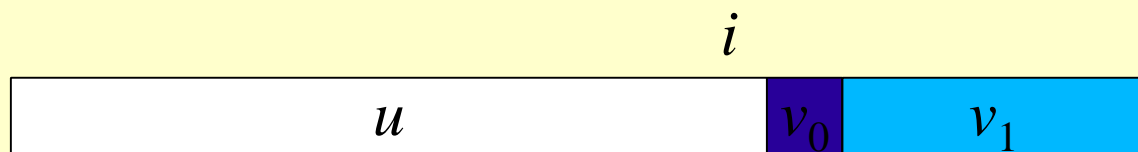


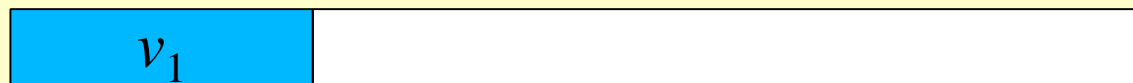
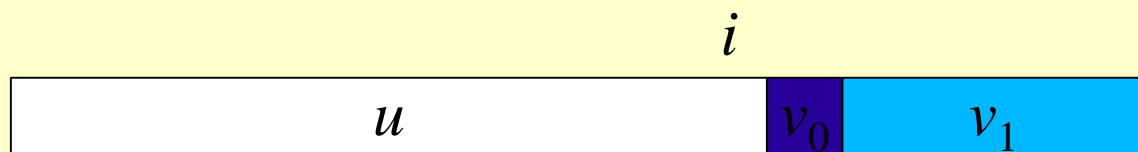


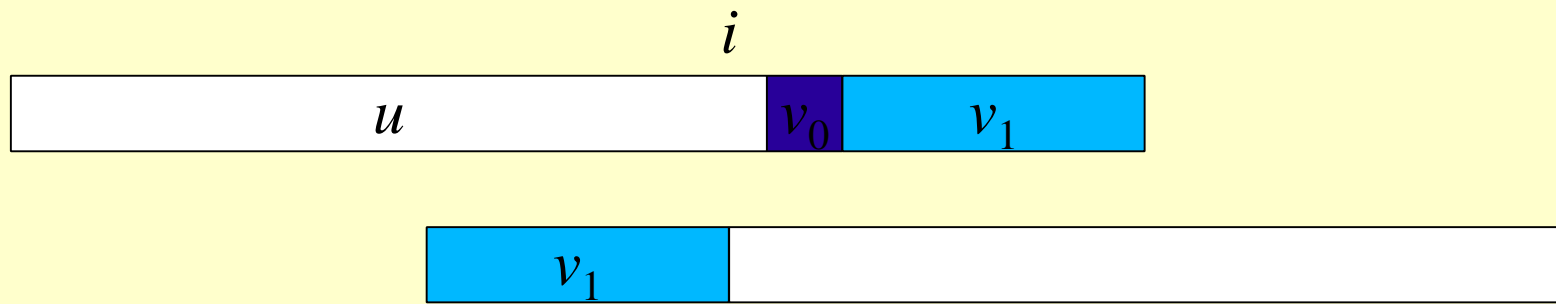


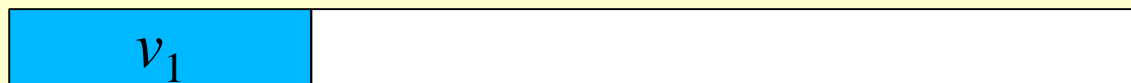
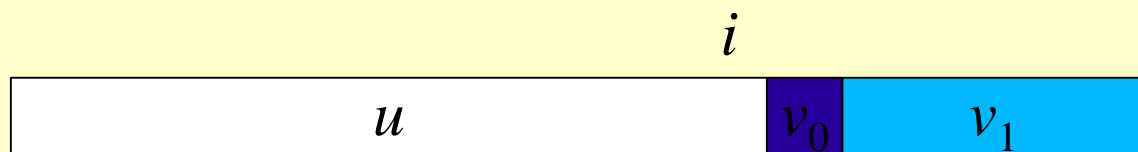


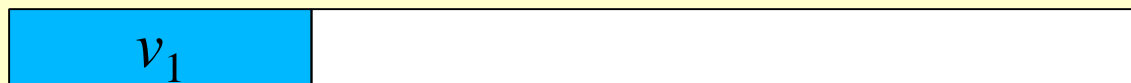
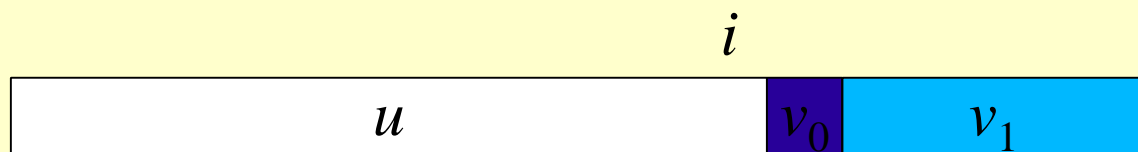


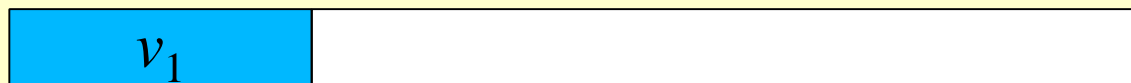
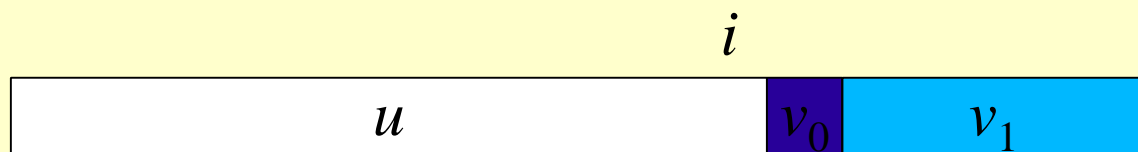


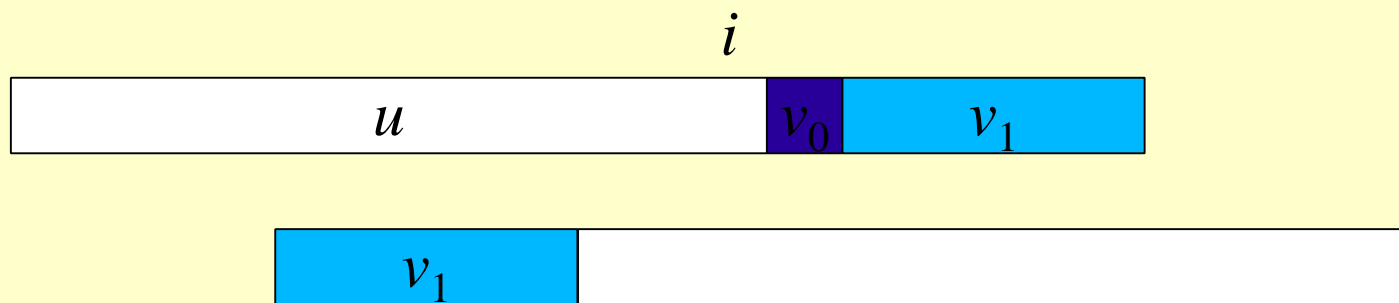


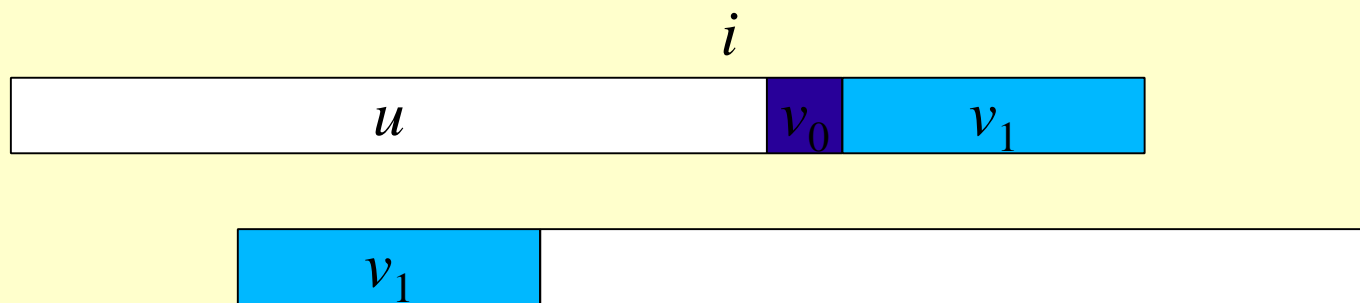


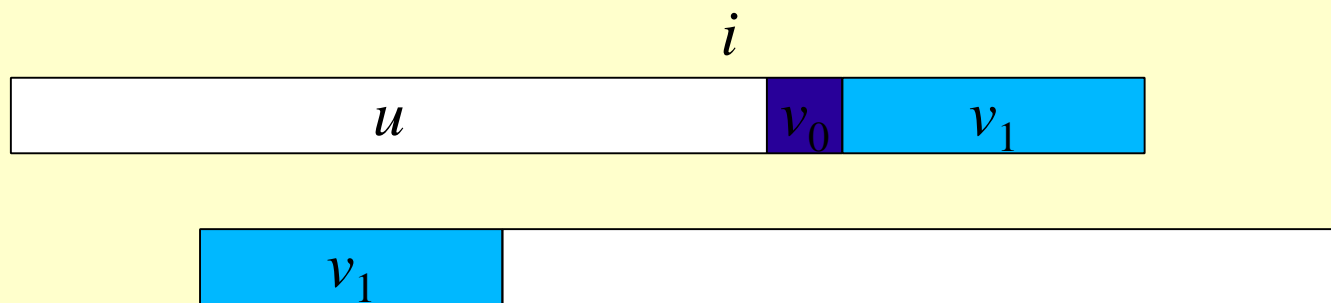


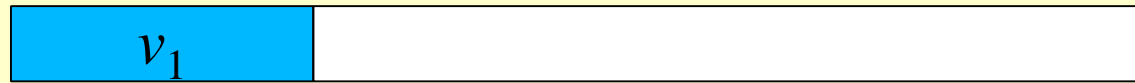
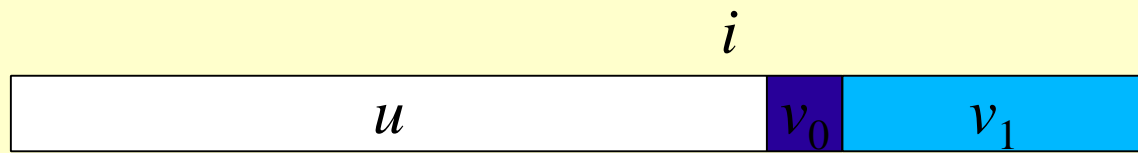


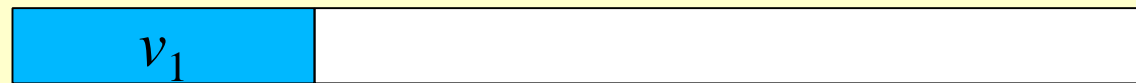
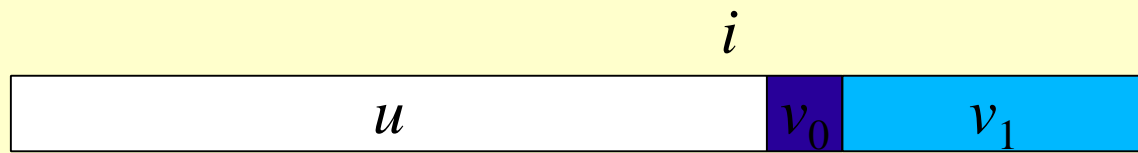


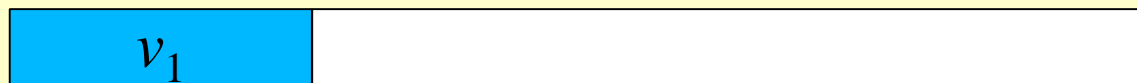
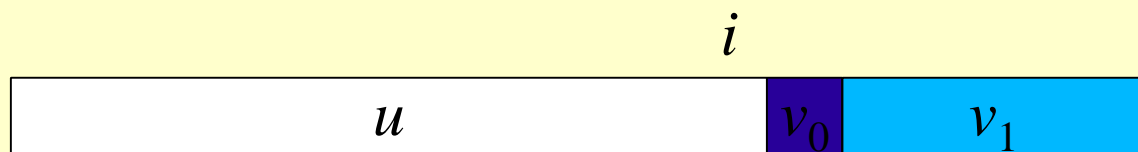


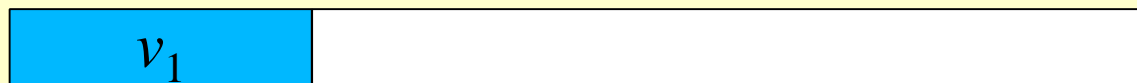
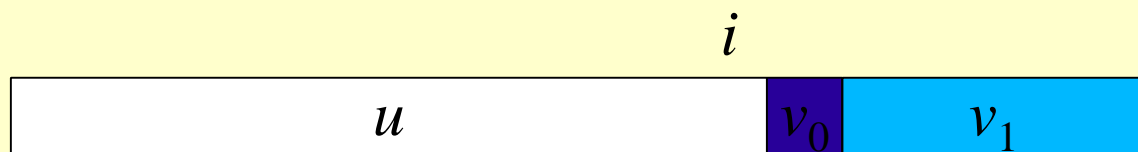


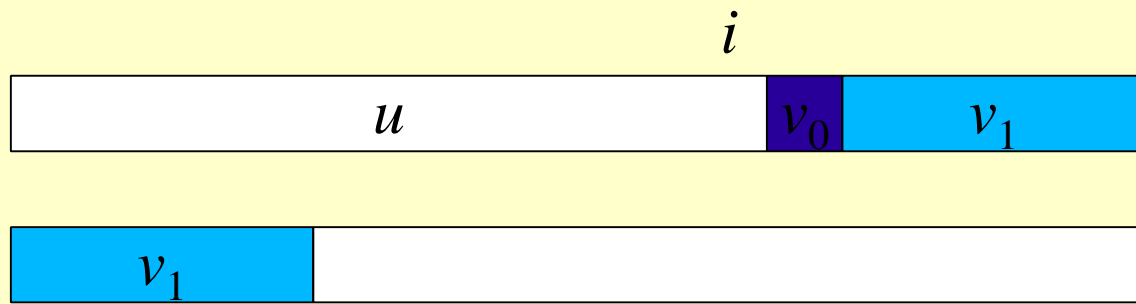




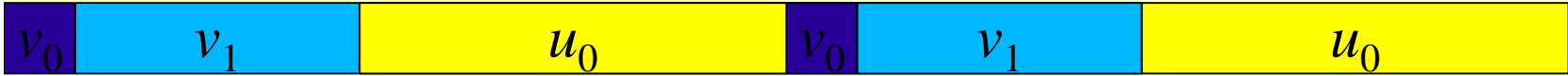


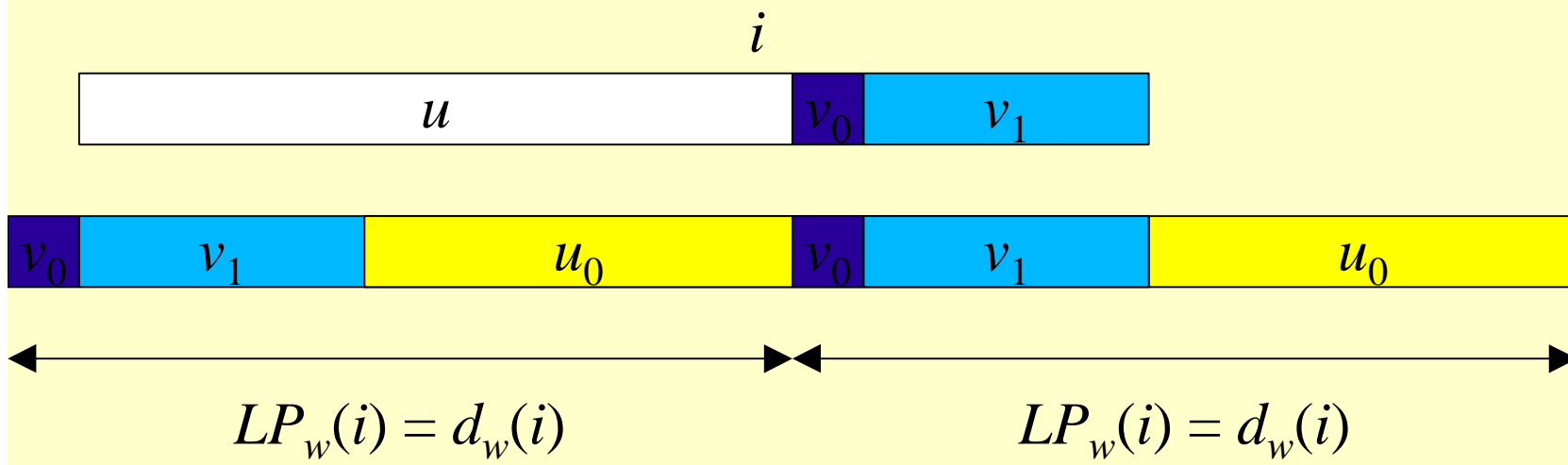






i





Lemma 2: Let $w = uv$ with $|u| < |v|$. If there is no internal square centered at $i = |u|$, then the minimal left external square has period $d_{wR}(|w| - i)$.

Theorem 3: In a word w of length n , all local periods $LP_w(i)$ can be computed in time $O(n)$.