

Calcul linéaire de toutes les périodes locales d'un mot

Thierry Lecroq

ABISS

Université de Rouen - France

Thierry.Lecroq@univ-rouen.fr

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

travail commun avec Jean-Pierre Duval (Rouen),
Roman Kolpakov (Moscou et Liverpool), Gregory
Kucherov (Nancy) et Arnaud Lefebvre (Rouen)

Notations

- Un mot $w = w[1..n]$ de longueur n ;
- On note $pér(w)$ la période de w .

Période locale

Définition 1 : Soit $w = uv$, et $|u| = i$. On dit qu'un carré non vide tt est centré à la position i (ou centré en i) de w ssi les deux conditions suivantes sont satisfaites :

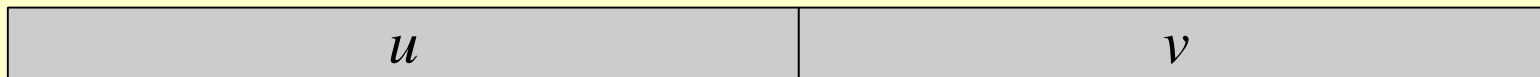
- i. t est un suffixe de u , ou u est un suffixe de t ,
- ii. t est un préfixe de v , ou v est un préfixe de t .



carré interne



carré externe droit



carré externe droit et gauche

Période Locale

Définition 2 : Le plus petit carré non vide centré à la position i de w est appelé le carré local minimal centré en i . La **période locale** à la position i de w , notée $LP_w(i)$, est la période du carré minimal centré en i .

Période Locale

Définition 2 : Le plus petit carré non vide centré à la position i de w est appelé le carré local minimal centré en i . La **période locale** à la position i de w , notée $LP_w(i)$, est la période du carré minimal centré en i .

Remarque : $1 \leq LP_w(i) \leq |w|$.

Période Locale

Définition 2 : Le plus petit carré non vide centré à la position i de w est appelé le carré local minimal centré en i . La **période locale** à la position i de w , notée $LP_w(i)$, est la période du carré minimal centré en i .

Remarque : $1 \leq LP_w(i) \leq pér(w) \leq |w|$.

Théorème de factorisation critique

Théorème : Pour tout mot w , il existe une position i (et la factorisation correspondante $w = uv$ avec $|u| = i$) telle que $LP_w(i) = pér(w)$. De plus, une telle position existe parmi n'importe quelles $pér(w)$ positions consécutives de w .

s-factorisation

Définition 3 : La s -factorisation sans copie chevauchante de w est la factorisation $w = f_1 f_2 \dots f_k$, où les f_i sont définis récursivement comme suit :

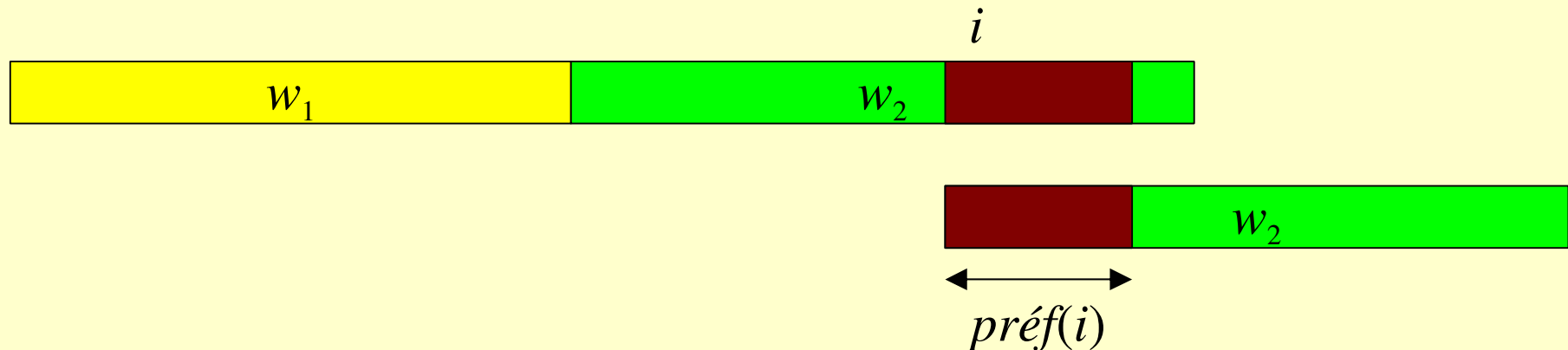
- i. $f_1 = w[1]$,
- ii. supposons calculés $f_1 f_2 \dots f_{i-1}$ ($i \geq 2$), et soit $w[j]$ la lettre qui suit immédiatement $f_1 f_2 \dots f_{i-1}$ (i.e. $j = |f_1 f_2 \dots f_{i-1}| + 1$). Si $w[j]$ n'apparaît pas dans $f_1 f_2 \dots f_{i-1}$, alors $f_i = w[j]$, sinon f_i est le plus long facteur de w commençant à la position j , qui possède une autre occurrence dans $f_1 f_2 \dots f_{i-1}$.

Fonctions d'extension

$$w = w_1[1..m]w_2[1..n]$$

$$\text{préf}(i) = \max \{ j \mid w_2[1..j] = w_2[i..i+j-1] \} \text{ pour } 2 \leq i \leq n$$

et $\text{préf}(n+1) = 0$

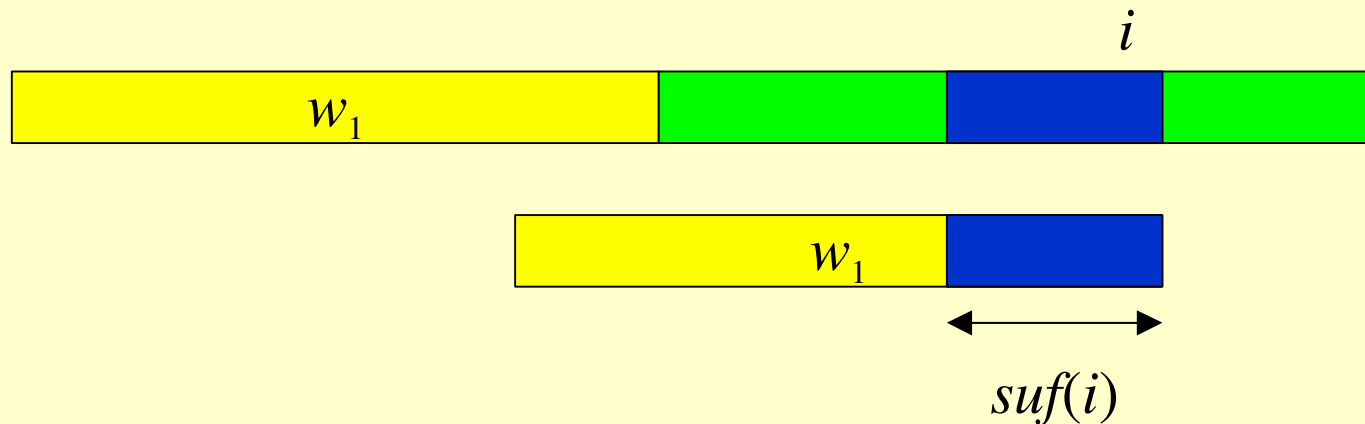


Fonctions d'extension

$$w = w_1[1..m]w_2[1..n]$$

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

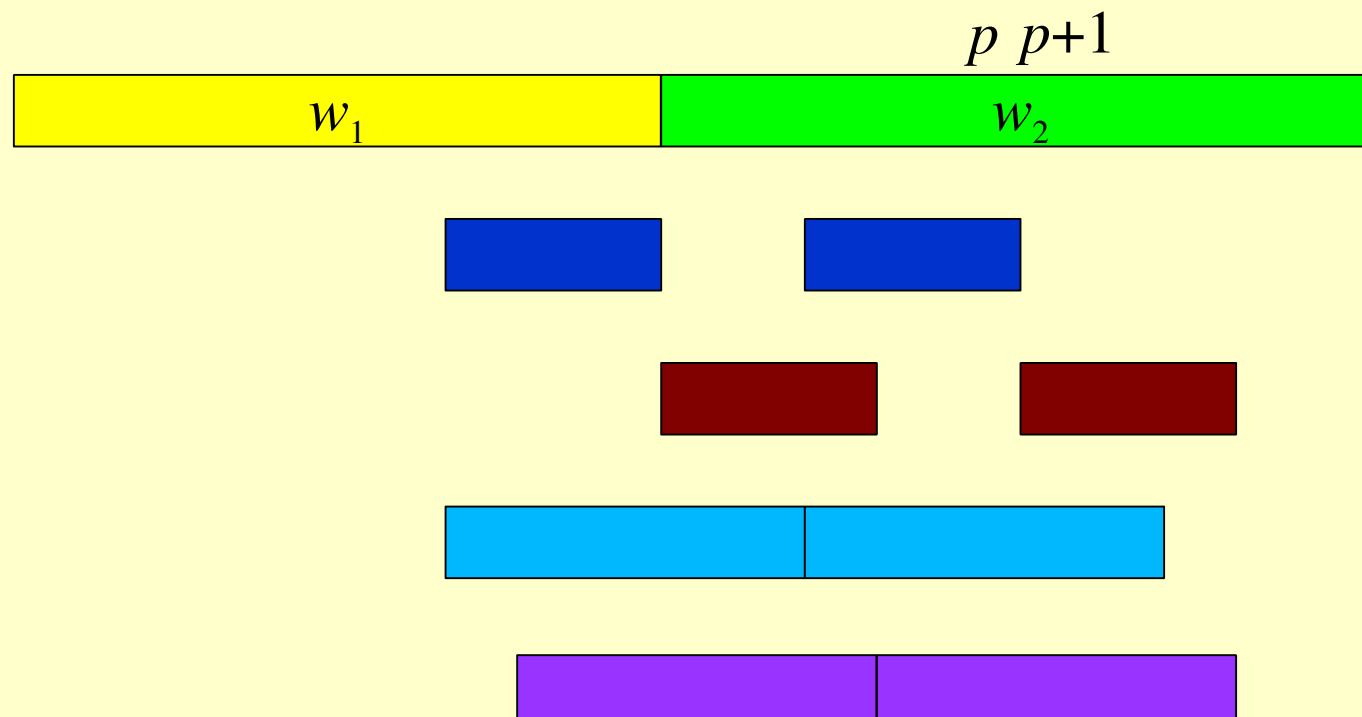
pour $1 \leq i \leq n$



Alors il existe un carré de période p ssi

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

[Main 1989]



À chaque position p où

$$suf(p) + pref(p+1) \geq p$$

est vérifié

il y a une série de carrés centrés à chaque position dans l'intervalle

$$[m - suf(p) + p, m + pref(p+1)] .$$

Cette série de carrés est une répétition maximale dans w [Kolpakov & Kucherov 1999].

Idée générale pour calculer toutes les périodes locales

Deux étapes :

- calcul de tous les carrés internes minimaux ;
- calcul de tous les carrés externes minimaux.

Idée générale pour calculer tous les carrés internes minimaux

- calcul de la s -factorisation et traitement des facteurs un par un de la gauche vers la droite ;
- pour chaque facteur f_r on considère séparément les carrés :
 - qui apparaissent entièrement à l'intérieur de f_r ;
 - qui se terminent dans f_r et s'étendent vers f_{r-1} .

Idée générale pour calculer tous les carrés internes minimaux

- Les carrés du premier type sont calculés en utilisant le fait que f_r possède une copie sur la gauche $\Rightarrow O(|f_r|)$.
- Les carrés du second type sont calculés en utilisant les fonctions d'extension et un lemme établissant que les carrés ne peuvent s'étendre vers la gauche de plus de $|f_r| + 2|f_{r-1}|$ lettres [Main 1989] $\Rightarrow O(|f_{r-1}| + |f_r|)$.
- Au total, trouver tous les carrés internes minimaux dans un mot de longueur n peut être effectué en temps $O(n)$.

Idée pour calculer tous les carrés internes minimaux apparaissant à l'intérieur de f_r

- D'abord on calcule la s -factorisation de w sans copie chevauchante et on garde pour chaque facteur f_r une référence vers sa copie (non chevauchante) à gauche.
- L'algorithme traite tous les facteurs f_r de gauche à droite et calcule pour chaque facteur f_r tous les carrés minimaux **se terminant** en f_r .
- Pour chaque carré minimal interne centré trouvé en position i , $LP_w(i)$ est initialisé.

Idée pour calculer tous les carrés internes minimaux apparaissant à l'intérieur de f_r

Après que la totalité du mot ait été traité, les positions i pour lesquelles les valeurs $LP_w(i)$ n'ont pas été assignées sont celles pour lesquelles il n'existe pas de carré minimal centré en i et $LP_w(i)$ est calculé avec une autre technique.

Calcul des carrés internes minimaux apparaissant à l'intérieur de f_r

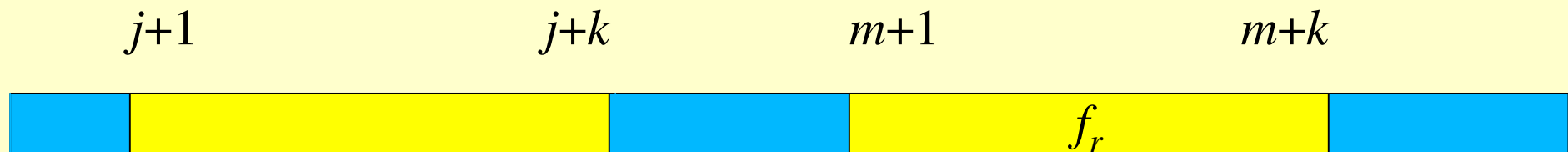
- Soit $f_r = w[m + 1 .. m + k]$ le facteur et $w[j + 1 .. j + k]$ sa copie gauche ($j + k \leq m$)
- Si pour une position $m + i$ ($1 \leq i < k$) le carré minimal centré en $m + i$ apparaît entièrement à l'intérieur du facteur f_r (i.e. $LP_w(m + i) \leq \min \{ i, k - i \}$)

- alors

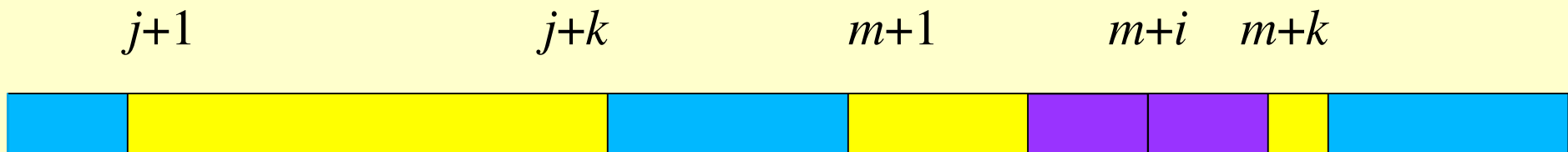
$$LP_w(m + i) = LP_w(j + i)$$

- $LP_w(j + i)$ a déjà été calculé donc on peut calculer toutes les valeurs $LP_w(m + i) \leq \min \{ i, k - i \}$ en temps $O(|f_r|)$.

Calcul des carrés internes minimaux apparaissant à l'intérieur de f_r



Calcul des carrés internes minimaux apparaissant à l'intérieur de f_r



Calcul des carrés internes minimaux apparaissant à l'intérieur de f_r

$j+1$

$j+i$

$j+k$

$m+1$

$m+i$

$m+k$



Il reste à trouver les valeurs $LP_w(m + i)$ qui correspondent aux carrés minimaux qui se terminent en f_r et s'étendent à gauche de la frontière entre f_r et f_{r-1} .

Il reste à trouver les valeurs $LP_w(m + i)$ qui correspondent aux carrés minimaux qui se terminent en f_r et s'étendent à gauche de la frontière entre f_r et f_{r-1} .

On partitionne ces carrés en deux catégories :

- ceux centrés à l'intérieur de f_r ;
- ceux centrés à gauche de f_r .

- On se concentre sur les carrés centrés en des positions dans $[m, m + k - 1]$ et commençant en des positions $\leq m$ et se terminant à l'intérieur de f_r .
- On calcule ces carrés en ordre croissant des périodes en utilisant les fonctions d'extension.
- Pour chaque $p \in [1, k - 1]$ on calcule la série de tous les carrés de période p centré en des positions dans $[m, m + k - 1]$ commençant en des positions $\leq m$ et se terminant à l'intérieur de f_r .

Supposons que :

- nous venons de calculer une série de carrés de période p ;
- $q < p$ est la plus grande période pour laquelle des carrés ont été précédemment trouvé.

$$p \geq 2q$$

- Si $p \geq 2q$ alors on vérifie pour chaque carré de la série s'il est minimal ou non en testant la valeur de $LP_w(i)$.
- Si ce carré n'est pas minimal, alors son centre i s'est déjà vu assigné une valeur $LP_w(i)$.
- Si aucune valeur n'a été précédemment assignée alors nous avons trouvé un carré minimal centré en i .

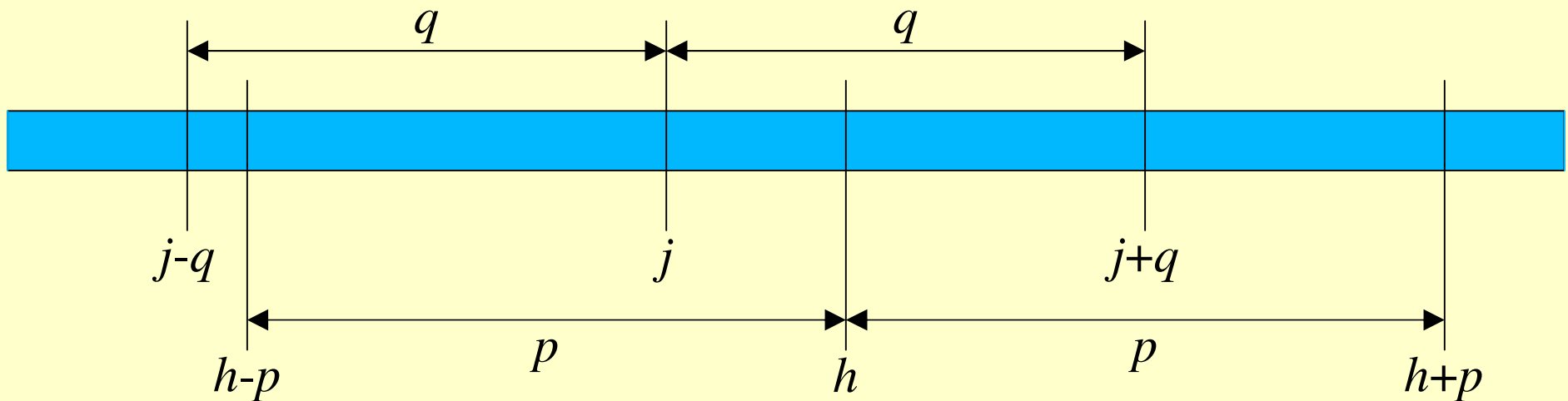
- Il y a au plus p carrés de période p (leurs centres appartenant à $[m, m + p - 1]$)
- Les vérifier tous nécessite au plus $2(p-q)$ tests individuels (puisque $q \leq p/2$ et $p-q \geq p/2$)

$$p < 2q$$

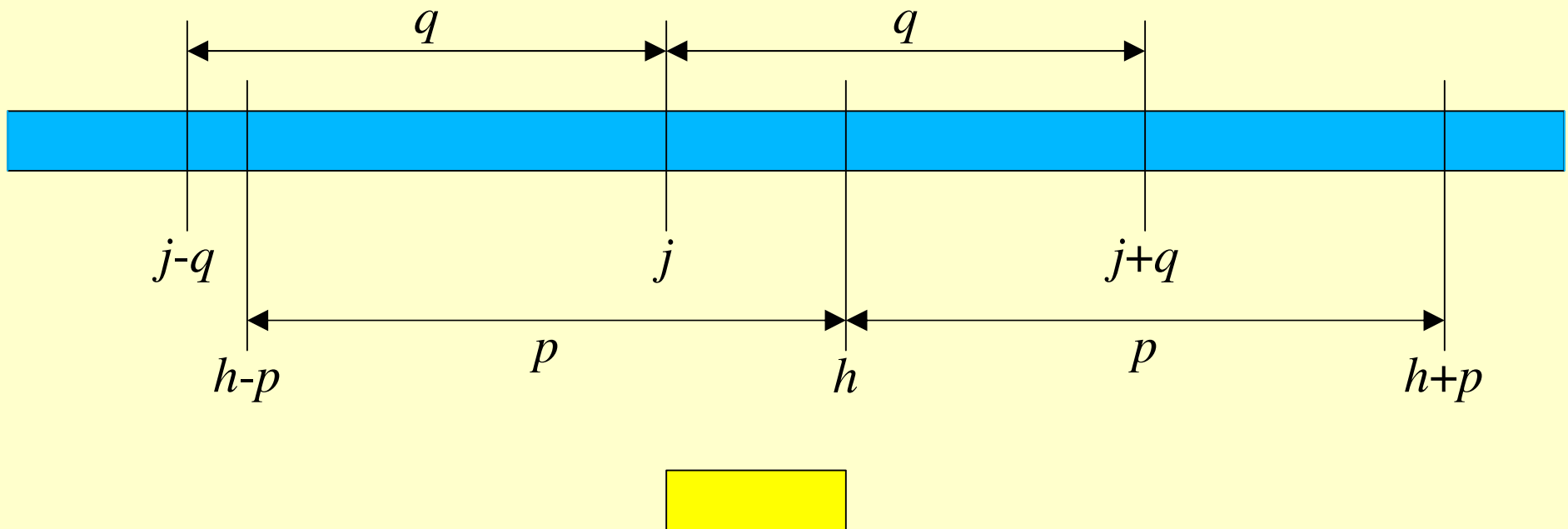
- considérons un carré $s_q = w[j - q + 1 .. j + q]$ de période q et centré en j
- Nous disons que nous avons besoin de vérifier la minimalité uniquement des carrés s_p de période p qui ont leur centre h vérifiant une des inégalités suivantes :
 - $|h - j| \leq p - q$ ou
 - $h \geq j + q$

h est situé à distance $p - q$ de j ou au-delà de la fin du carré s_q .

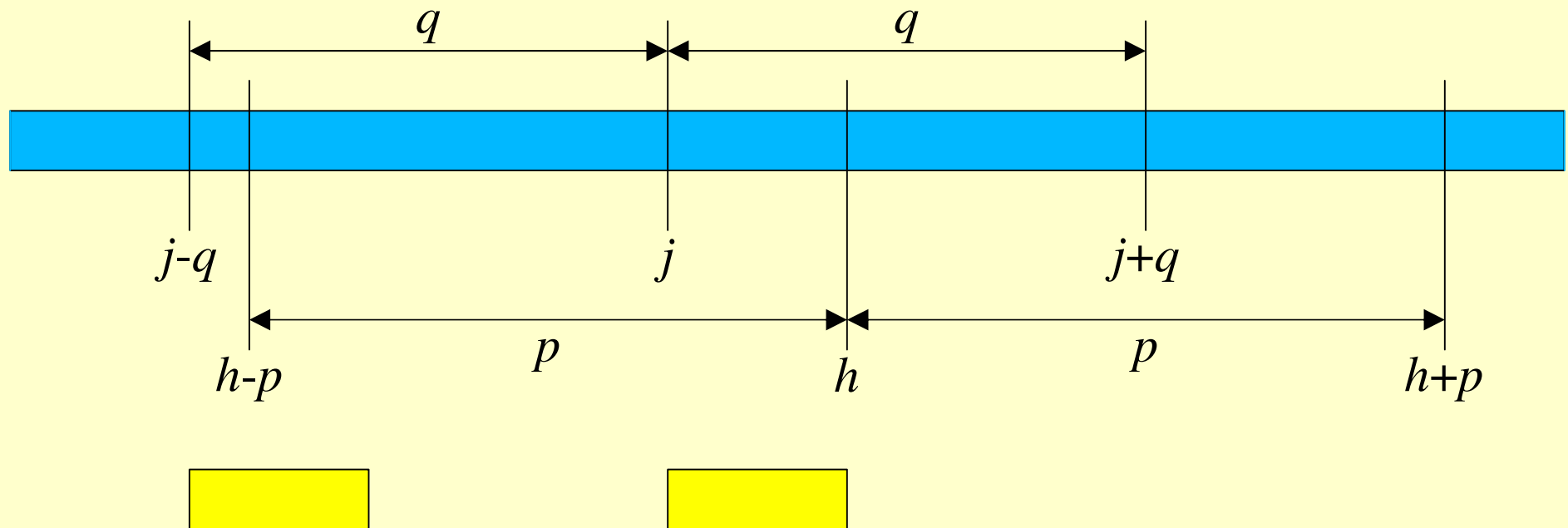
- Preuve par l'absurde : supposons que $|h - j| > p - q$ et $h < j + q$
- 2 cas symétriques : $h > j$ ou $h < j$
- $h > j$



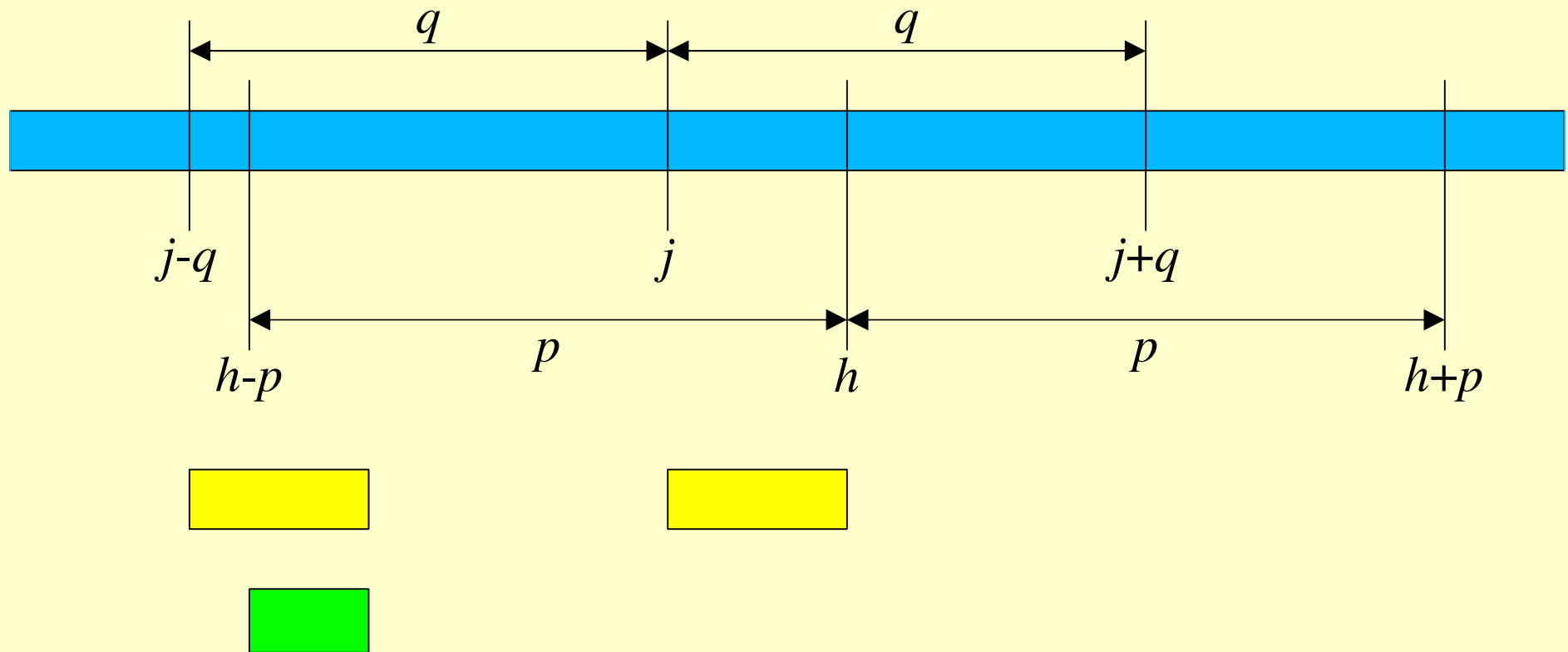
- Preuve par l'absurde : supposons que $|h - j| > p - q$ et $h < j + q$
- 2 cas symétriques : $h > j$ ou $h < j$
- $h > j$



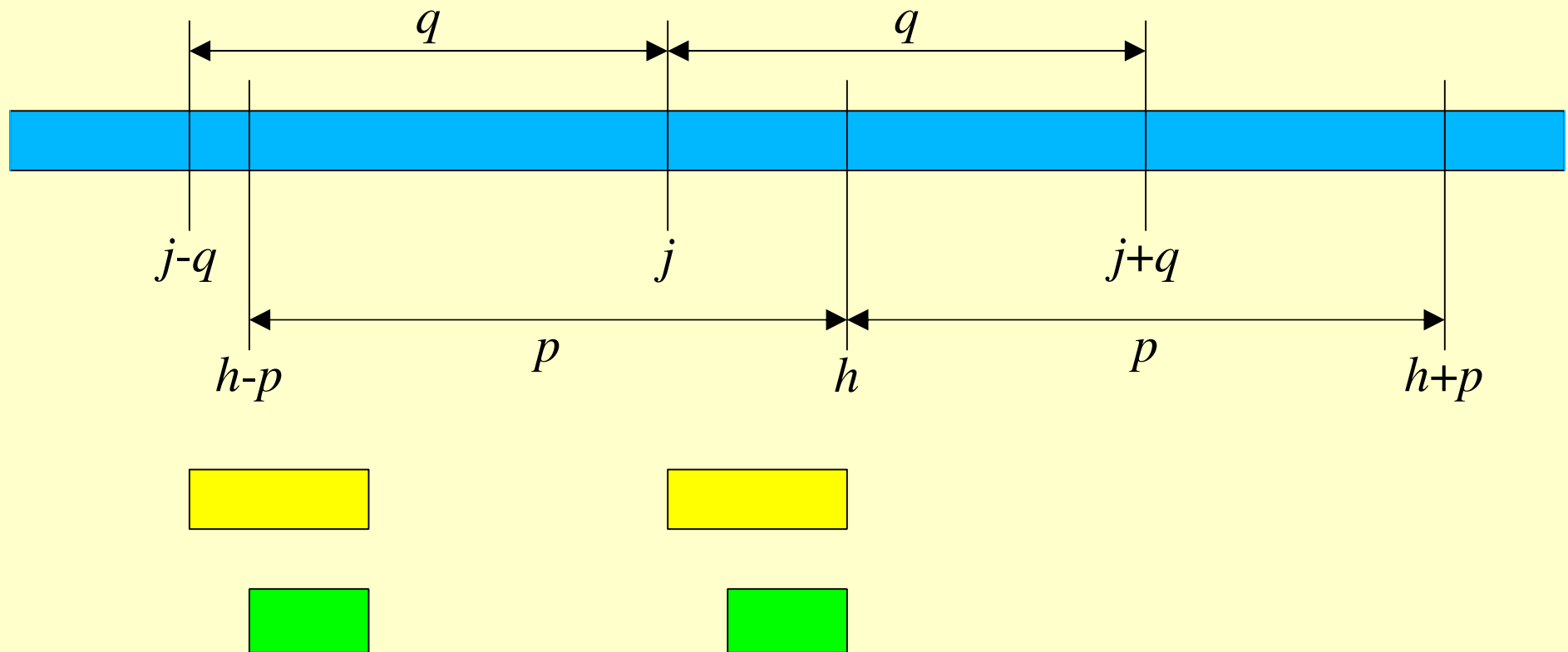
- Preuve par l'absurde : supposons que $|h - j| > p - q$ et $h < j + q$
- 2 cas symétriques : $h > j$ ou $h < j$
- $h > j$



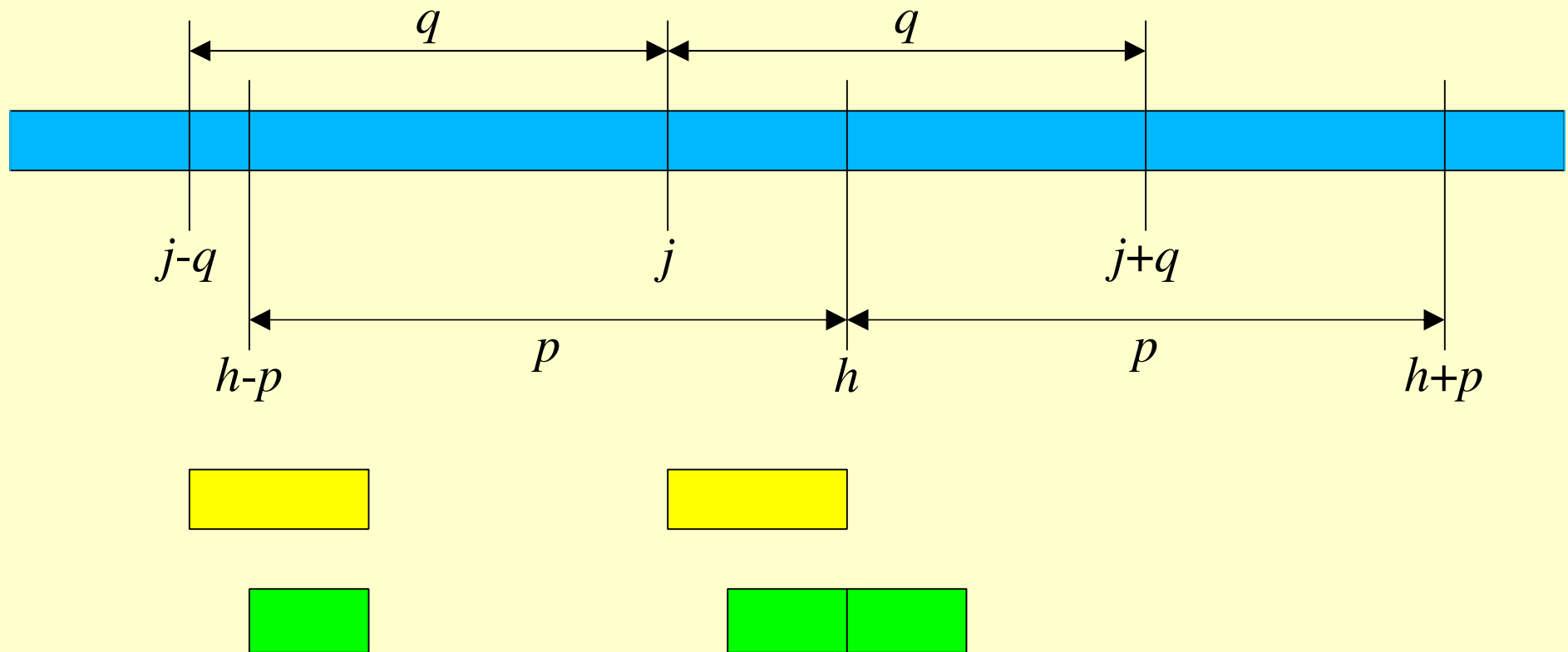
- Preuve par l'absurde : supposons que $|h - j| > p - q$ et $h < j + q$
- 2 cas symétriques : $h > j$ ou $h < j$
- $h > j$



- Preuve par l'absurde : supposons que $|h - j| > p - q$ et $h < j + q$
- 2 cas symétriques : $h > j$ ou $h < j$
- $h > j$



- Preuve par l'absurde : supposons que $|h - j| > p - q$ et $h < j + q$
- 2 cas symétriques : $h > j$ ou $h < j$
- $h > j$



- il y a au plus $2(p-q)$ carrés s_p vérifiant $|h - j| \leq p - q$
- il y a au plus $p - q$ carrés s_p vérifiant $h \geq j + q$ puisque s_p doit commencer avant m ($h \leq m + p$)
- il y a au plus $3(p - q)$ carrés de période p dont on doit vérifier la minimalité
- il y a au plus $O(|f_r|)$ carrés à vérifier pour le facteur courant

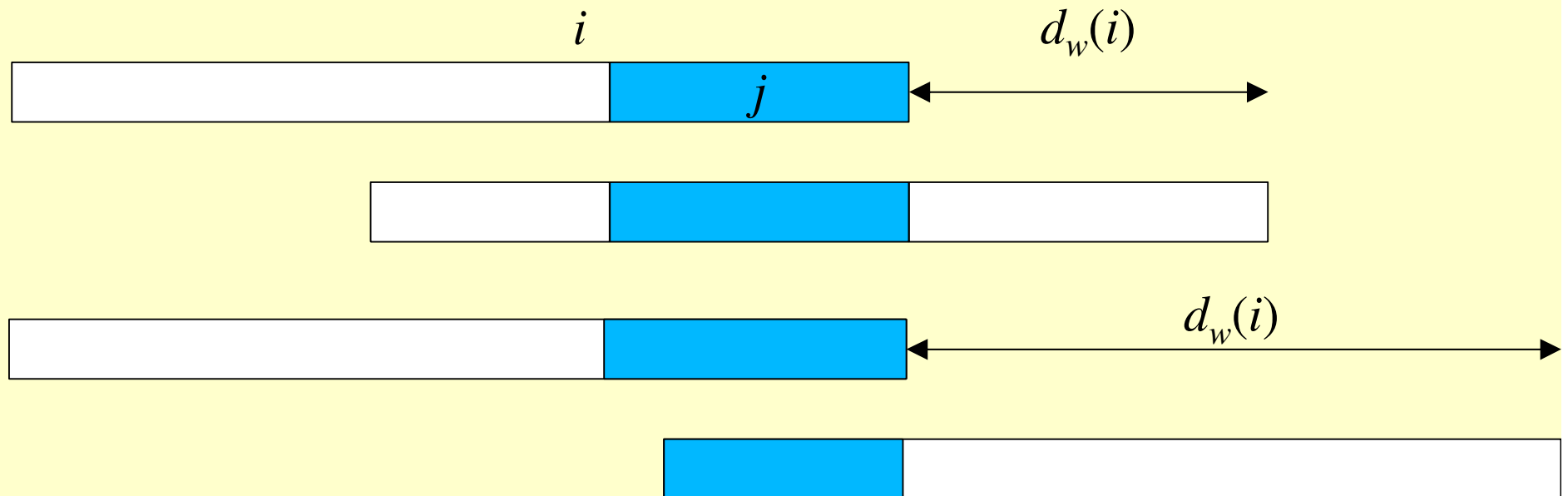
- Un argument similaire s'applique aux carrés centrés à gauche de f_r
- $O(|f_{r-1}| + |f_r|)$ tests pour les carrés franchissant la frontière entre f_{r-1} et f_r
- $O(|f_r|)$ pour les carrés à l'intérieur de f_r

Théorème 2 : Tous les carrés internes minimaux d'un mot de longueur n peuvent être calculés en temps $O(n)$.

Fonction de décalage simplifiée de Boyer-Moore

Définition 4 : Pour un mot w de longueur n la fonction de décalage simplifiée de Boyer-Moore est définie comme suit :

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

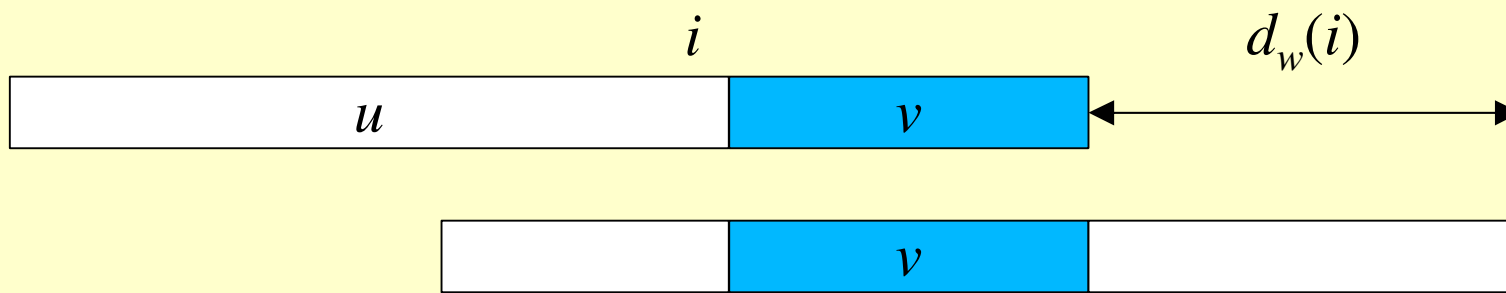


Lemme 1 : Soit $w = uv$ avec $|u| \geq |v|$. S'il n'y a pas de carré interne centré en $i = |u|$, alors le carré externe minimal droit a pour période $d_w(i)$.

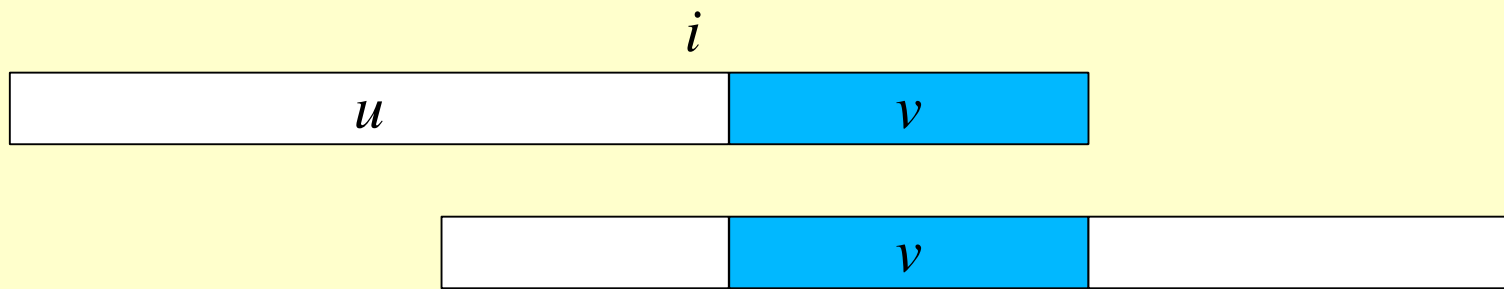
Preuve :

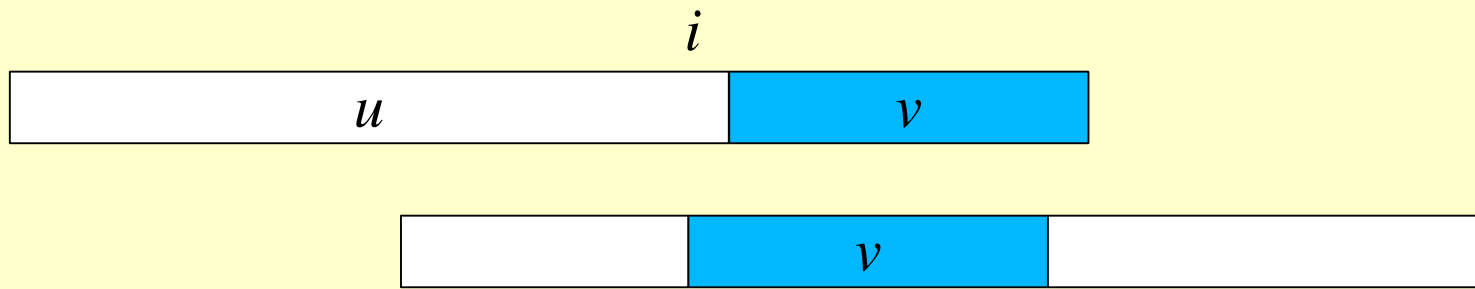
2 cas

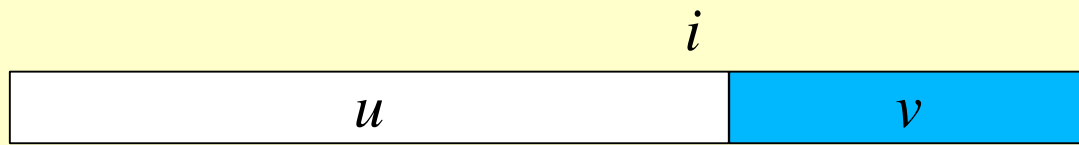
- $d_w(i) \leq |u|$;
- $d_w(i) > |u|$.

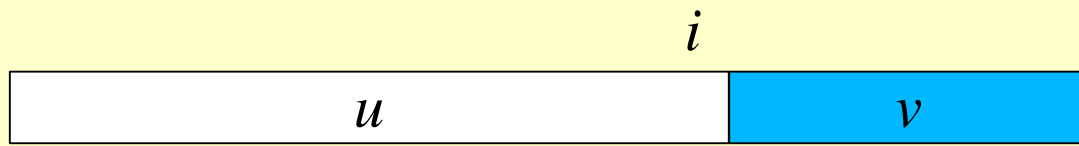


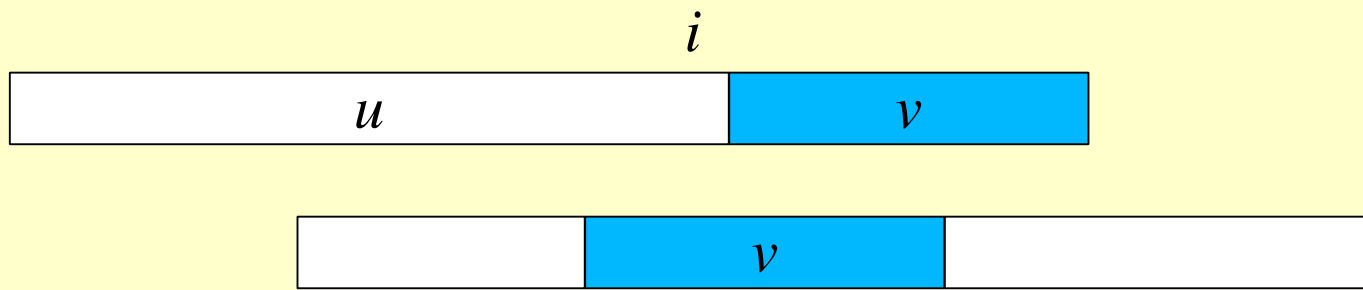
$$d_w(i) \leq |u|$$

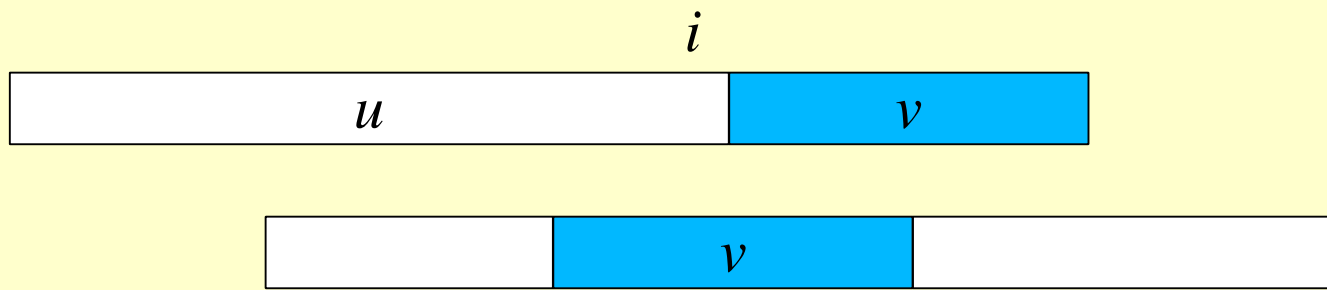


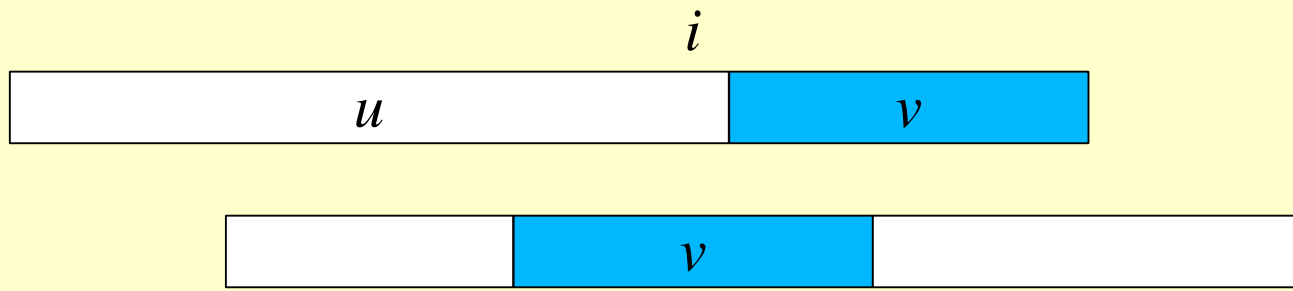


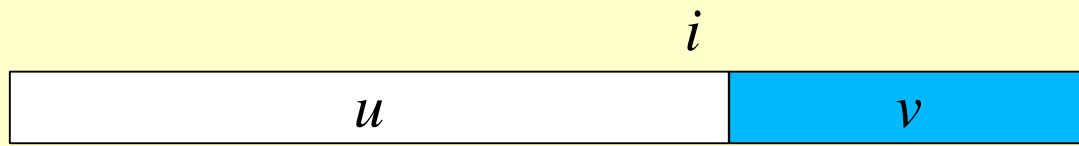


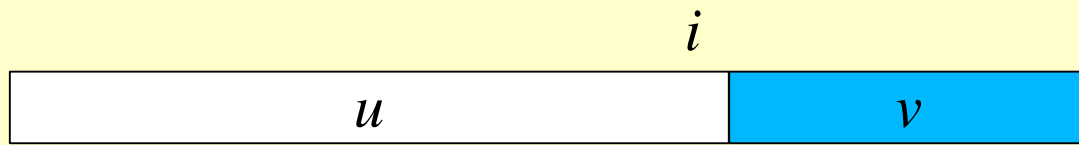


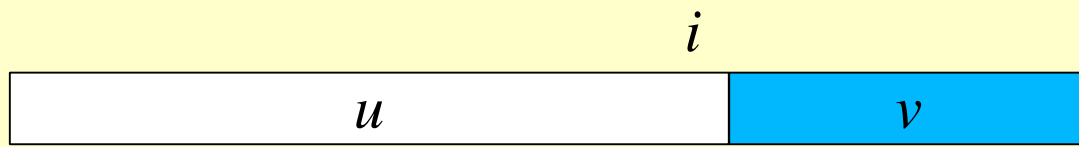


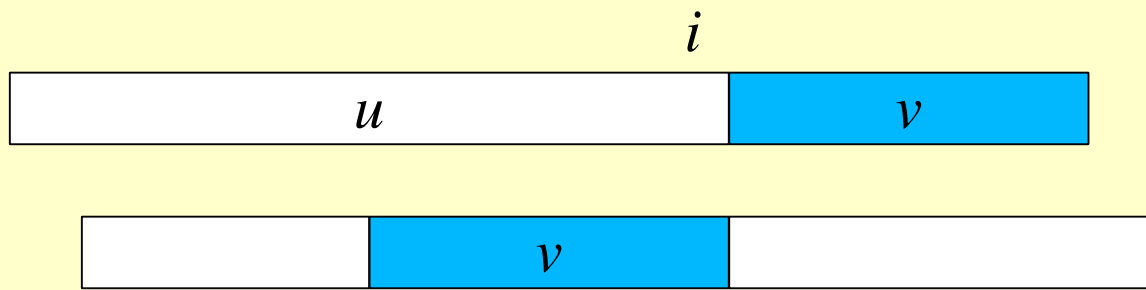


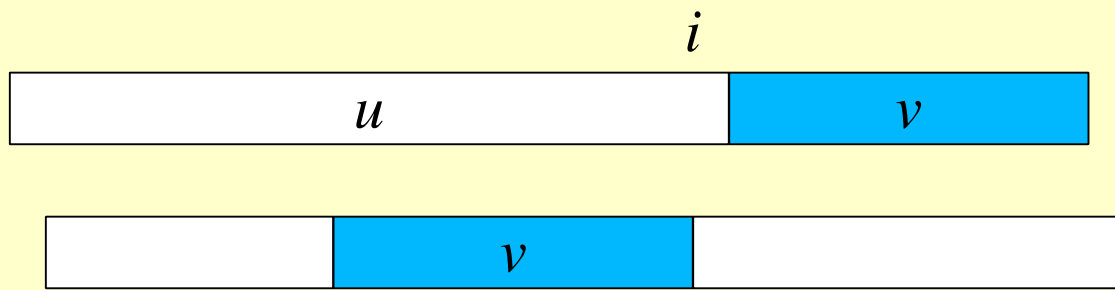


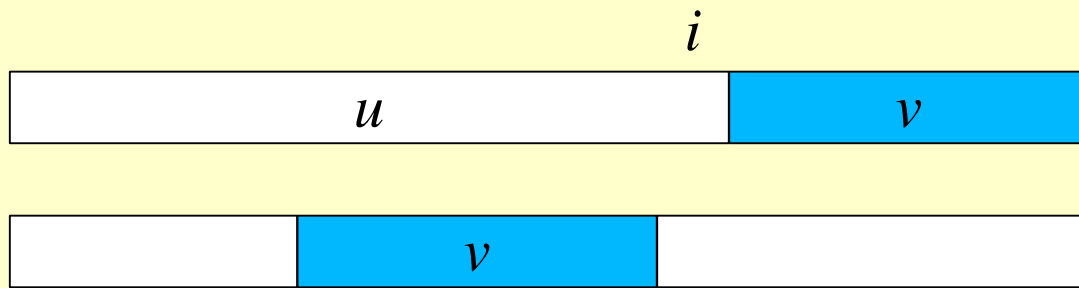


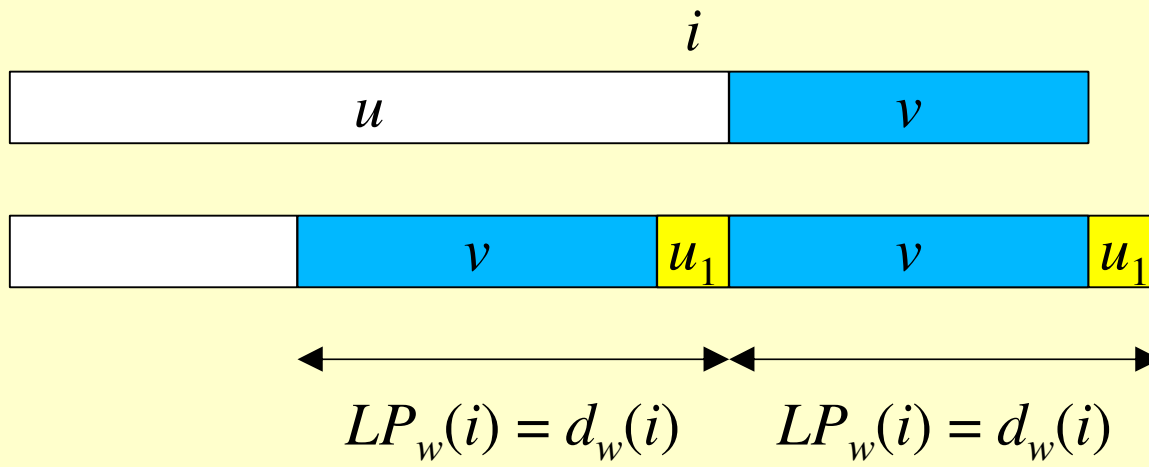


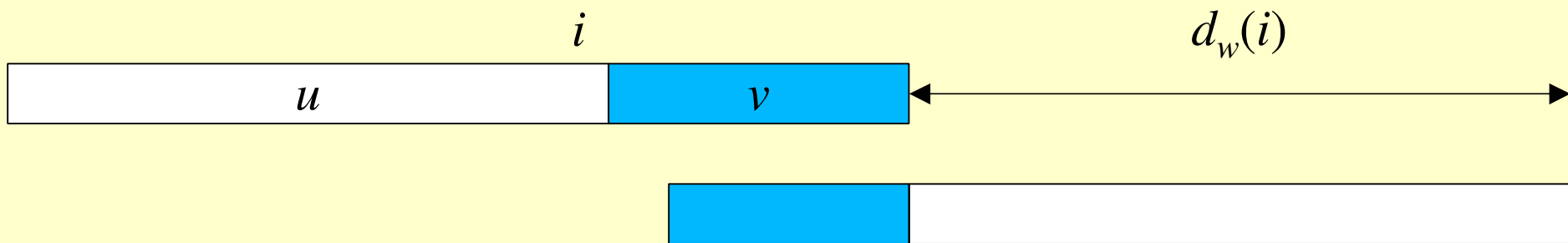




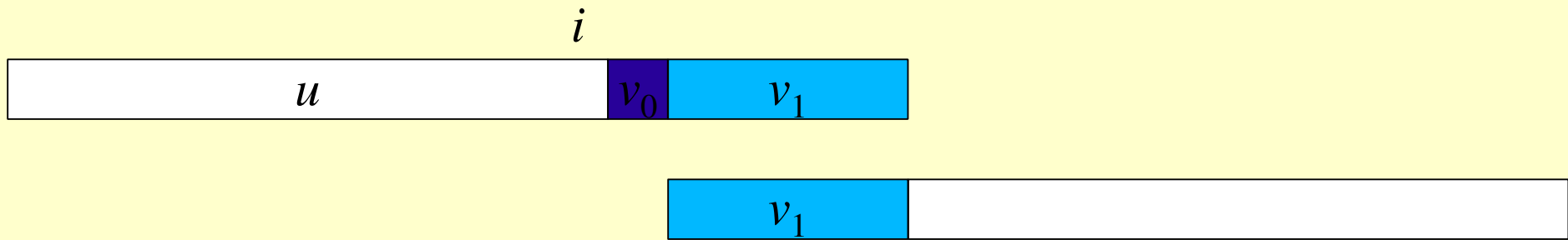


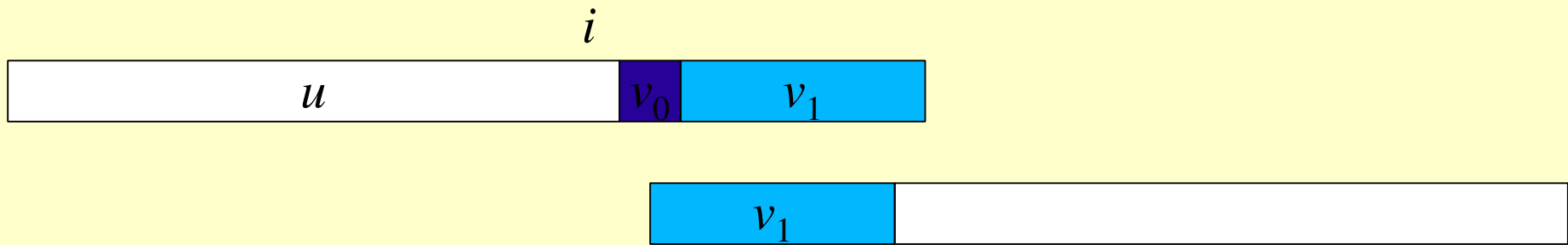


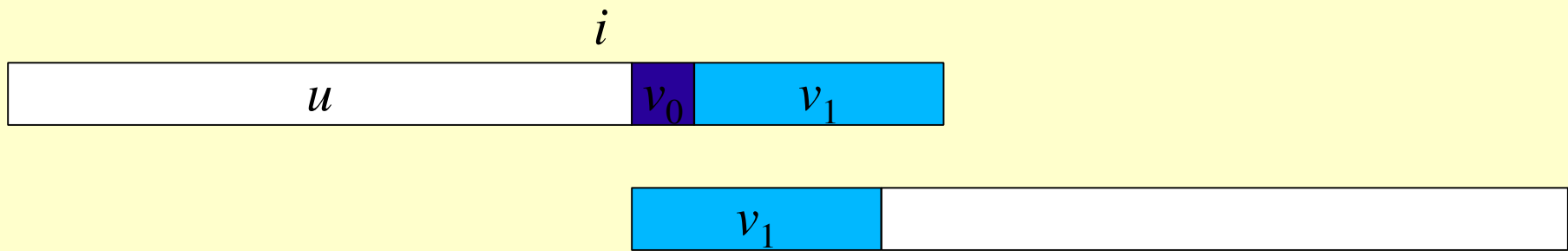


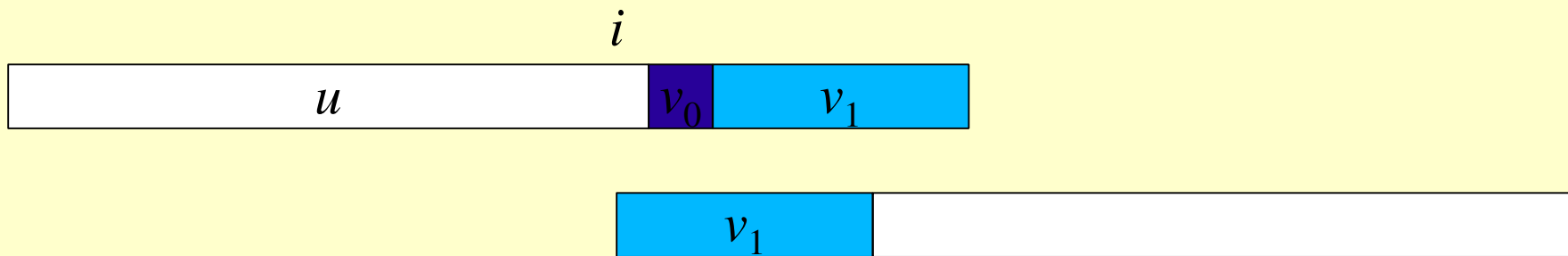


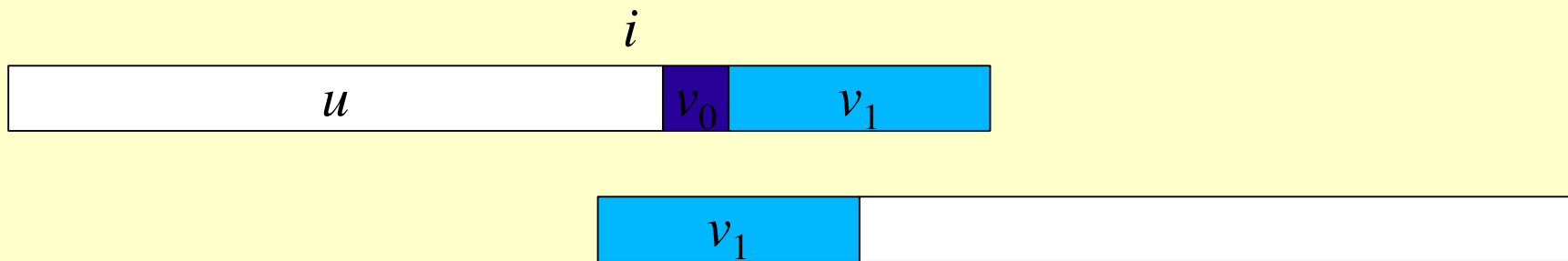
$$d_w(i) > |u|$$

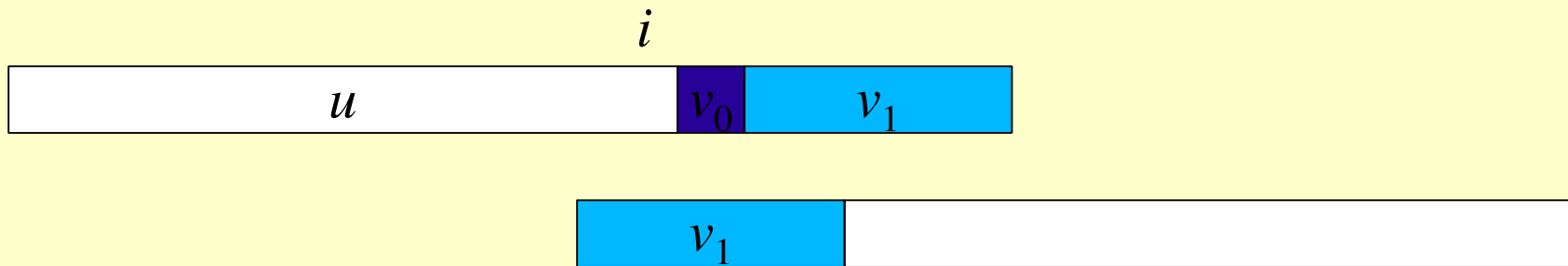


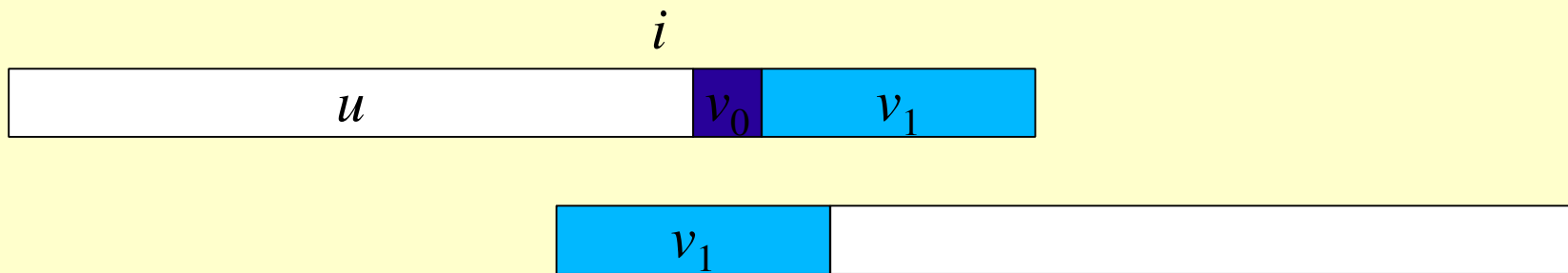




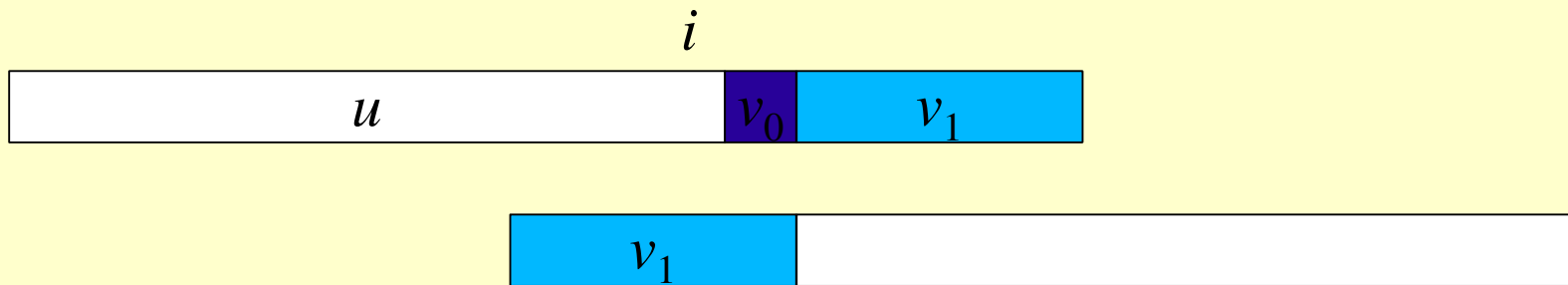


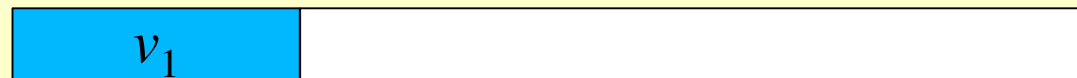


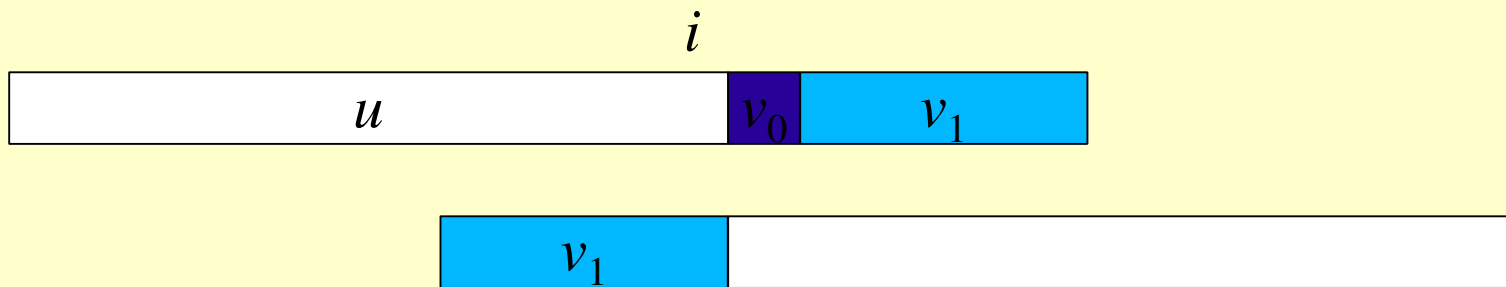


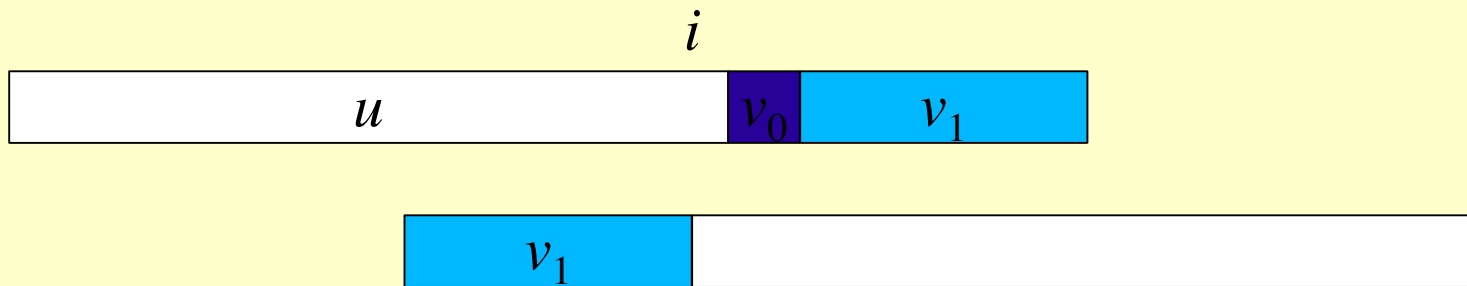


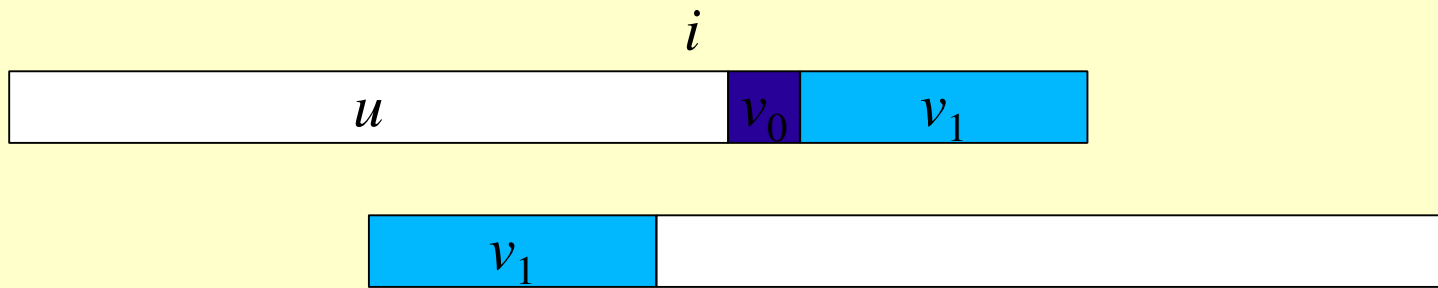


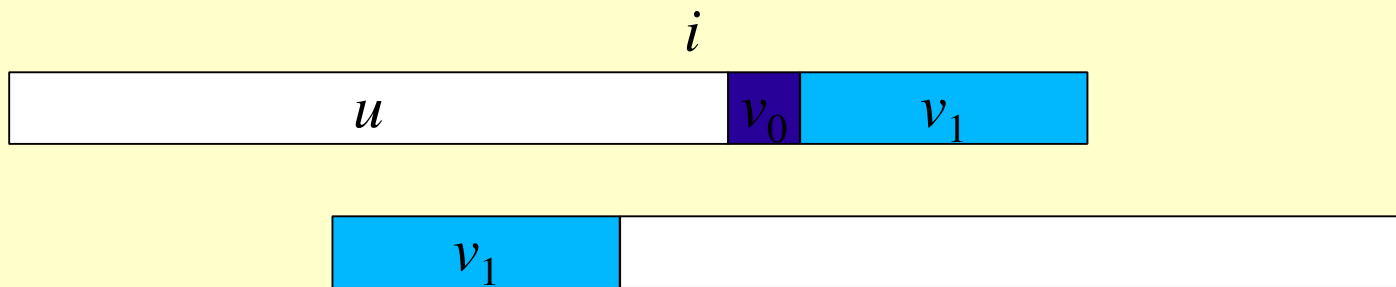




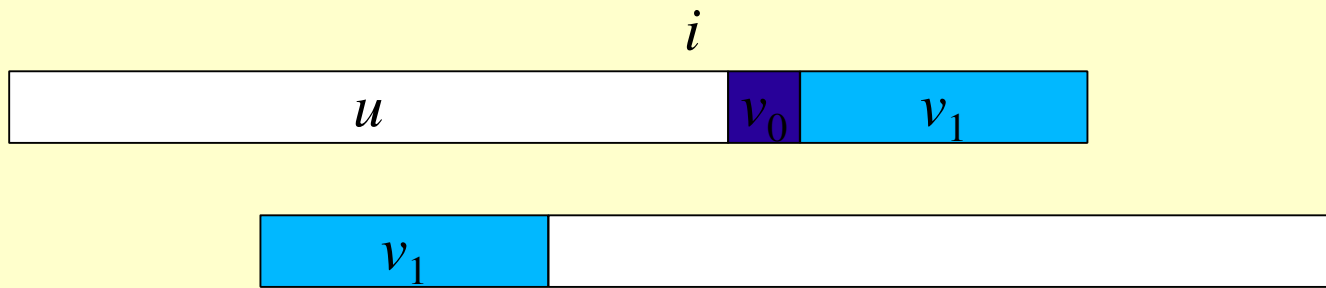


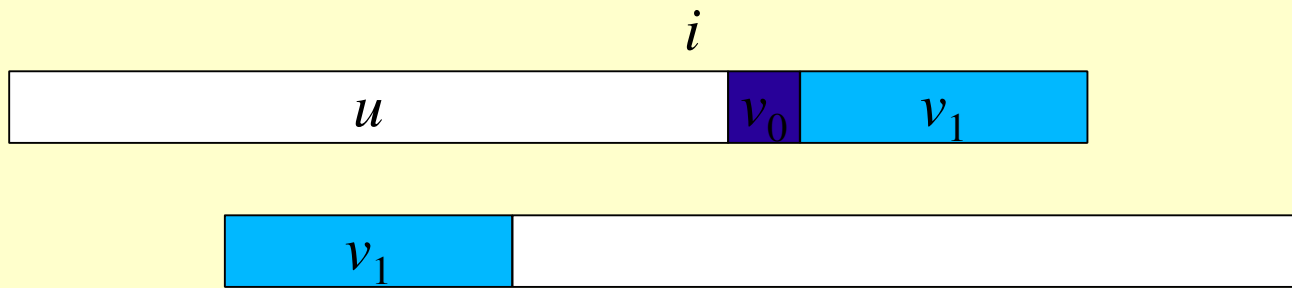




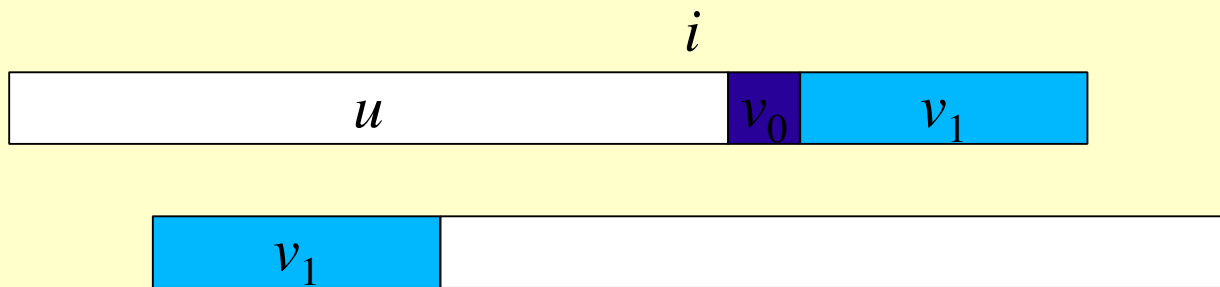


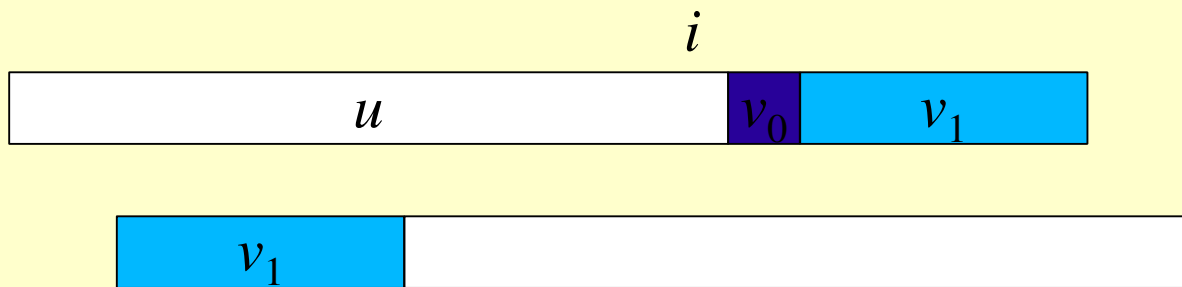


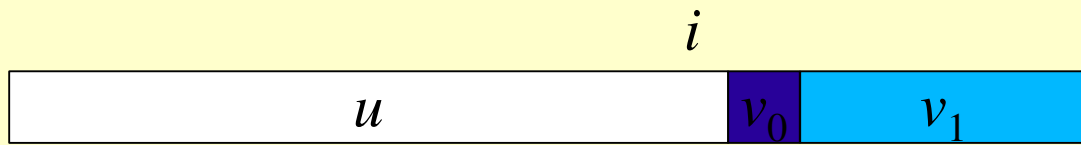


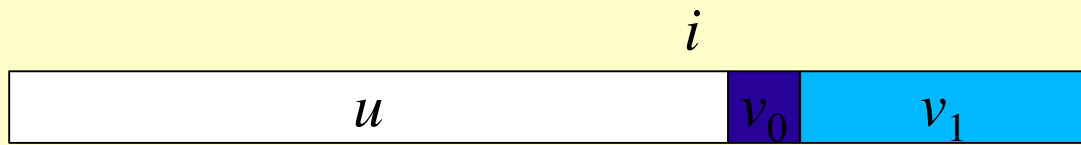


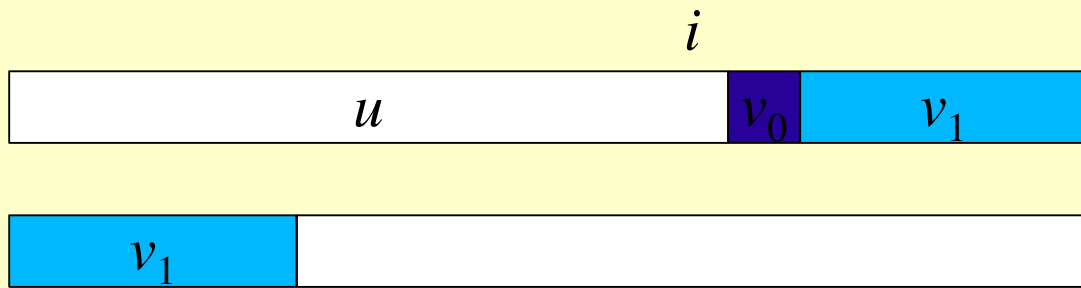




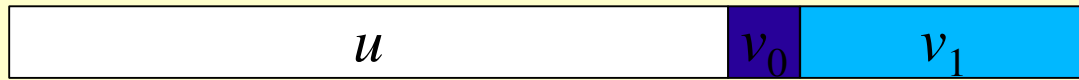


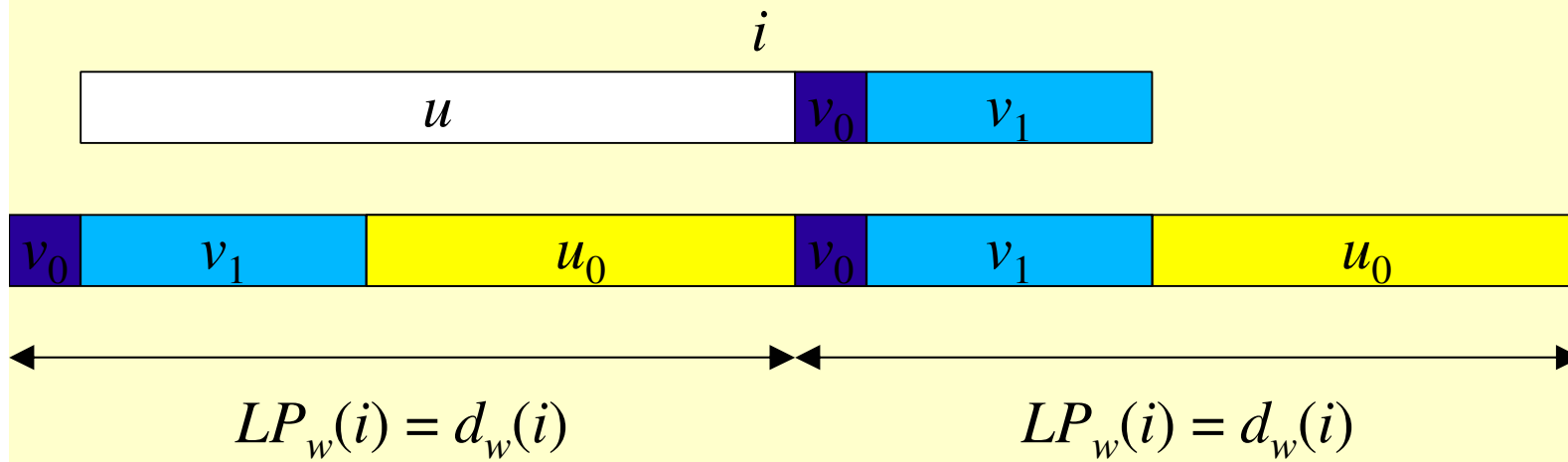






i





Remarque

Si $d_w(i) > i$ alors

$$d_w(i) = \text{per}(w)$$

et

$$\text{per}(w) > |w|/2.$$

Lemme 2 : Soit $w = uv$ avec $|u| < |v|$. S'il n'y a pas de carrés internes centré en $i = |u|$, alors le carré externe minimal gauche a pour période $d_{wR}(|w|-i)$.

Théorème 3 : Toutes les périodes locales d'un mot de longueur n peuvent être calculées en temps $O(n)$.

Conclusions

- Le calcul de toutes les périodes locales d'un mot peut être effectué en temps linéaire ;
- Cela inclut le calcul de la période globale (Théorème de factorisation critique) ;
- Cela permet de trouver toutes les factorisations critiques d'un mot.

Perspectives

- génération et caractérisation de tous les ensembles de périodes locales ;
- extension aux périodes locales en utilisant une distance (Hamming, édition, ...).