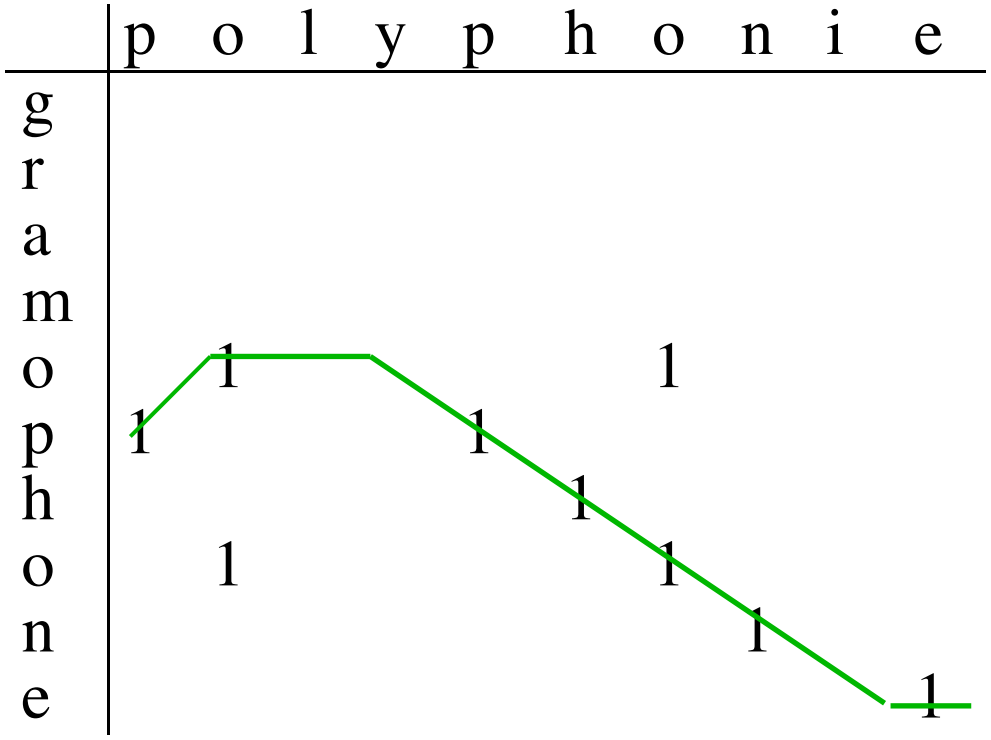
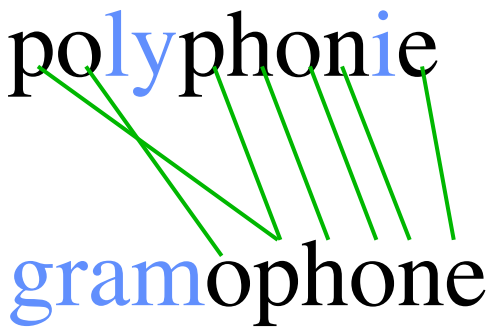
Homologies approximatives

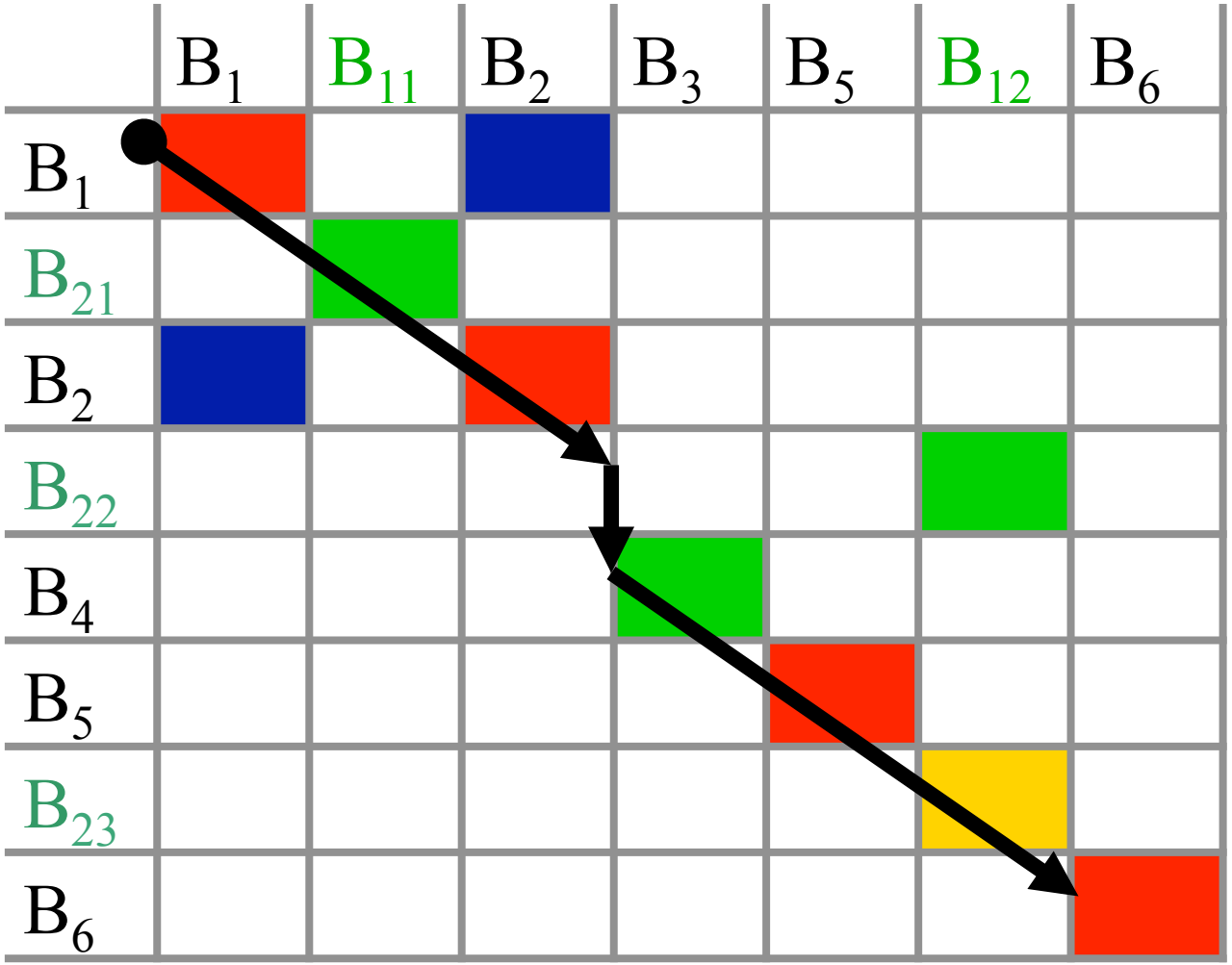




Utilisation de **Distances d'édition**

LIP6 CNRS

Calcul des transformations, *distance d'édition*



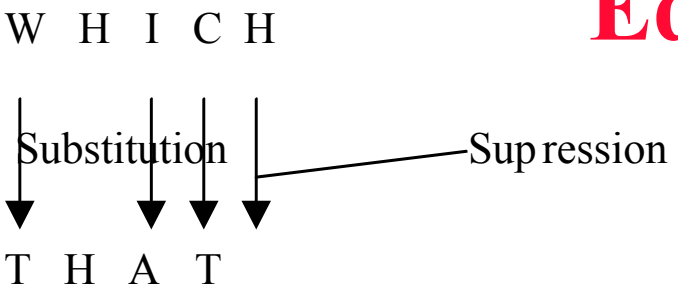
Couleur	Coût
Rouge	0
Jaune	1
Vert	2
Bleu	3
Blanc	4

Coûts = Distance entre blocs

Calcul de la distance d'édition = minimisation de la somme des coûts des transformations

Distances d ’édition

Édition = Transformation



Ensemble d’édition standard: {*insertion, destruction, remplacement*}

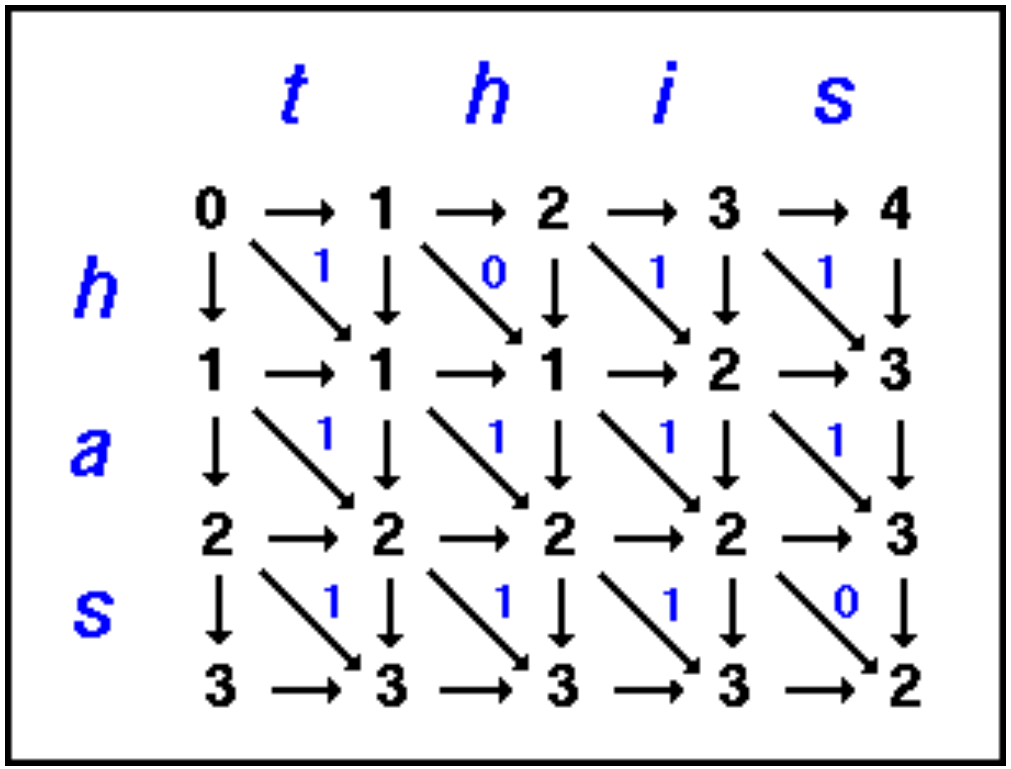
Distance d ’édition

=

Nombre minimal d ’édition

Un exemple

- En supposant que les coûts d'addition, de destruction et de substitutions sont égaux à 1, la distance entre '*this*' et '*has*' est 2

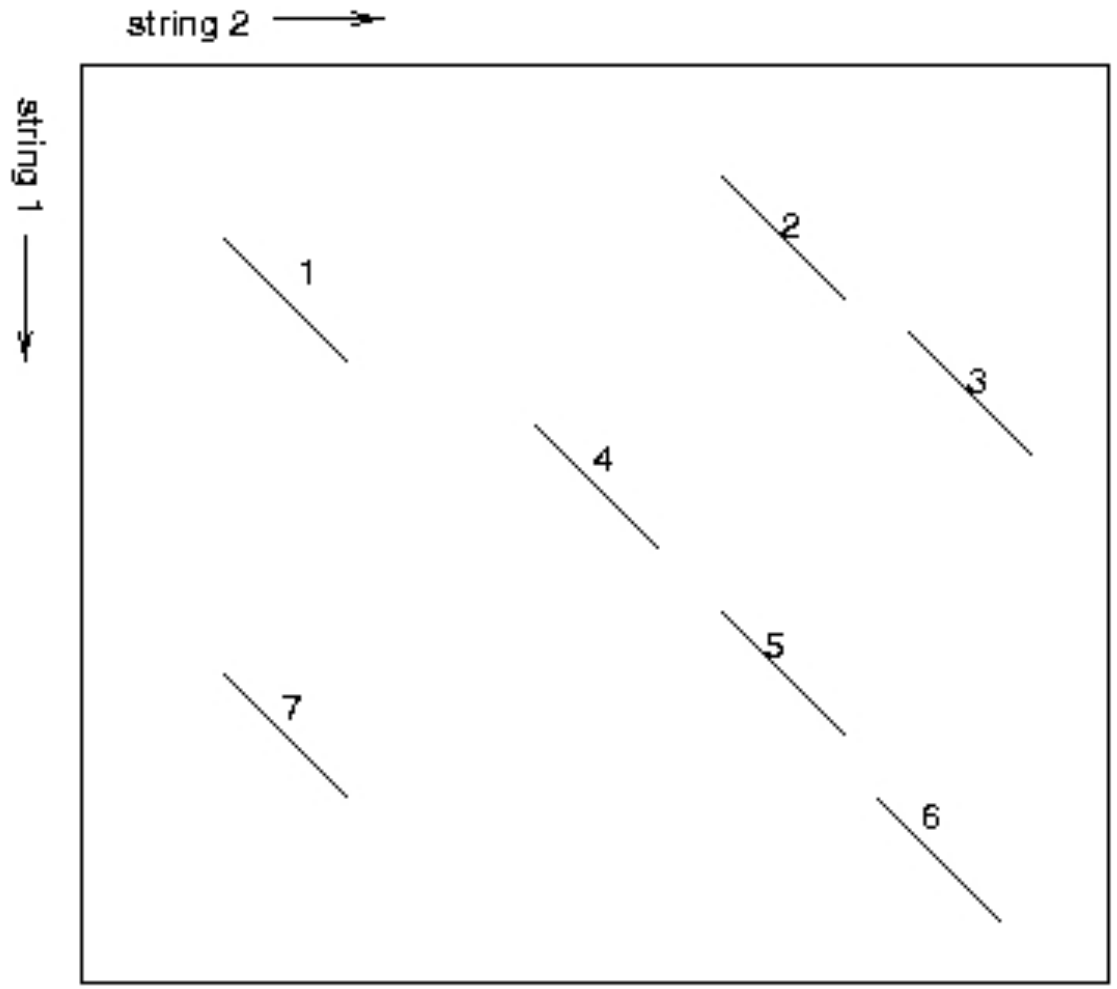


Un autre exemple

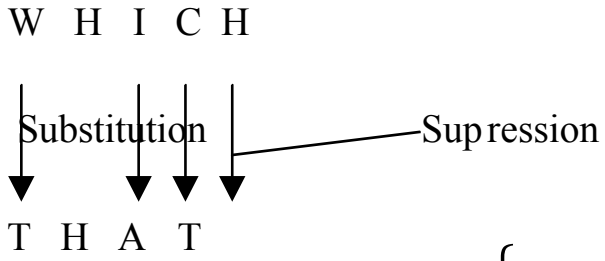
- Distance ‘meilenstein’ et ‘levenshtein’
- Sous-chaînes communes: diagonales
  - ‘ens’
  - ‘tein’

		m	e	i	l	e	n	s	t	e	i	n
	0	1	2	3	4	5	6	7	8	9	10	11
l	1	1	2	3	3	4	5	6	7	8	9	10
e	2	2	1	2	3	3	4	5	6	7	8	9
v	3	3	2	2	3	4	4	5	6	7	8	9
e	4	4	3	3	3	3	4	5	6	6	7	8
n	5	5	4	4	4	4	3	4	5	6	7	7
s	6	6	5	5	5	5	4	3	4	5	6	7
h	7	7	6	6	6	6	5	4	4	5	6	7
t	8	8	7	7	7	7	6	5	4	5	6	7
e	9	9	8	8	8	7	7	6	5	4	5	6
i	10	10	9	8	9	8	8	7	6	5	4	5
n	11	11	10	9	9	9	8	8	7	6	5	4

Sous-chaînes communes



Distances d ’édition *(exemple)*



Ensemble d’édition standard
(limité aux listes)

$$\text{edit}_{As}(x[1..i],y[1..j]) = \min\begin{cases} \text{edit}_{As}(x[1..i-1],y[1..j]) + deletion(x(i)) \\ \text{edit}_{As}(x[1..i-1],y[1..j-1]) + substitution(x(i),y(j)) \\ \text{edit}_{As}(x[1..i],y[1..j-1]) + insertion(y(j)) \end{cases}$$

		T	H	A	T
	0	$del(T)$			
W	$ins(W)$	$subst(W,T)$			
H			$subst(W,T)$		
I				$subst(W,T)+subst(A,I)$	
C				$subst(W,T)+subst(A,I)$ $+ins(C)$	$subst(W,T)+subst(A,I)$ $+subst(T,C)$
H					$subst(W,T)+subst(A,I)$ $+\min\begin{cases} subst(T,C)+ins(H) \\ ins(C)+subst(T,H) \end{cases}$

Calcul effectif

- Complexité théorique: **n*m**
- Algorithme efficace: évaluation paresseuse, uniquement lorsque nécessaire

Evaluation paresseuse

		A	C	G	T	A	C	G	T	A	C	G	T
	0												
A		0											
C			0										
A													
T													
A													
C													
T													
T													
G													
T													
A													
C													
T													

Evaluation paresseuse

		A	C	G	T	A	C	G	T	A	C	G	T
	0	1											
A	1	0	1										
C		1	0	1									
A			1	1									
T													
A													
C													
T													
T													
G													
T													
A													
C													
T													

Evaluation paresseuse

		A	C	G	T	A	C	G	T	A	C	G	T
	0	1											
A	1	0	1										
C		1	0	1									
A			1	1									
T					1								
A						1							
C							1						
T													
T													
G													
T													
A													
C													
T													

Evaluation paresseuse

		A	C	G	T	A	C	G	T	A	C	G	T
	0	1	**2**										
A	1	0	1	**2**									
C	**2**	1	0	1	**2**								
A		**2**	1	1	**2**								
T			**2**	**2**	1	**2**							
A					**2**	1	**2**						
C						**2**	1	**2**					
T							**2**	**2**	**2**				
T									**2**				
G													
T													
A													
C													
T													

Evaluation paresseuse

		A	C	G	T	A	C	G	T	A	C	G	T
	0	1	2	**3**									
A	1	0	1	2	**3**								
C	2	1	0	1	2	**3**							
A	**3**	2	1	1	2	**3**							
T		**3**	2	2	1	2	**3**						
A			**3**	**3**	2	1	2	**3**					
C					**3**	2	1	2	**3**				
T						**3**	2	2	2	**3**			
T							**3**	**3**	2	**3**			
G								**3**	**3**	**3**			
T									**3**				
A										**3**			
C											**3**		
T													

Evaluation paresseuse

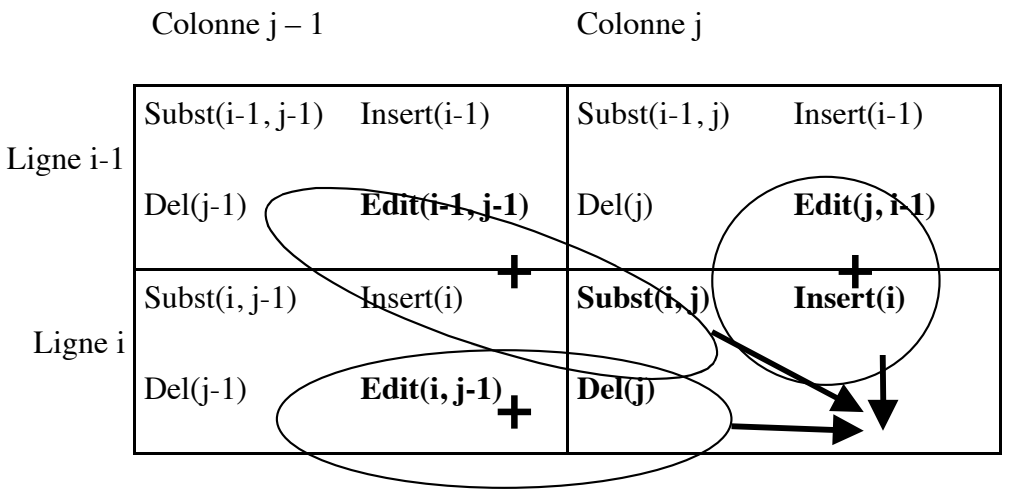
		A	C	G	T	A	C	G	T	A	C	G	T
	0	1	2	3	**4**								
A	1	0	1	2	3	**4**							
C	2	1	0	1	2	3	**4**						
A	3	2	1	1	2	3	**4**						
T	**4**	3	2	2	1	2	3	**4**					
A		**4**	3	3	2	1	2	3	**4**				
C			**4**	**4**	3	2	1	2	3	**4**			
T					**4**	3	2	2	2	3	**4**		
T						**4**	3	3	2	3	**4**		
G							**4**	3	3	3	**4**		
T								**4**	3	**4**	**4**		
A									**4**	3	**4**		
C										**4**	3	**4**	
T											**4**	**4**	**4**

Matrice d'édition

Col. j

Line i

substitution(line i, col j)	insertion(line i)
deletion(col j)	edit(:$_T$O$_i$, :$_T$,O$_j$)



Propagation de contraintes

Calcul local de la distance d'édition au moyen d'un degré

Deux phases

1- Propagation :

– Toutes les modifications sont propagées de proche en proche sur les cellules adjacentes à l'**état courant**

2- Progression :

– Lorsqu'il n'y a plus de propagation, progression sur les cellules adjacentes à l'**état courant**

Différentes données séquentielles

- **Séquences simples**
  - Biologie moléculaires → Distance d'édition
  - Editions non standards (déplacements) Alignement par fragments
- **Séquences multiples**
  - Musique (mélodie, rythme, …) → MEVM

  Modèle d'Édition Multi-Valué [Rolland & Ganascia]
- **Arbres**
  - Cas général: trop complexe
  - Cas particulier: « **SOT** » (**S**tratified **O**rdered **T**ree)
- **Données numérique**
  - Descripteurs qualitatifs