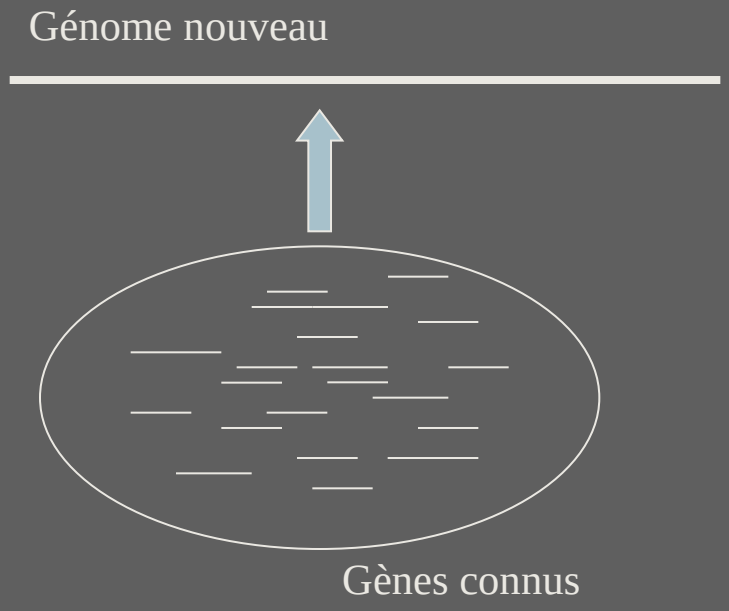
Comparaison de séquences 2 à 2

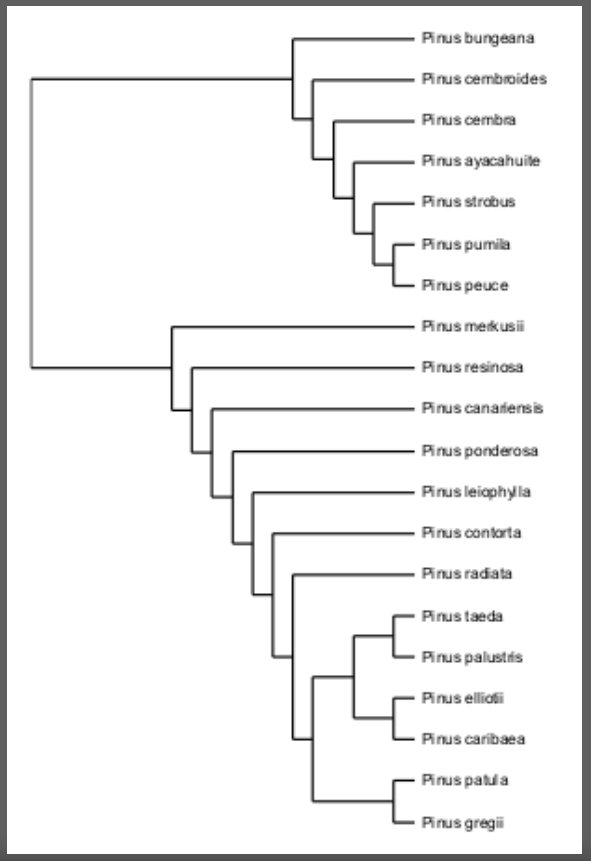
Alain Denise
Bioinformatique
LRI Orsay
Université Paris-Sud 11

Motivations

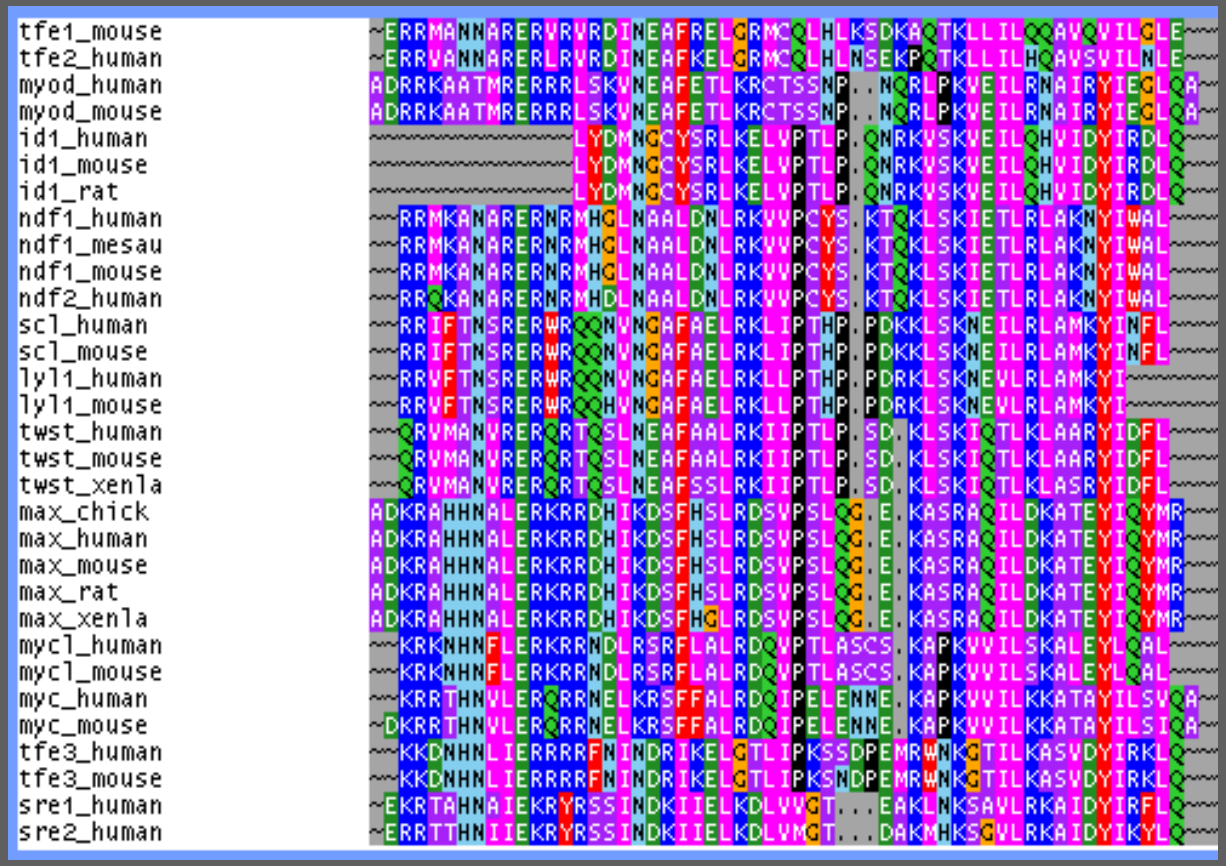
- Détection de gènes :



- Phylogénie :



(Alignements multiples)



Distance d'édition

Deux séquences $v = v_1 v_2 \ldots v_n$ et $w = w_1 w_2 \ldots w_m$

Opérations d'édition :

- ins(x,i)
- suppr(x,i)
- subs(x,y,i)

CHAT - suppr(C,1) → HAT - subs(H,R,1) → RAT

Distance d'édition

• Chaque modification a un poids, dépendant de l'opération et des lettres en cause.

• Distance d'édition entre v et w : poids minimal d'une suite d'opérations permettant de transformer v en w.

CHAT - suppr(C,1) → HAT - subs(H,R,1) → RAT

Alignement

X : alphabet des séquences
X' = X ∪ {-}
X'' = X' X' \ {-,-}

2 morphismes de monoïdes :

F_1: X''* → X*
(x,y) → x si x ≠ -
ε sinon

F_2: X''* → X*
(x,y) → y si x ≠ -
ε sinon

Un alignement de $v = v_1v_2\ldots v_n$ et $w = w_1w_2\ldots w_m$ est un mot A de X'' tel que $F_1(A) = v$ et $F2(A) = w$.

v = aatca w = agca

aa-tca
agc-a- OK

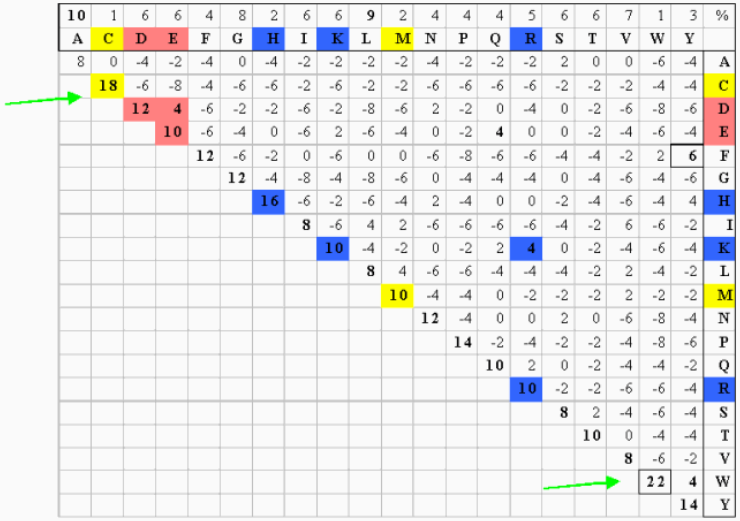
aatc-a
-agc-a ~~OK~~

aatca
-agca OK

Score d'alignement

- Le score d'un alignement est la somme des scores lettre à lettre.
- On cherche l'alignement de score maximal.

Scoring matrix based on large set of distantly related blocks: Blosum62

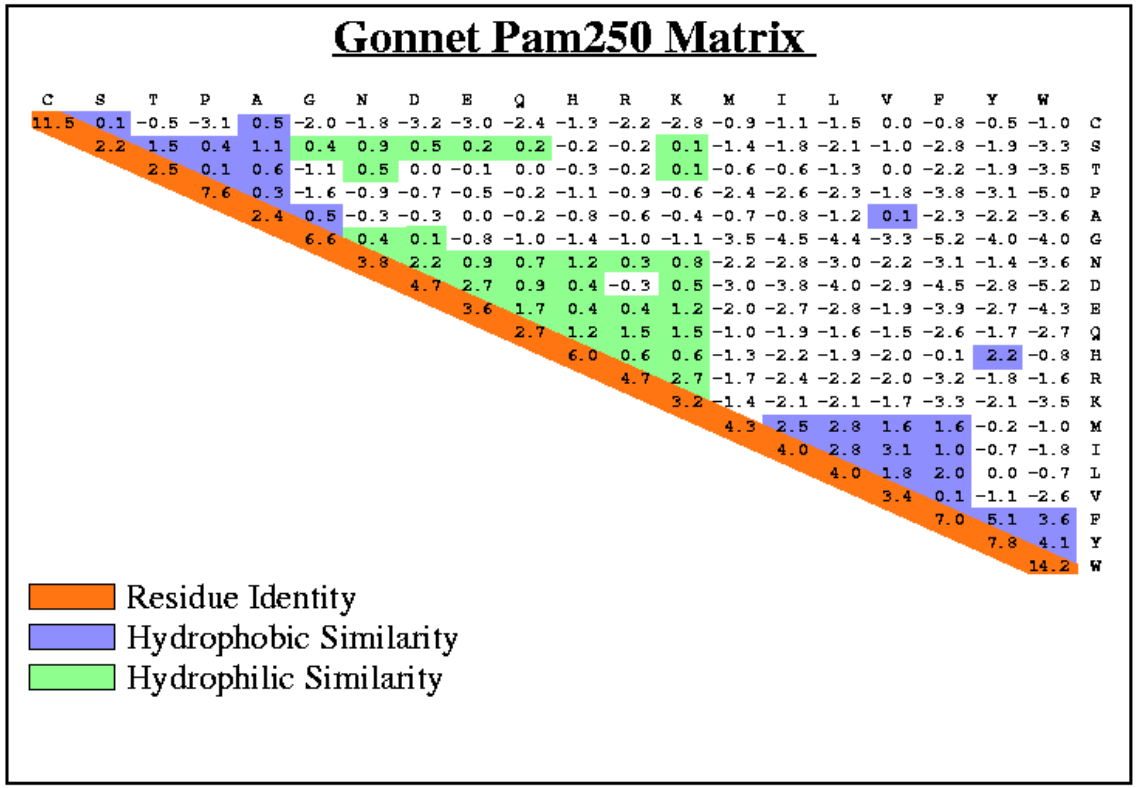


10	1	6	6	4	8	2	6	6	9	2	4	4	4	5	6	6	7	1	3	%
A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y	
8	0	-4	-2	-4	0	-4	-2	-2	-2	-2	-4	-2	-2	-2	2	0	0	-6	-4	A
	18	-6	-8	-4	-6	-6	-2	-6	-2	-2	-6	-6	-6	-6	-2	-2	-2	-4	-4	C
		12	4	-6	-2	-2	-6	-2	-8	-6	2	-2	0	-4	0	-2	-6	-8	-6	D
			10	-6	-4	0	-6	2	-6	-4	0	-2	4	0	0	-2	-4	-6	-4	E
				12	-6	-2	0	-6	0	0	-6	-8	-6	-6	-4	-4	-2	2	6	F
					12	-4	-8	-4	-8	-6	0	-4	-4	-4	0	-4	-6	-4	-6	G
						16	-6	-2	-6	-4	2	-4	0	0	-2	-4	-6	-4	4	H
							8	-6	4	2	-6	-6	-6	-6	-4	-2	6	-6	-2	I
								10	-4	-2	0	-2	2	4	0	-2	-4	-6	-4	K
									8	4	-6	-6	-4	-4	-4	-2	2	-4	-2	L
										10	-4	-4	0	-2	-2	-2	2	-2	-2	M
											12	-4	0	0	2	0	-6	-8	-4	N
												14	-2	-4	-2	-2	-4	-8	-6	P
													10	2	0	-2	-4	-4	-2	Q
														10	-2	-2	-6	-6	-4	R
															8	2	-4	-6	-4	S
																10	0	-4	-4	T
																	8	-6	-2	V
																		22	4	W
																			14	Y

Score d'alignement

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

	A	C	G	T
A	3	0	1	0
C	0	3	0	1
G	1	0	3	0
T	0	0	1	3



Un alignement de deux protéines

```
seq 1:       3 EHYYSEKPSVKSNKQTWSFRLRNKDFTFTSDSGVFSKKEVDFGSRLLIDSFEEPEVEGGI 62
                ||||  |.|     .     |   | | . |||||    | |. |||..      .  |
seq 2:       2 GHYYSREPNVPLKTKQIDVCIRGYCFKFITASGVFSFGKLDRGTELLIENM-ILKPDWKI 60

seq 1:      63 LDVGCGYGPIGLSLASDFKDRTIHMIDVNERAVELSNENAEQNGITNVKIYQSDLFSNVD 122
               || ||||| ||   || |    . | |.| |||     |.  ||. |. .    |.  |.
seq 2:      61 LDLGCGYGVIGI-VASRFVNYVV-MTDINKRAVQIARKNIKINGVKNAEVRLGNLYEPVE 118

seq 1:     123 SAQTFASILTNPPIRAGKKVVHAIFEKSAEHLKASGELWIVIQKKQGAPSAIEKLEELFD 182
                .  | || ||||..||| . . |       |.  | |  ||  | ||  . .   . |
seq 2:     119 -GEKFHSIITNPPVHAGKDILREIVINAPNYLHDGGMLQLVIKTKLGAKFIKDLMKDTFT 177

seq 1:     183 EVSVVQKKKGY 193
               ||    |  ||
seq 2:     178 EVVELAKGSGY 188
```

Equivalence édition - alignement

```
AACTA-CGAT
A-GTACCGTT
```

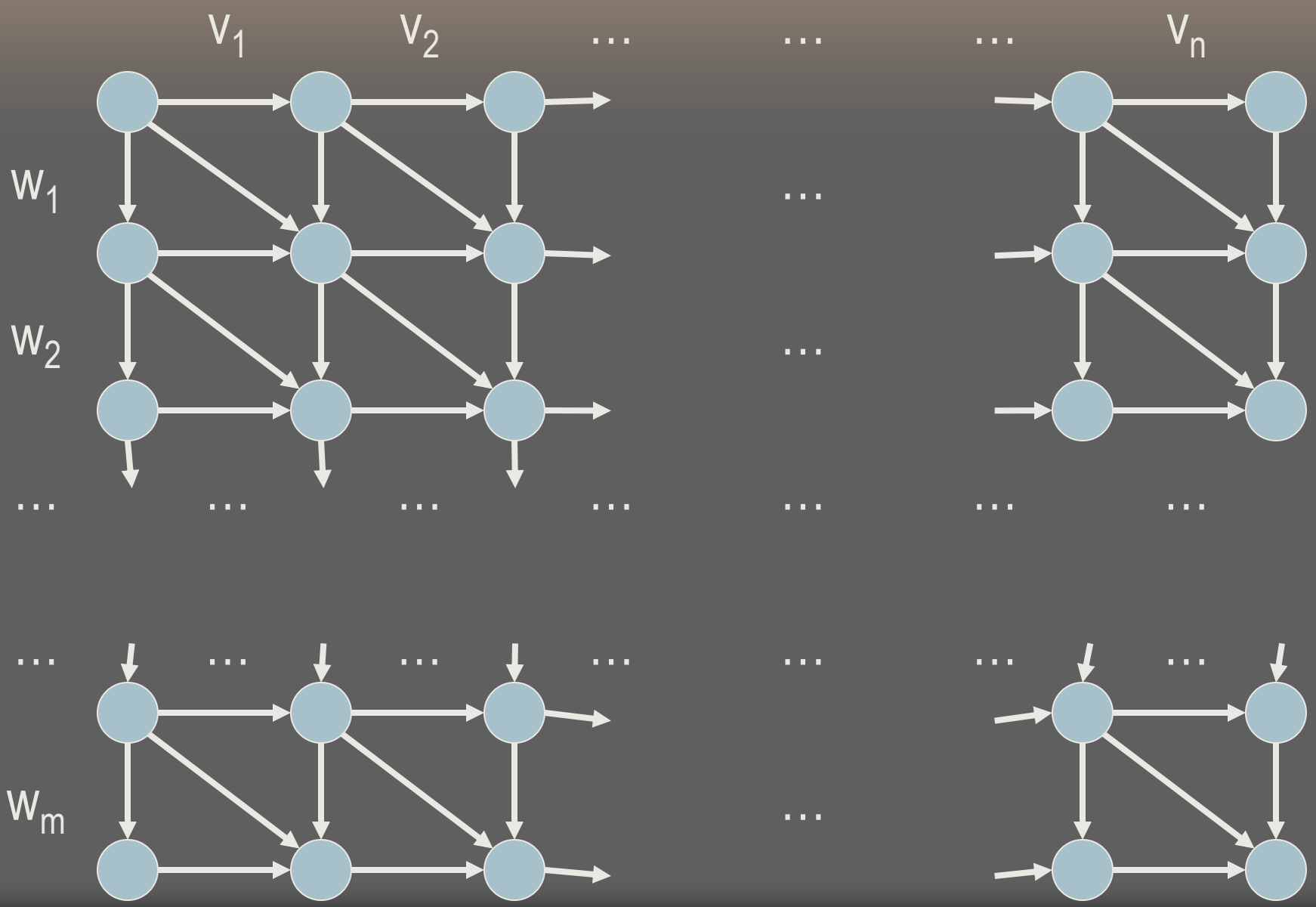
suppr(A,2)
subs(C,G,2)
ins(C,5)
subs(A,T,8)

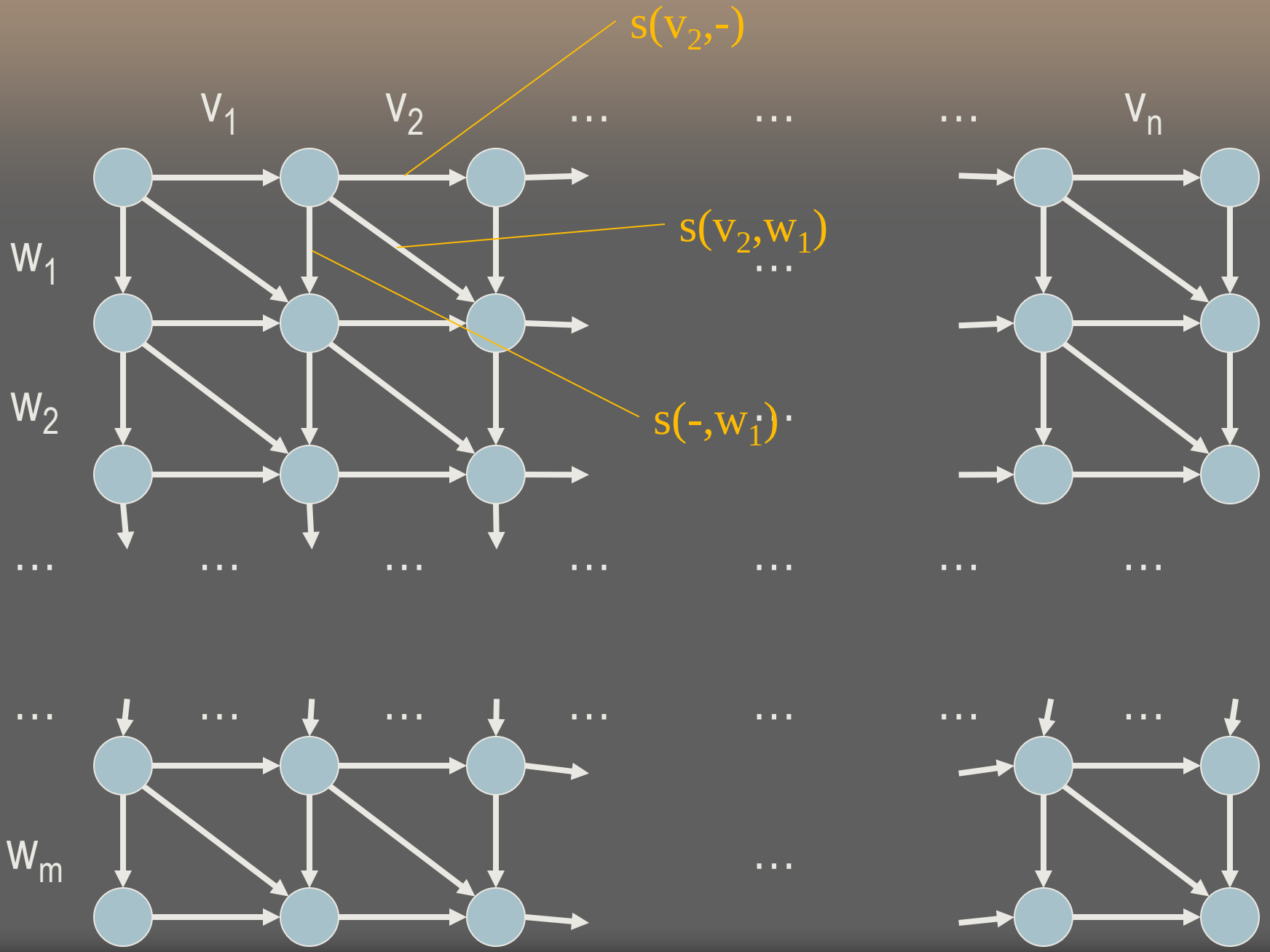
Algorithme de Needleman et Wunsch (1970), Gotoh (1982)

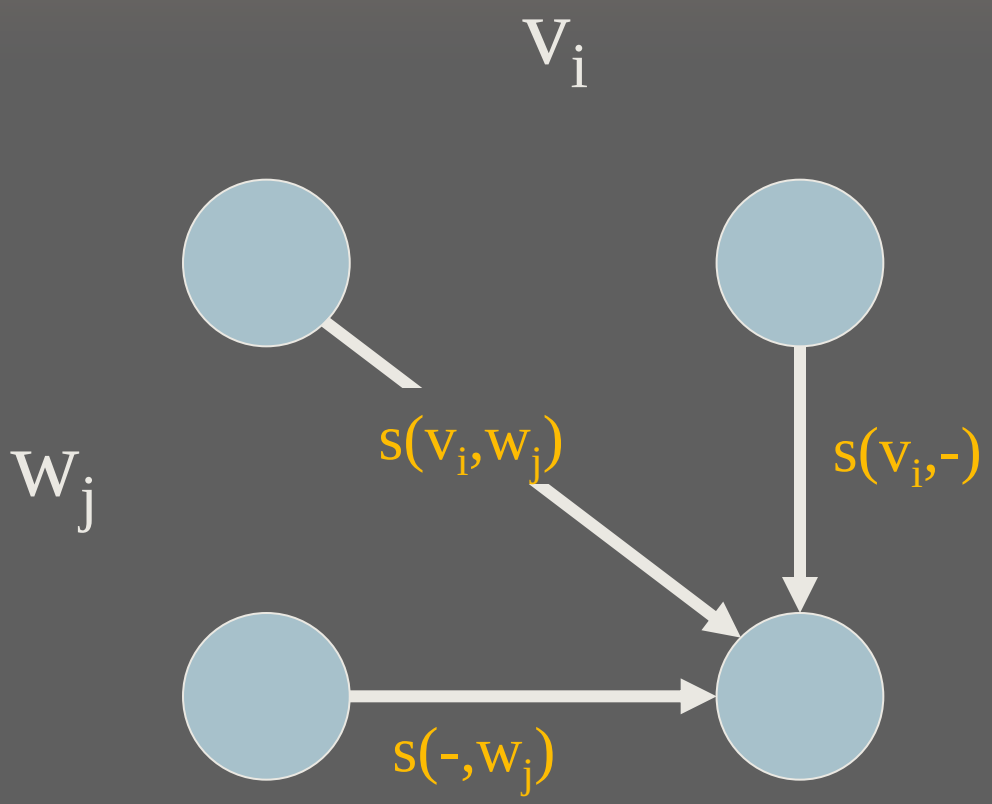
$v = v_1 v_2 \ldots v_n$
$w = w_1 w_2 \ldots w_m$

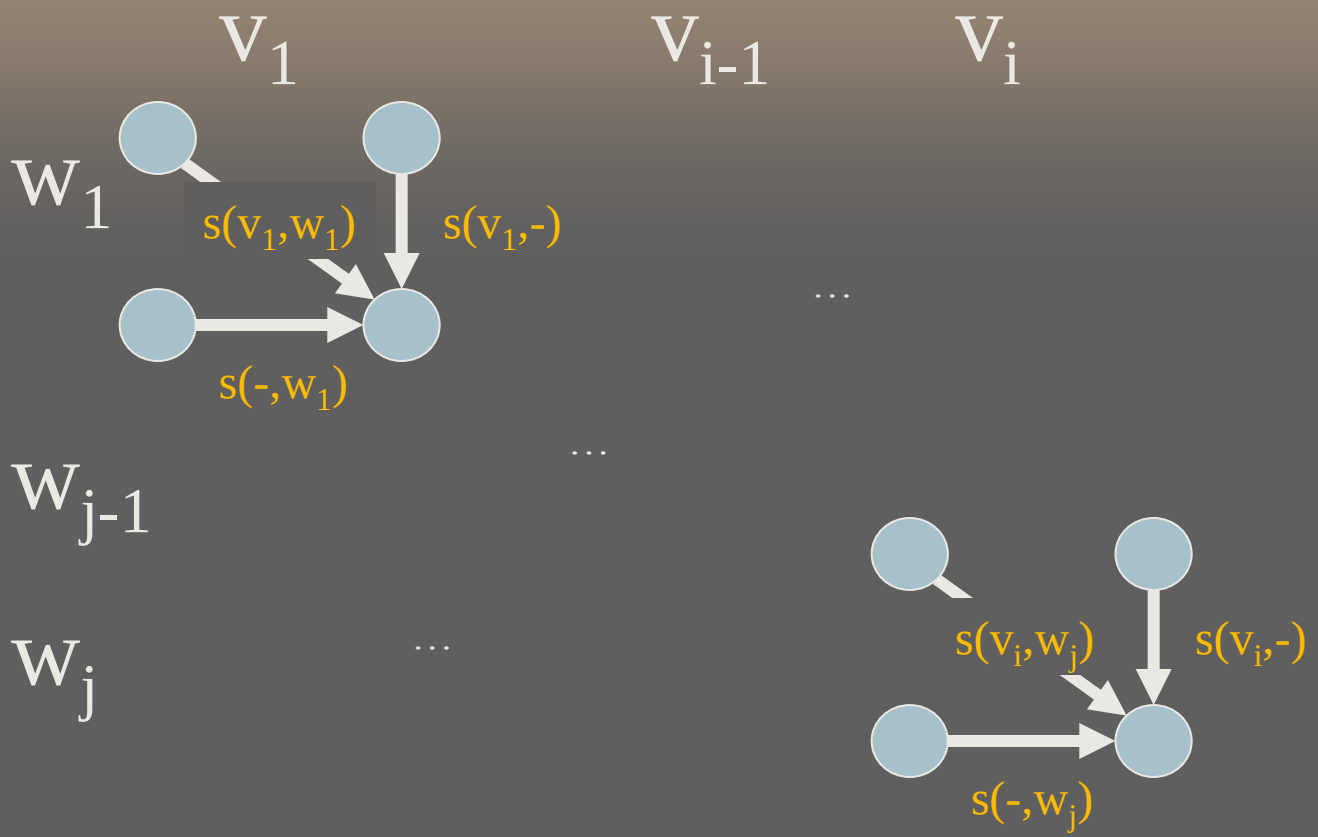
s(x,y) : score de substitution de x en y
s(x,-) : score de suppression de x
s(-,y) : score d'insertion de y

S(v,w) : score d'alignement de v et w


v_1
v_2
…
…
…
v_n
w_1
w_2
w_m


s(v2,-)
v1
v2
…
…
…
vn
s(v2,w1)
w1
…
w2
s(-,w1)
…
…
…
…
…
…
…
…
…
…
…
…
…
…
…
wm
…


v_i
w_j
$s(v_i,w_j)$
$s(v_i,-)$
$s(-,w_j)$



$$S(v_1 \ldots v_i, w_1 \ldots w_j) = \text{Max} \{$$

$$S(v_1 \ldots v_{i-1}, w_1 \ldots w_{j-1}) + s(v_i, w_j)$$
$$S(v_1 \ldots v_{i-1}, w_1 \ldots w_j) + s(v_i, -)$$
$$S(v_1 \ldots v_i, w_1 \ldots w_{j-1}) + s(-, w_j)$$

$$\}$$

	W	I	N	D	O	X	S	
	0	-1	-2	-3	-4	-5	-6	-7
L	-1							
I	-2							
N	-3							
U	-4							
X	-5							

s(x,y) = 2 si x=y
-1 sinon

s(x,-) = s(-,x) = -1

LINUX \ WINDOWS	0	-1	-2	-3	-4	-5	-6	-7
-1		-1						
-2								
-3								
-4								
-5								

s(x,y) = 2 si x=y
-1 sinon

s(x,-) = s(-,x) = -1

LINUX \ WINDOWS	0	-1	-2	-3	-4	-5	-6	-7
-1	-1	-2						
-2								
-3								
-4								
-5								

s(x,y) = 2 si x=y
-1 sinon

s(x,-) = s(-,x) = -1

	W	I	N	D	O	X	S	
L	0	-1	-2	-3	-4	-5	-6	-7
I	-1	-1	-2	-3	-4	-5	-6	-7
N	-2	-2						
U	-3							
X	-4							
	-5							

s(x,y) = 2 si x=y
-1 sinon

s(x,-) = s(-,x) = -1

	W	I	N	D	O	X	S
L	0	-1	-2	-3	-4	-5	-6	-7
I	-1	-1	-2	-3	-4	-5	-6	-7
N	-2	-2	1					
U	-3							
X	-4							
	-5							

s(x,y) = 2 si x=y
-1 sinon

s(x,-) = s(-,x) = -1

LINUX \ WINDOWS	W	I	N	D	O	X	S
L	0	-1	-2	-3	-4	-5	-6	-7
I	-1	-1	-2	-3	-4	-5	-6	-7
N	-2	-2	1 →	0				
U	-3							
X	-4							
	-5							

s(x,y) = 2 si x=y
-1 sinon

s(x,-) = s(-,x) = -1

	W	I	N	D	O	X	S
L	0	-1	-2	-3	-4	-5	-6	-7
I	-1	-1	-2	-3	-4	-5	-6	-7
N	-2	-2	1	0	-1	-2	-3	-4
U	-3	-3	0	3	2	1	0	-1
X	-4	-4	-1	2	2	1	0	-1
	-5	-5	-2	1	1	1	3	2

s(x,y) = 2 si x=y
-1 sinon

s(x,-) = s(-,x) = -1

	W	I	N	D	O	X	S	
L	0	-1	-2	-3	-4	-5	-6	-7
I	-1	-1	-2	-3	-4	-5	-6	-7
N	-2	-2	1	0	-1	-2	-3	-4
U	-3	-3	0	3	2	1	0	-1
X	-4	-4	-1	2	2	1	0	-1
	-5	-5	-2	1	1	1	3 →	2

s(x,y) = 2 si x=y
-1 sinon

s(x,-) = s(-,x) = -1

	W	I	N	D	O	X	S
L	0	-1	-2	-3	-4	-5	-6	-7
I	-1	-1	-2	-3	-4	-5	-6	-7
N	-2	-2	1	0	-1	-2	-3	-4
U	-3	-3	0	3	2	1	0	-1
X	-4	-4	-1	2	2	1	0	-1
	-5	-5	-2	1	1	1	3 →	2

s(x,y) = 2 si x=y
-1 sinon

s(x,-) = s(-,x) = -1

	W	I	N	D	O	X	S	
L	0	-1	-2	-3	-4	-5	-6	-7
I	-1	-1	-2	-3	-4	-5	-6	-7
N	-2	-2	1	0	-1	-2	-3	-4
U	-3	-3	0	3	2	1	0	-1
X	-4	-4	-1	2	2	1	0	-1
	-5	-5	-2	1	1	1	3	2

$s(x,y) = 2$ si $x=y$
-1 sinon

$s(x,-) = s(-,x) = -1$

	W	I	N	D	O	X	S
L	0	-1	-2	-3	-4	-5	-6	-7
I	-1	-1	-2	-3	-4	-5	-6	-7
N	-2	-2	1	0	-1	-2	-3	-4
U	-3	-3	0	3	2	1	0	-1
X	-4	-4	-1	2	2	1	0	-1
	-5	-5	-2	1	1	1	3	2

s(x,y) = 2 si x=y
-1 sinon

s(x,-) = s(-,x) = -1

	W	I	N	D	O	X	S
L	0	-1	-2	-3	-4	-5	-6	-7
I	-1	-1	-2	-3	-4	-5	-6	-7
N	-2	-2	1	0	-1	-2	-3	-4
U	-3	-3	0	3	2	1	0	-1
X	-4	-4	-1	2	2	1	0	-1
	-5	-5	-2	1	1	1	3	2

s(x,y) = 2 si x=y
-1 sinon

s(x,-) = s(-,x) = -1

	W	I	N	D	O	X	S
L	0	-1	-2	-3	-4	-5	-6	-7
I	-1	-1	-2	-3	-4	-5	-6	-7
N	-2	-2	1	0	-1	-2	-3	-4
U	-3	-3	0	3	2	1	0	-1
X	-4	-4	-1	2	2	1	0	-1
	-5	-5	-2	1	1	1	3	2

s(x,y) = 2 si x=y
-1 sinon

s(x,-) = s(-,x) = -1

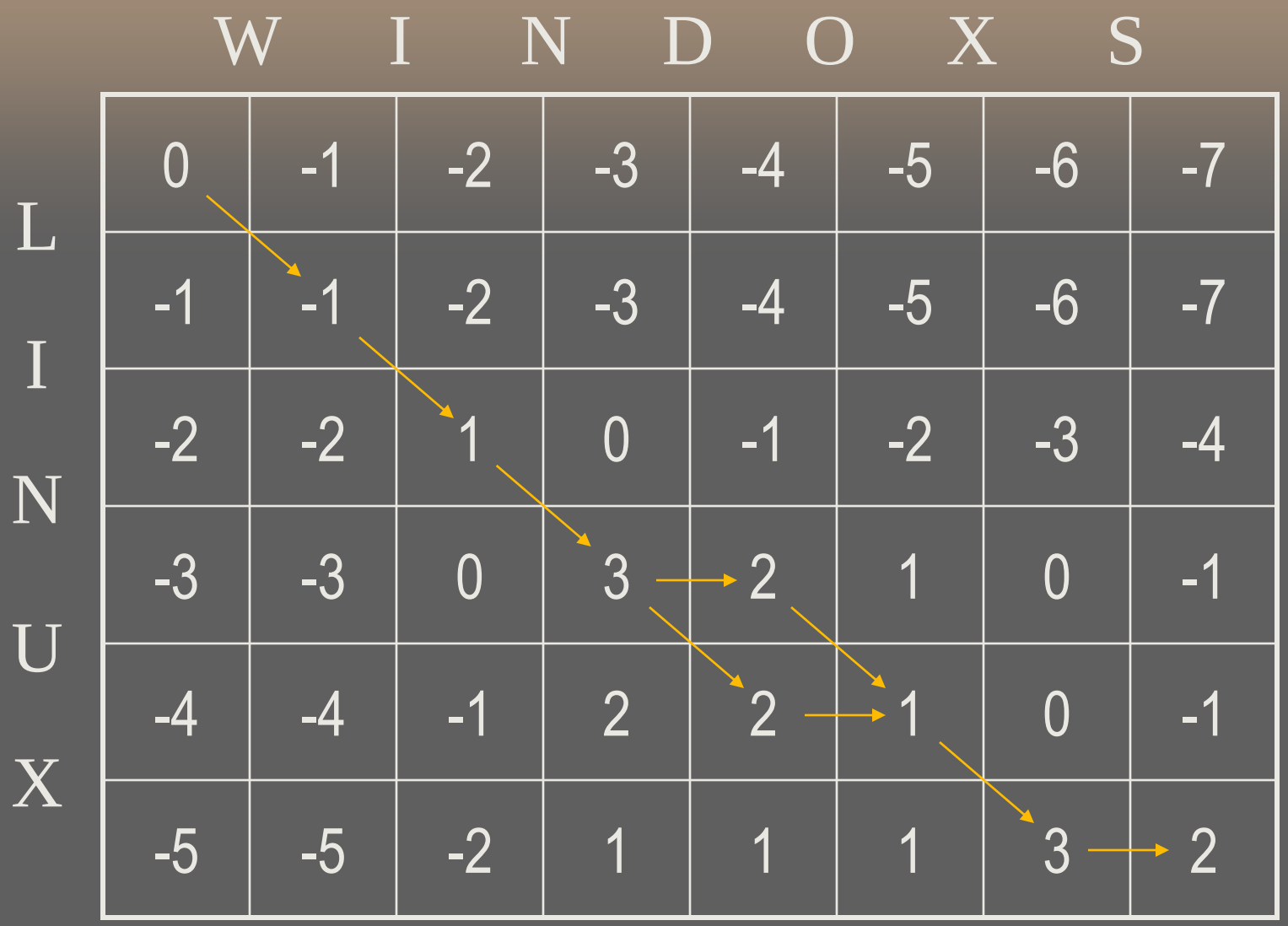
L \ W	I	N	D	O	X	S	
0	-1	-2	-3	-4	-5	-6	-7
-1	-1	-2	-3	-4	-5	-6	-7
-2	-2	1	0	-1	-2	-3	-4
-3	-3	0	3	2	1	0	-1
-4	-4	-1	2	2	1	0	-1
-5	-5	-2	1	1	1	3	2

s(x,y) = 2 si x=y
-1 sinon

s(x,-) = s(-,x) = -1

	W	I	N	D	O	X	S
L	0	-1	-2	-3	-4	-5	-6	-7
I	-1	-1	-2	-3	-4	-5	-6	-7
N	-2	-2	1	0	-1	-2	-3	-4
U	-3	-3	0	3	2	1	0	-1
X	-4	-4	-1	2	2	1	0	-1
	-5	-5	-2	1	1	1	3	2

W I N D O X S
L I N – U X –



W I N D O X S
L I N – U X –

W I N D O X S
L I N U – X –

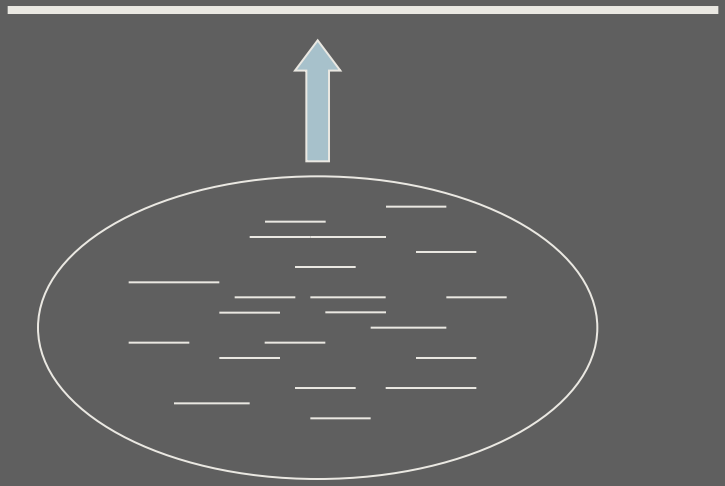
Complexité

En espace : m n

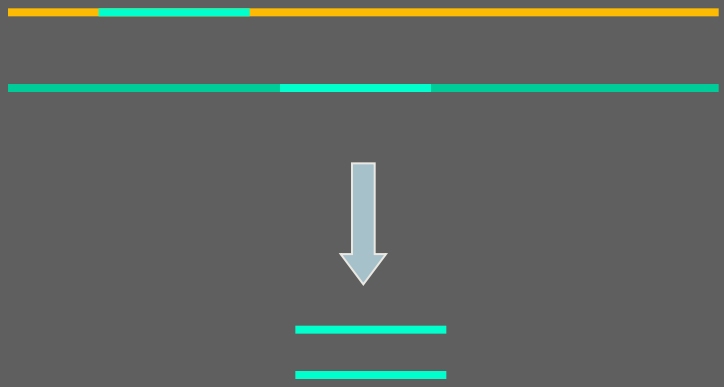
En temps : m n

Alignement local (Smith et Waterman 1981)

Aligner des petites séquences sur des grandes :



Trouver des similarités locales :

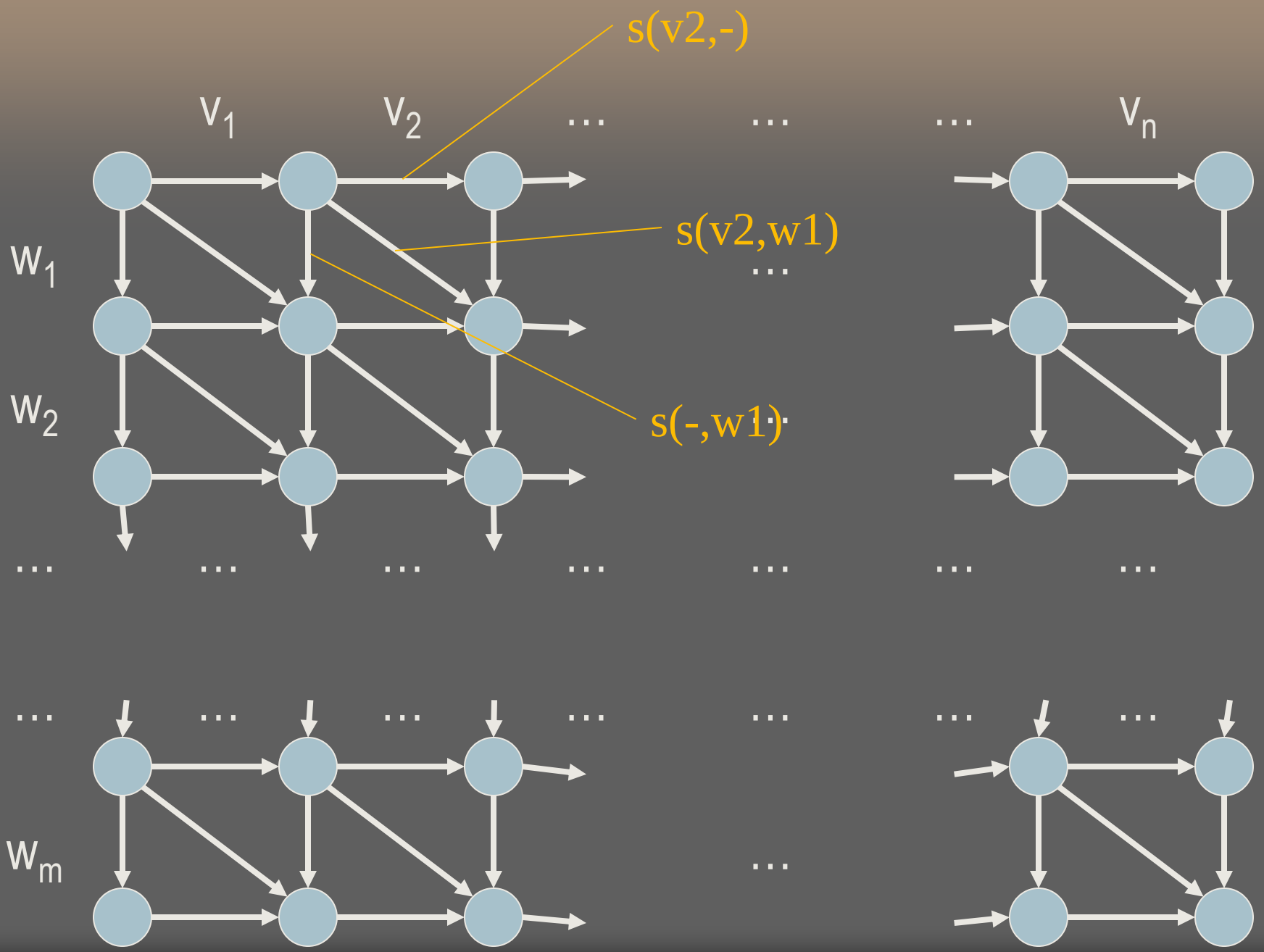


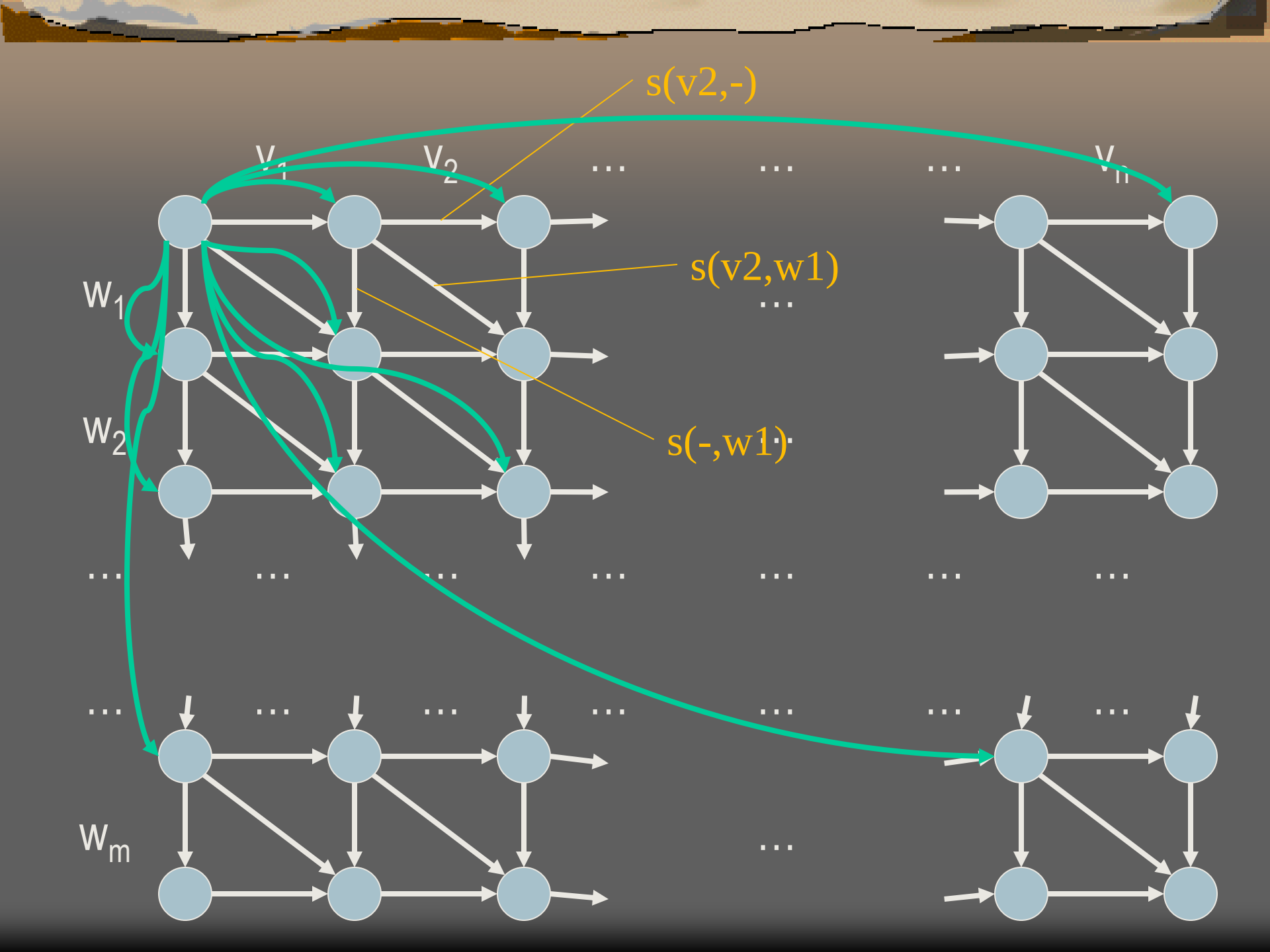
Alignement local de deux séquences d'ADN

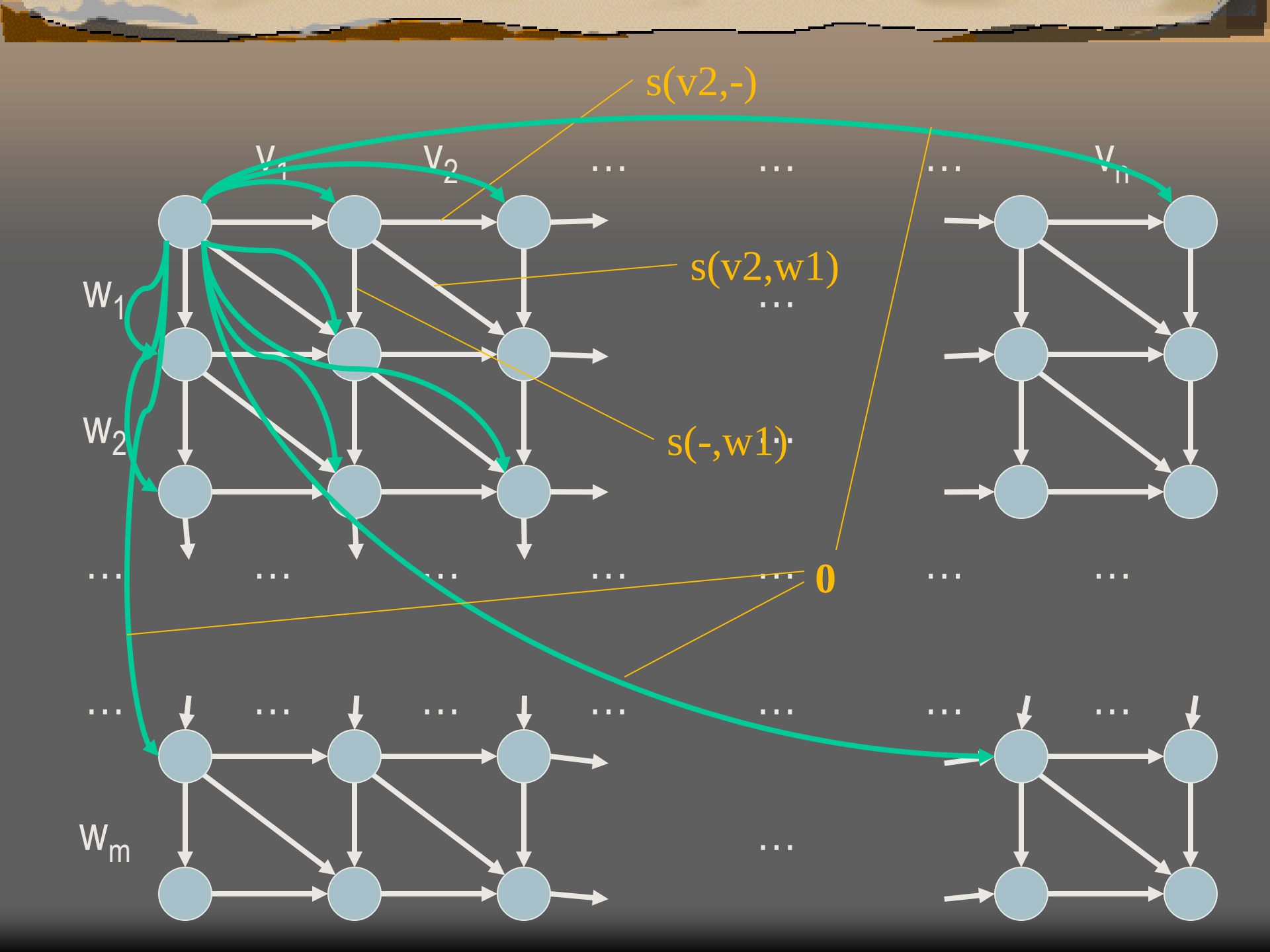
ACCAGTTAAAGCTAGTGATTTATTTTATGATCCCCAACCACTTCTTAATTCTATCAGTTT
TTATAGTATATAACAACGTAAATCTGTAACGACGTCAGTTGTTCCAATATCTATTTTCTT
ATTTGTGGAAAAACATATCTACCTTACTTAATAAGTTGCAGGTATGTTTATTAAAACACA
AGCAAAGTAGCTATATATCTCATTTTATCCTGCAAATTGTAAGATCGAAAATGATACTAT
AAAGGATGAGAACACCGACTTAATAATACGTTCATATTCTAGTTACAATCAATCTAACTC
TAAGCGTTGATTTTAACTTAAGACTAGCGGTGTTATAATTTTAATATGTATGAATTGCAA

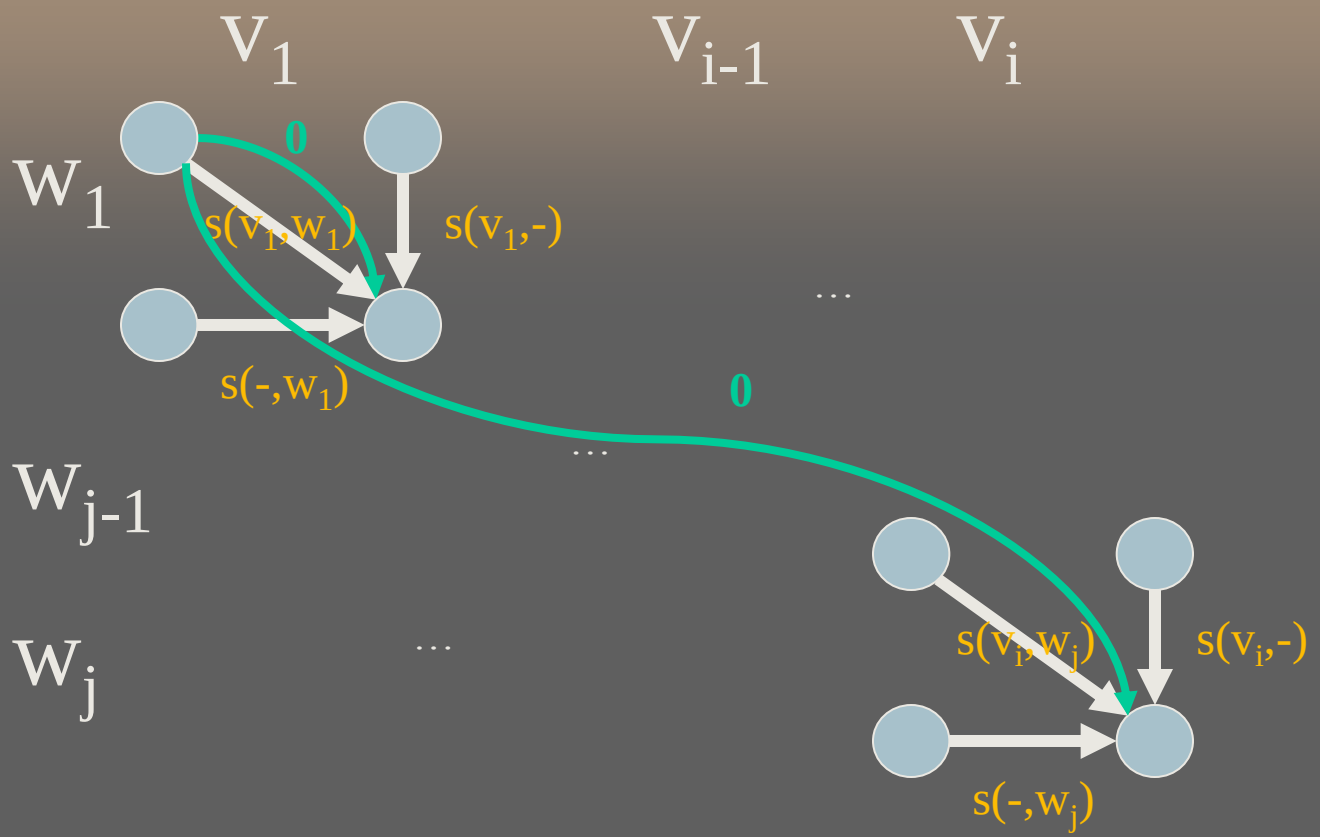
TTCTCTGCCAATCCCCTATACTATACATGGCTAGCTCAATAGGGATGAGATGGGAGTGAT
GGTAACTAGTAATGTAACGTTTTATATAATTACCTAAAAAAATACTTATGCTAAGTTCGT
AGCACCCTCTTAGGTATAAAATTTATTCCAACGTGGCATATAAATGTATTAATATTTAAC
TAAAGTATCTAAAAATCTCAATTTATAATATAACAGTTATTGTTGAAATCAGGCGGAATG
TTTATCACCATGAATATATTTTTAGTCGAAACATGTGATTAGAATAATTAACGTAAATAT

```
seq 1:     153 AAGTTGCAGGTATGTTTATTAAAACACAAGCAAAGTAGCTATATATCTCATTTTAT 208
               | || |||  ||  | |||||| |   ||  ||||||| ||| | ||||||| |||||
seq 2:     151 ACGTGGCATATAAATGTATTAATATTTAACTAAAGTATCTAAAAATCTCAATTTAT 206
```


s(v2,-)
s(v2,w1)
s(-,w1)
v1
v2
vn
w1
w2
wm


s(v2,-)
v_1
v_2
v_n
s(v2,w1)
w_1
w_2
s(-,w1)
w_m


s(v2,-)
v_1
v_2
…
…
…
v_n
s(v2,w1)
w_1
…
w_2
s(-,w1)
…
…
…
…
…
…
…
…
0
…
…
…
…
…
…
…
…
w_m
…



$$S(v_1 \ldots v_i, w_1 \ldots w_j) = \text{Max} \{$$

$$0$$

$$S(v_1 \ldots v_{i-1}, w_1 \ldots w_{j-1}) + s(v_i, w_j)$$

$$S(v_1 \ldots v_{i-1}, w_1 \ldots w_j) + s(v_i, -)$$

$$S(v_1 \ldots v_i, w_1 \ldots w_{j-1}) + s(-, w_j)$$

$$\}$$

	W	I	N	D	O	X	S	
	0	0	0	0	0	0	0	0
L	0							
I	0							
N	0							
U	0							
X	0							

s(x,y) = 2 si x=y
-1 sinon

s(x,-) = s(-,x) = -1

	W	I	N	D	O	X	S	
	0	0	0	0	0	0	0	0
L	0	0						
I	0							
N	0							
U	0							
X	0							

s(x,y) = 2 si x=y
-1 sinon

s(x,-) = s(-,x) = -1

	W	I	N	D	O	X	S	
L	0	0	0	0	0	0	0	0
I	0	0	0	0	0	0	0	0
N	0	0	2					
U	0							
X	0							
	0							

s(x,y) = 2 si x=y
-1 sinon

s(x,-) = s(-,x) = -1

	W	I	N	D	O	X	S	
L	0	0	0	0	0	0	0	0
I	0	0	0	0	0	0	0	0
N	0	0	2	1	0	0	0	0
U	0	0	1	4	3	2	1	0
X	0	0	0	3	3	2	1	0
	0	0	0	2	2	2	4	3

s(x,y) = 2 si x=y
-1 sinon

s(x,-) = s(-,x) = -1

	W	I	N	D	O	X	S	
L	0	0	0	0	0	0	0	0
I	0	0	0	0	0	0	0	0
N	0	0	2	1	0	0	0	0
U	0	0	1	**4**	3	2	1	0
X	0	0	0	3	3	2	1	0
	0	0	0	2	2	2	**4**	3

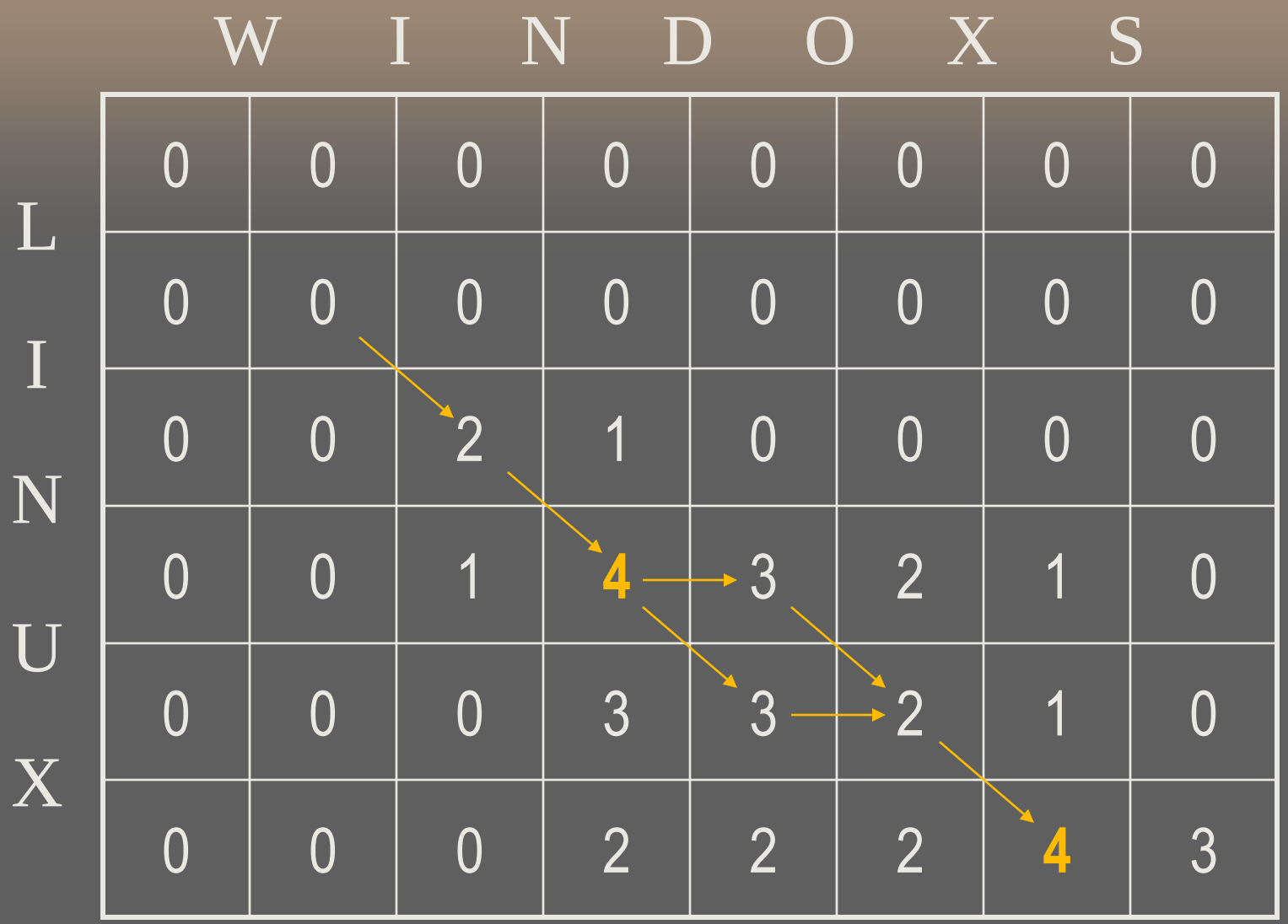
s(x,y) = 2 si x=y
-1 sinon

s(x,-) = s(-,x) = -1

	W	I	N	D	O	X	S	
L	0	0	0	0	0	0	0	0
I	0	0	0	0	0	0	0	0
N	0	0	2	1	0	0	0	0
U	0	0	1	**4**	3	2	1	0
X	0	0	0	3	3	2	1	0
	0	0	0	2	2	2	**4**	3

s(x,y) = 2 si x=y
-1 sinon

s(x,-) = s(-,x) = -1



I N
I N

I N D O X
I N – U X

I N D O X
I N U – X