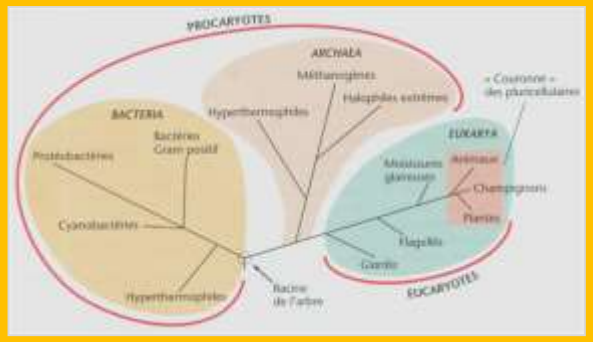
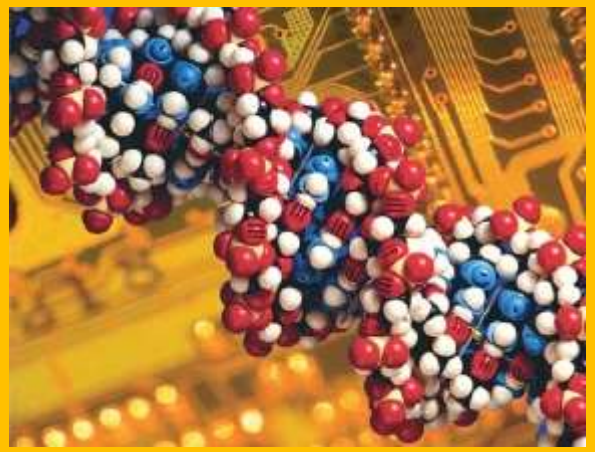
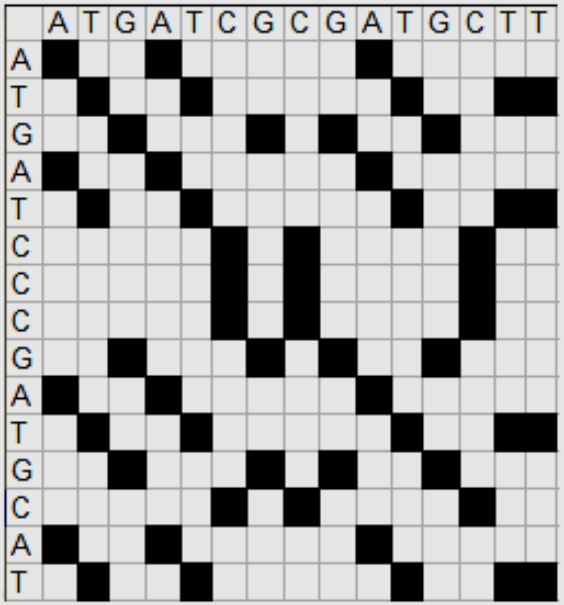
Université Frères Mentouri Constantine 1
Institut de la Nutrition, de l'Alimentation et des Technologies Agro-alimentaires (INATAA)
1e année Master Biotechnologie alimentaire
2019-2020

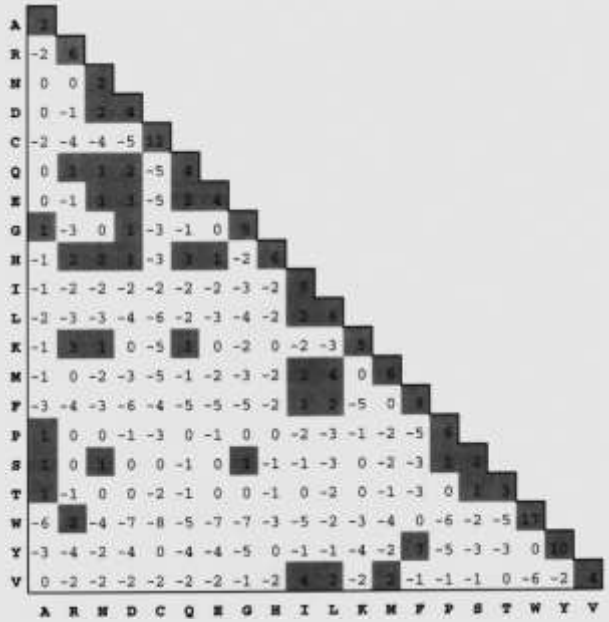


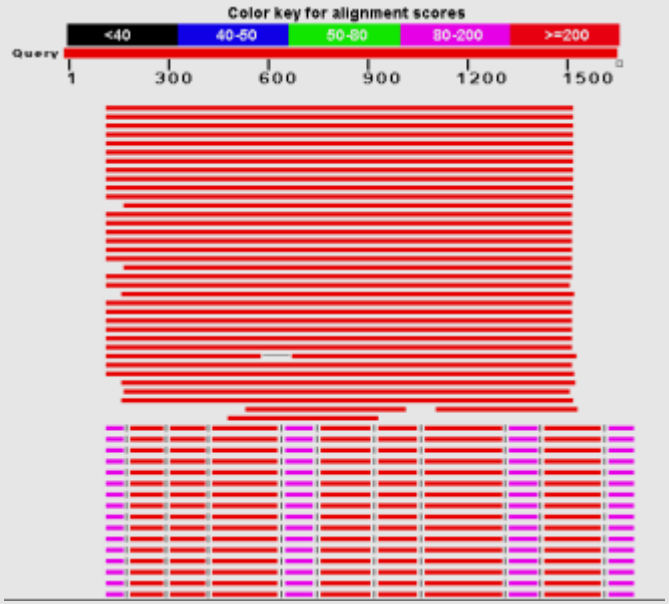
COURS DE BIOINFORMATIQUE

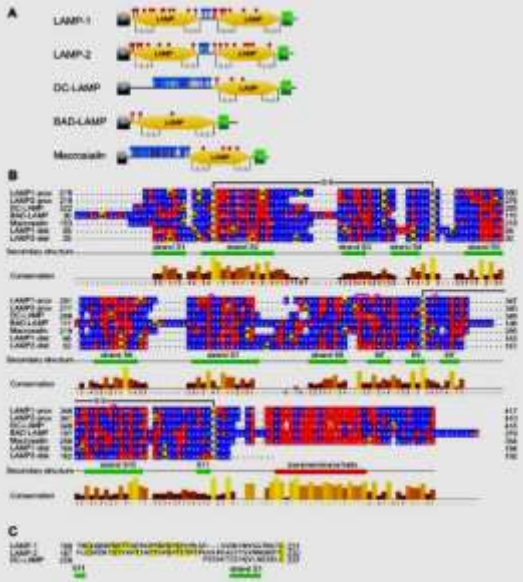


CHAPITRE II
LES ALIGNEMENTS









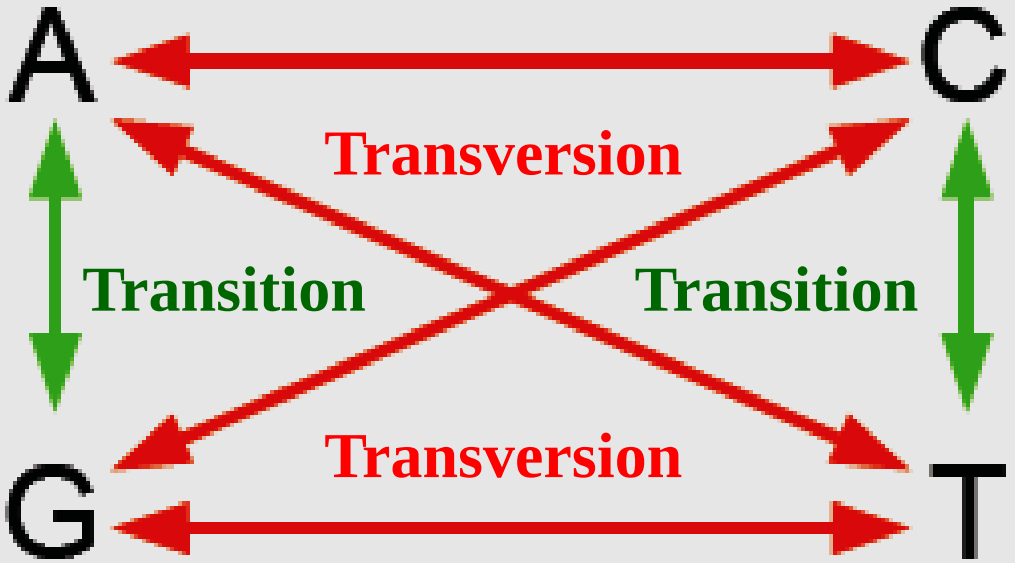
ALIGNEMENT DES SÉQUENCES: DÉFINITIONS

- L'évolution des gènes laisse une trace parfaitement visible lorsque l'on compare leurs séquences ;

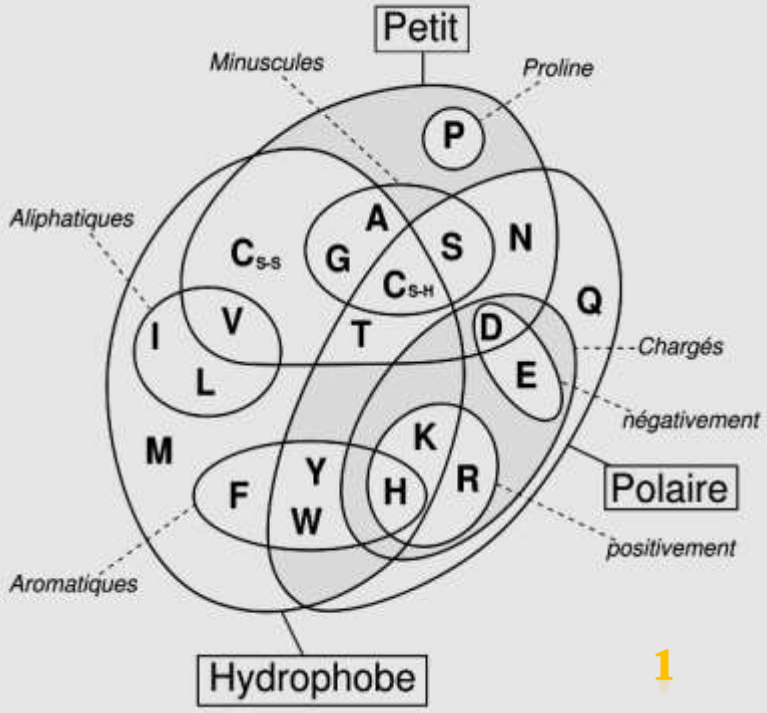
- L'évolution des gènes s'effectue essentiellement par mutations ponctuelles :

✓ les insertions-délétions ou indels (apparition ou disparition d'une nucléotide/un acide aminé) ;

✓ les substitutions (remplacement d'un résidu par un autre) :

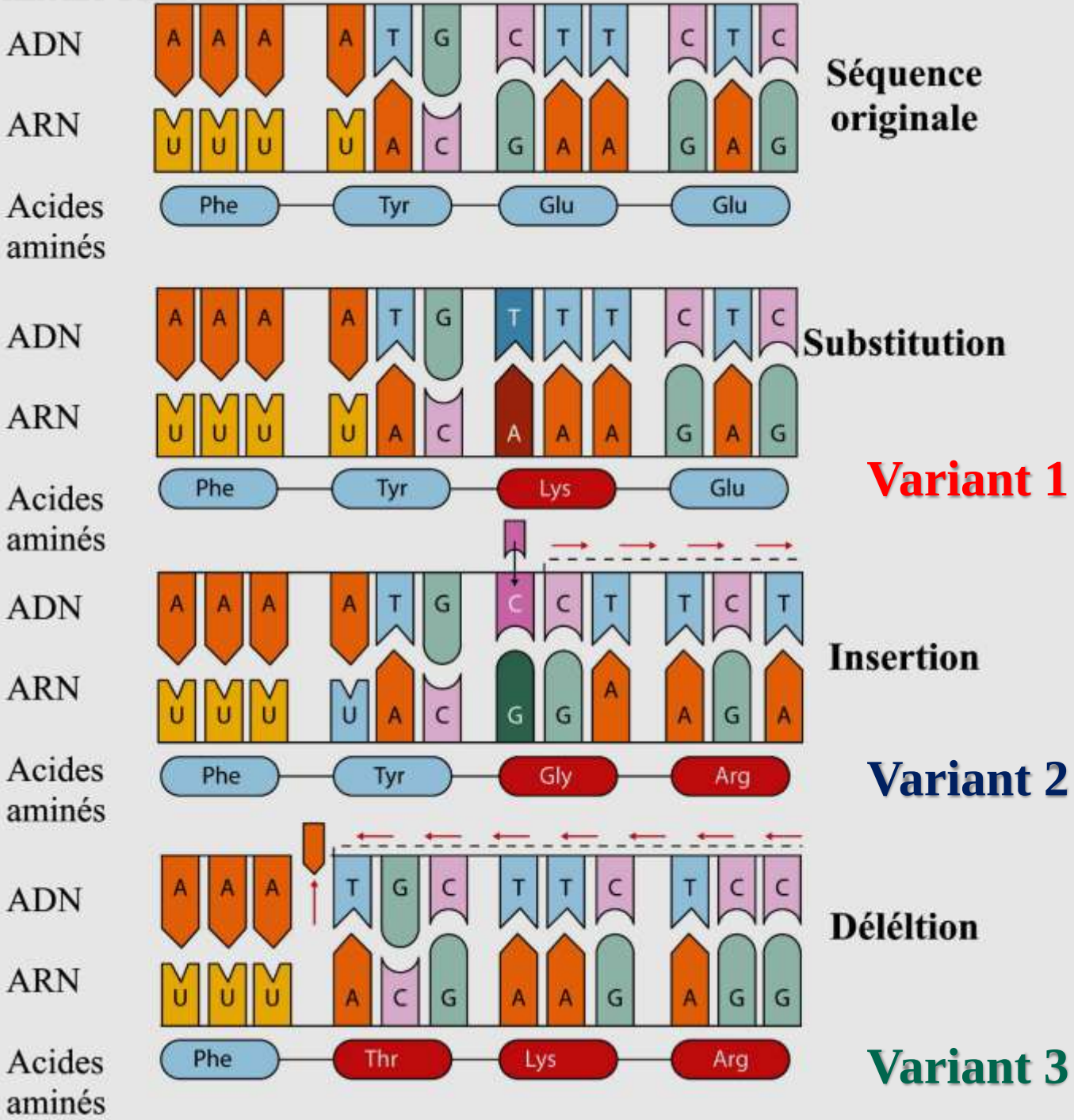


purines **pyrimidines**



ALIGNEMENT DES SÉQUENCES: DÉFINITIONS

COÛT DES MUTATIONS PONCTUELLES



ALIGNEMENT DES SÉQUENCES: DÉFINITIONS

PRINCIPE DE L'ÉVOLUTION DES GÈNES

- ✓ On peut déduire la fonction de la plupart des gènes par comparaison avec les gènes «homologues» d'autres espèces, c'est-à-dire issus d'un ancêtre commun.
- ✓ Les régions fonctionnelles des gènes (codant pour des sites catalytiques, de fixation, séquences motifs de régulation, etc.) sont soumises à sélection. Elles sont relativement préservées par l'évolution car des mutations trop radicales sont désavantageuses.
- ✓ Les régions non-fonctionnelles ne subissent aucune pression de sélection et divergent rapidement à mesure que s'accumulent les mutations.

ALIGNEMENT DES SÉQUENCES: DÉFINITIONS

PRINCIPE DE L'ALIGNEMENT

En bioinformatique, la comparaison des séquences (ADN, ARN et/ou protéines) repose dans la plupart des applications sur la notion d'**alignement**.

L'**alignement** est une opération qui vise à identifier des **zones communes** entre un groupe de *k* séquences. La présence de zones communes au niveau de ces séquences pourrait indiquer que :

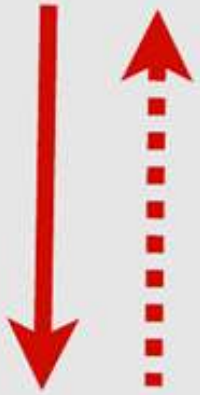
- leur structure (primaire, secondaire ou tertiaire) est semblable ;
- leurs fonctions biologiques sont proches ou différentes (dans le cas de la dissémilarité) ;
- leur origine est commune ou éloignée (notion d'homologie), etc.

ALIGNEMENT DES SÉQUENCES: DÉFINITIONS

PRINCIPE DE L'ALIGNEMENT

similarités dans la séquence

similarités dans la structure



similarités dans la fonction

ALIGNEMENT DES SÉQUENCES: DÉFINITIONS

✓ Selon la taille des séquences comparées on distingue deux types d'alignements :

Alignement local ou global ?

Finalités différentes

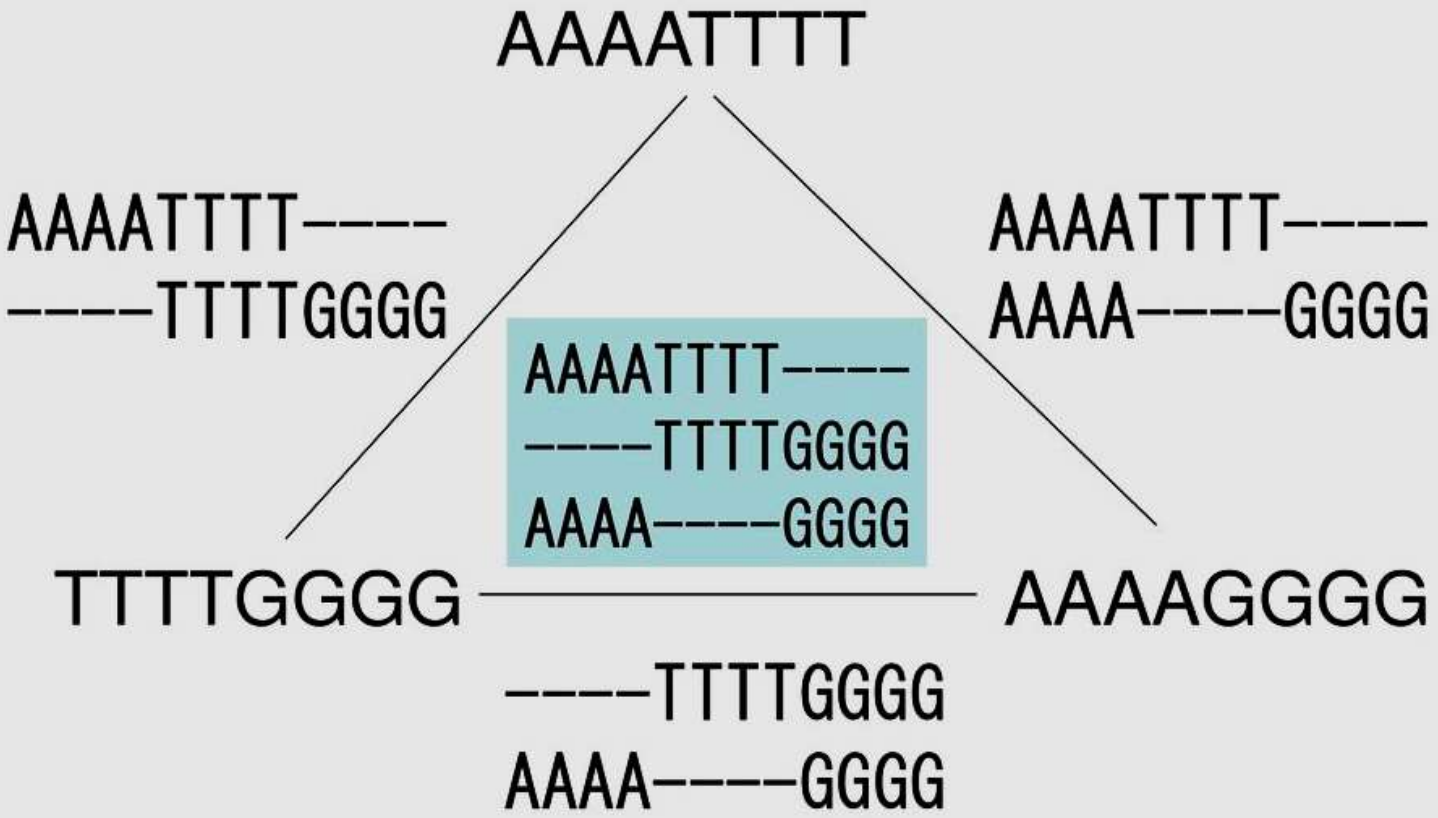
L'alignement **global** est conçu pour comparer des séquences apparentées sur toute leur longueur.

L'alignement **local** est conçu pour rechercher dans la séquence A des régions semblables à la séquence B (ou à des parties de la séquence B)

ALIGNEMENT DES SÉQUENCES: DÉFINITIONS

✓ Selon le nombre de séquences comparées on distingue deux types d'alignements :

- alignement **par paire** : on aligne 2 séquences
- alignement **multiple** : on aligne plus de 2 séquences



Les méthodes d'alignement sont retrouvées dans la plupart des applications de la bioinformatique :

- études phylogénétiques ;
- assemblage, études comparatives (structure, fonction) des génomes ;
- annotation des gènes ;
- prédiction de la structure et de la fonction des ARN ;
- prédiction de la structure 2D/3D des protéines ;
- caractérisation de la fonction des protéines ;

……

ALIGNEMENT DES SÉQUENCES: DÉFINITIONS

NOTIONS DE SIMILARITÉ, D'IDENTITÉ ET D'HOMOLOGIE

Il existe plusieurs termes permettant de nommer la ressemblance entre deux séquences biologiques:

- ✓ La **similarité** est une quantité qui se mesure en **% d'identité** ou par un **score de similarité**. L'identité elle-même peut être définie comme étant une ressemblance parfaite entre deux séquences.
- ✓ L'**homologie** est une propriété (évolutive) des séquences : deux séquences sont dites homologues si elles possèdent un ancêtre commun. L'homologie présente la particularité d'être **transitive**. Si A est homologue à B et B homologue à C, alors A est homologue à C même si A et C se ressemblent très peu.

ALIGNEMENT DES SÉQUENCES: DÉFINITIONS

NOTIONS DE SIMILARITÉ, D'IDENTITÉ ET D'HOMOLOGIE

- ✓ En pratique, l'homologie entre deux séquences est inférée avec confiance lorsque le pourcentage d'identité elles est **≥ 30 %** et que l'alignement **couvre ≥ 70 %** des deux séquences.
- ✓ En revanche, des protéines partageant moins de 30 % d'identité peuvent être homologues (dans ce cas le taux de mutation a été élevé au cours de l'évolution). L'homologie est donc une propriété intrinsèque et ne peut être qualifiée de forte ou de faible.

La **matrice de point** ou ***dot plot*** est une méthode simple de représentation **visuelle (graphique)** des positions de similarités entre deux séquences (ou sur la même séquence).

Objectifs

• Dans le cas de la comparaison d'une séquence avec elle-même, il s'agit de détecter les répétitions internes.

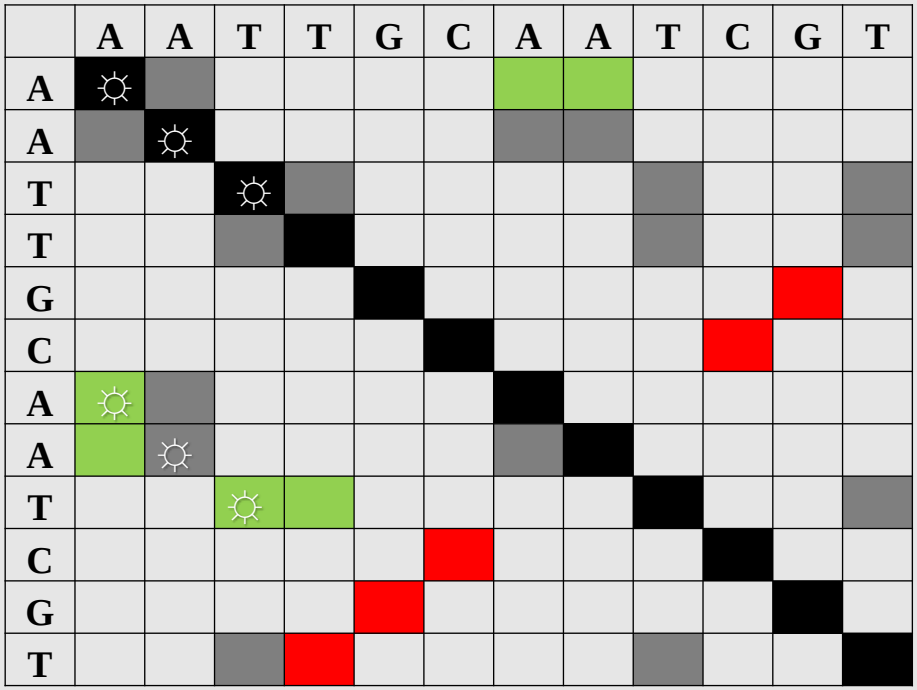
• Dans le cas où la comparaison implique deux séquences différentes, il est possible d'identifier des régions de similarité.

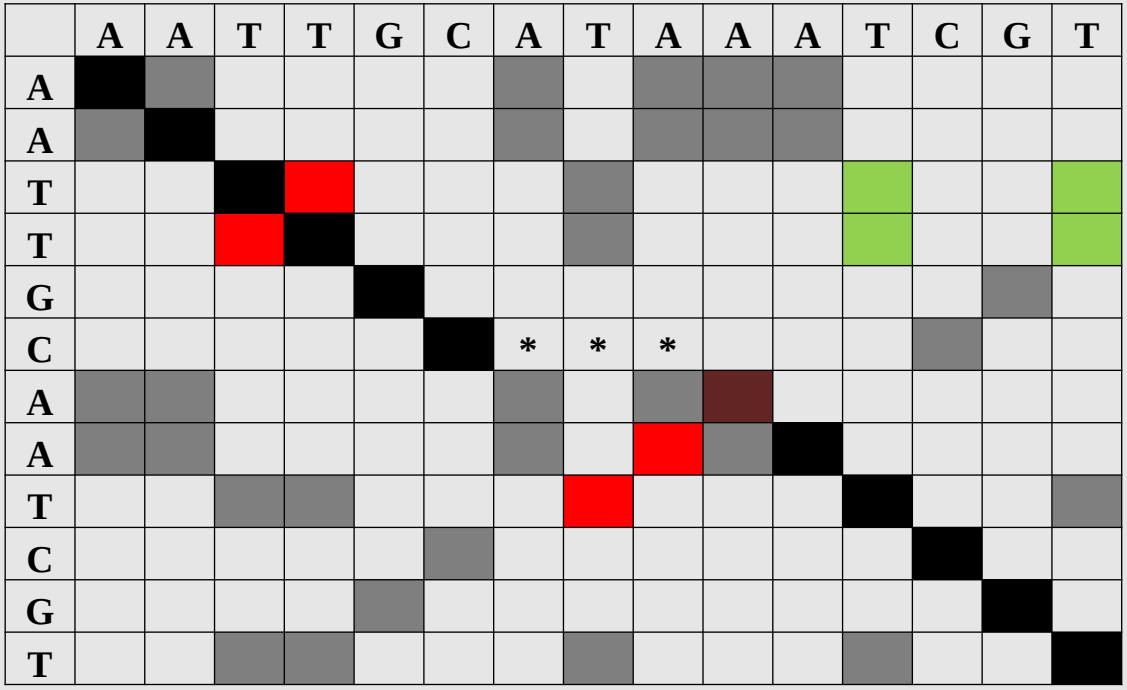
Principe

Les séquences sont positionnées perpendiculairement dans un tableau et on met un point à chaque appariement. La multiplicité des points forme des diagonales. Les décalages correspondent à des indels et les segments parallèles indiquent des répétitions. L'examen de la diagonale (de gauche à droite) permet de trouver le meilleur alignement.

ALIGNEMENT DE DEUX SÉQUENCES

LE DOT PLOT





Zones communes (diagonale)

Palindromes

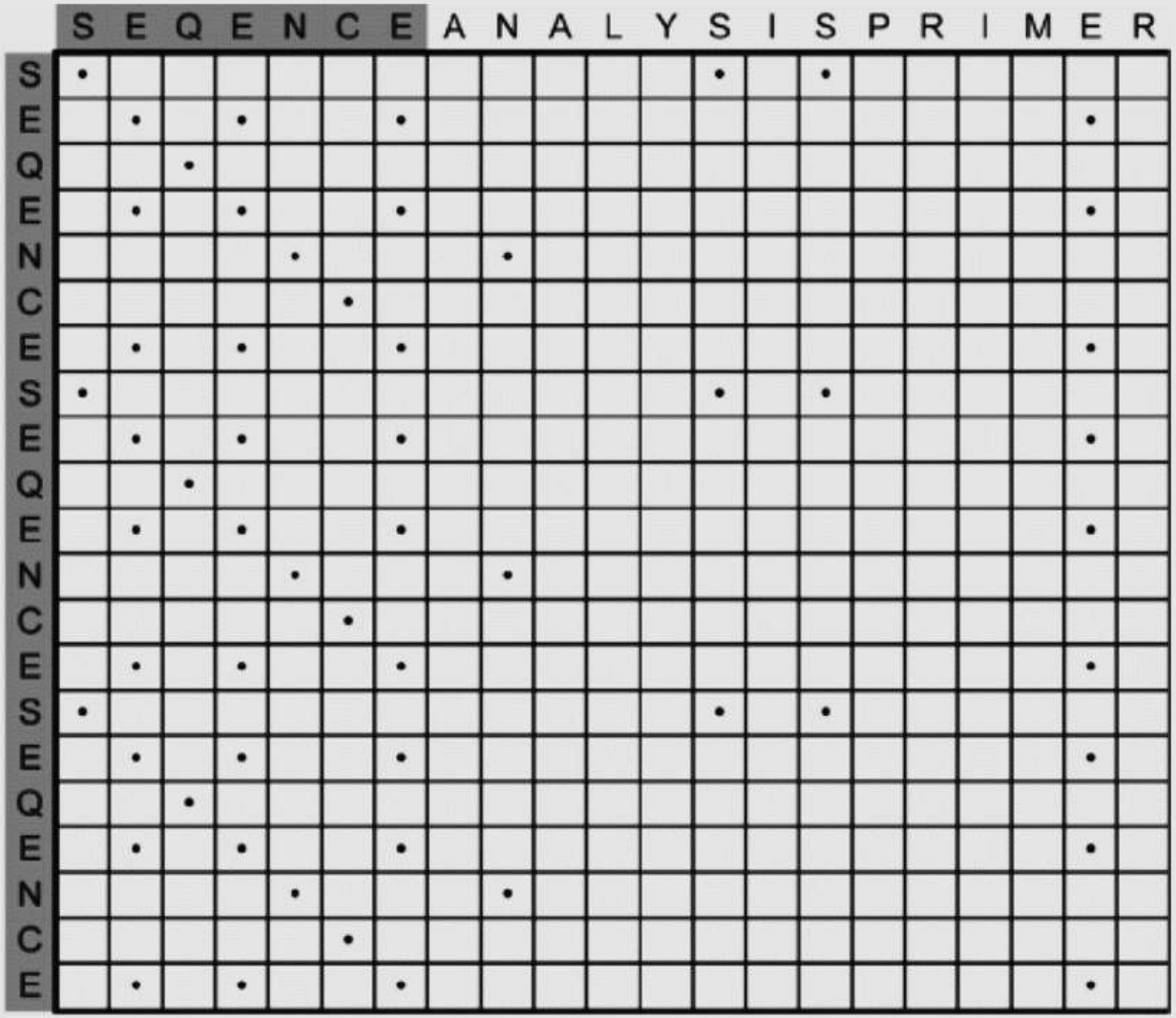
Répétitions du résidu

☼ Répétition interne

* Indels

ALIGNEMENT DE DEUX SÉQUENCES

LE DOT PLOT

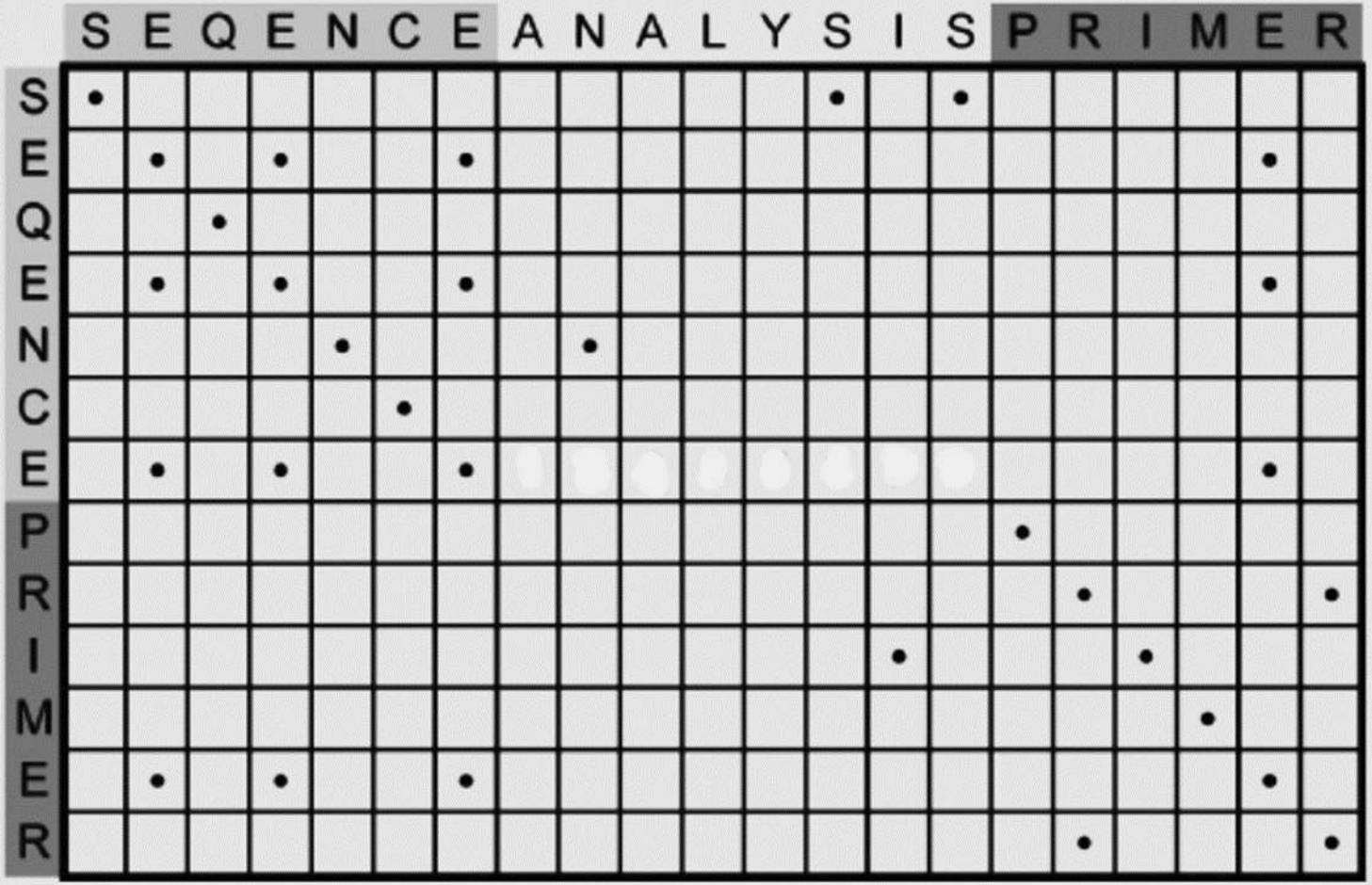


répétition interne du segment « SEQENCE »

Exemples de phénomènes biologiques détectés par le dot plot entre deux séquences

ALIGNEMENT DE DEUX SÉQUENCES

LE DOT PLOT

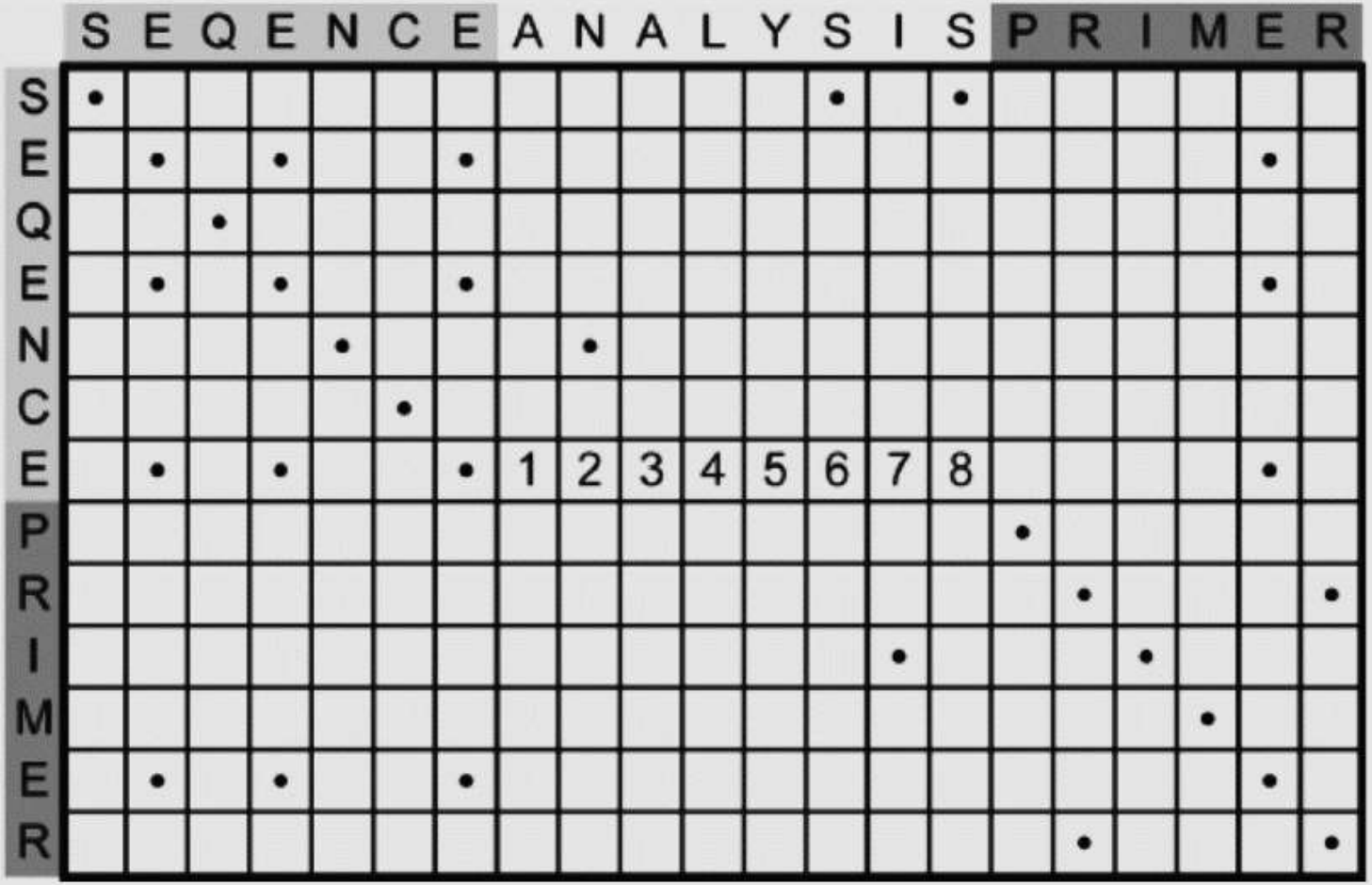


insertion de 8 acides aminés (N° 1 à 8) dans la séquence horizontale. La séquence « ANALYSIS » a été insérée dans la séquence A ou perdue dans la séquence B.

Exemples de phénomènes biologiques
détectés par le dot plot entre deux séquences

ALIGNEMENT DE DEUX SÉQUENCES

LE DOT PLOT

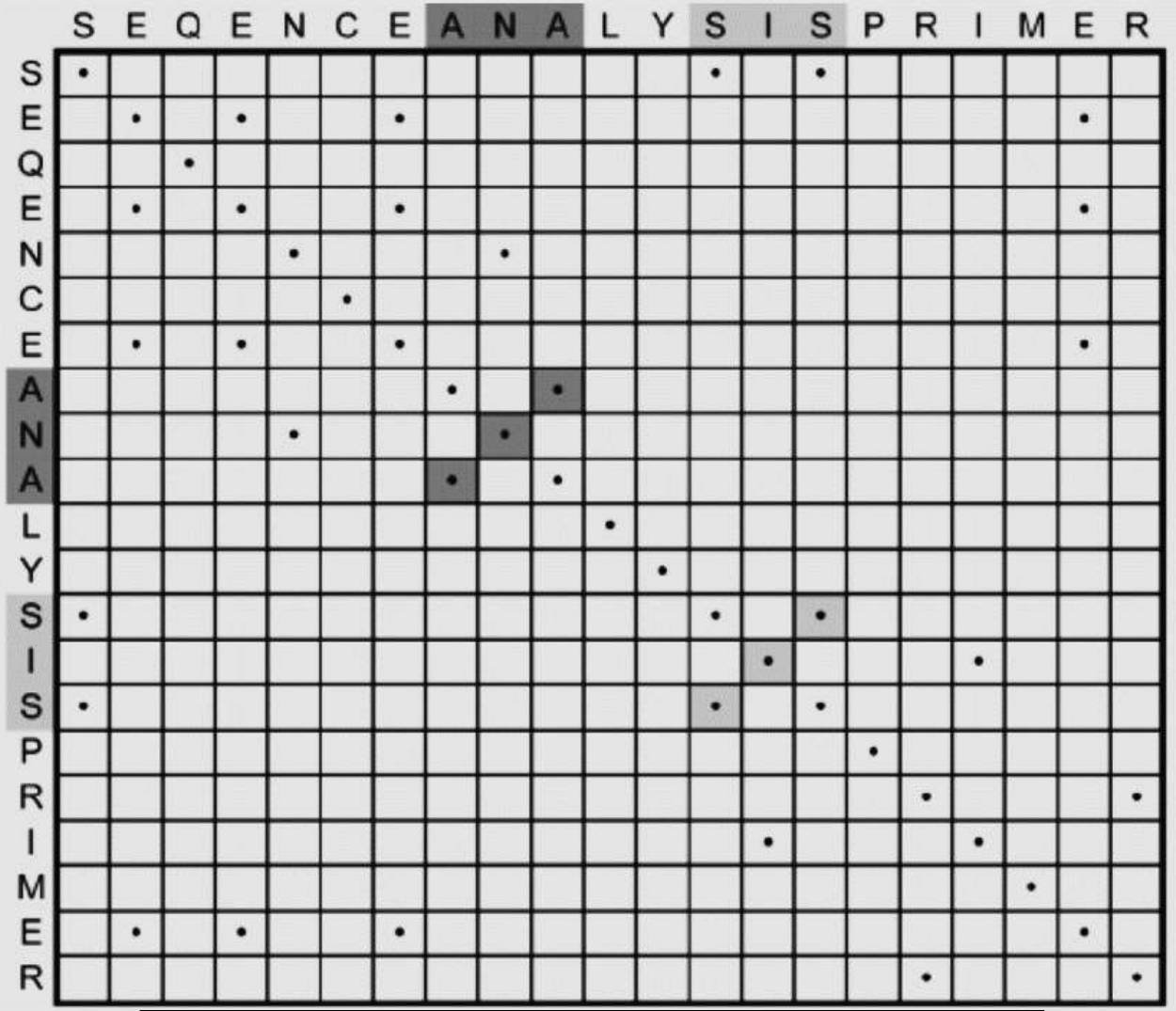


insertion de 8 acides aminés (N° 1 à 8) dans la séquence horizontale. La séquence « ANALYSIS » a été insérée dans la séquence A ou perdue dans la séquence B.

Exemples de phénomènes biologiques
détectés par le dot plot entre deux séquences

ALIGNEMENT DE DEUX SÉQUENCES

LE DOT PLOT

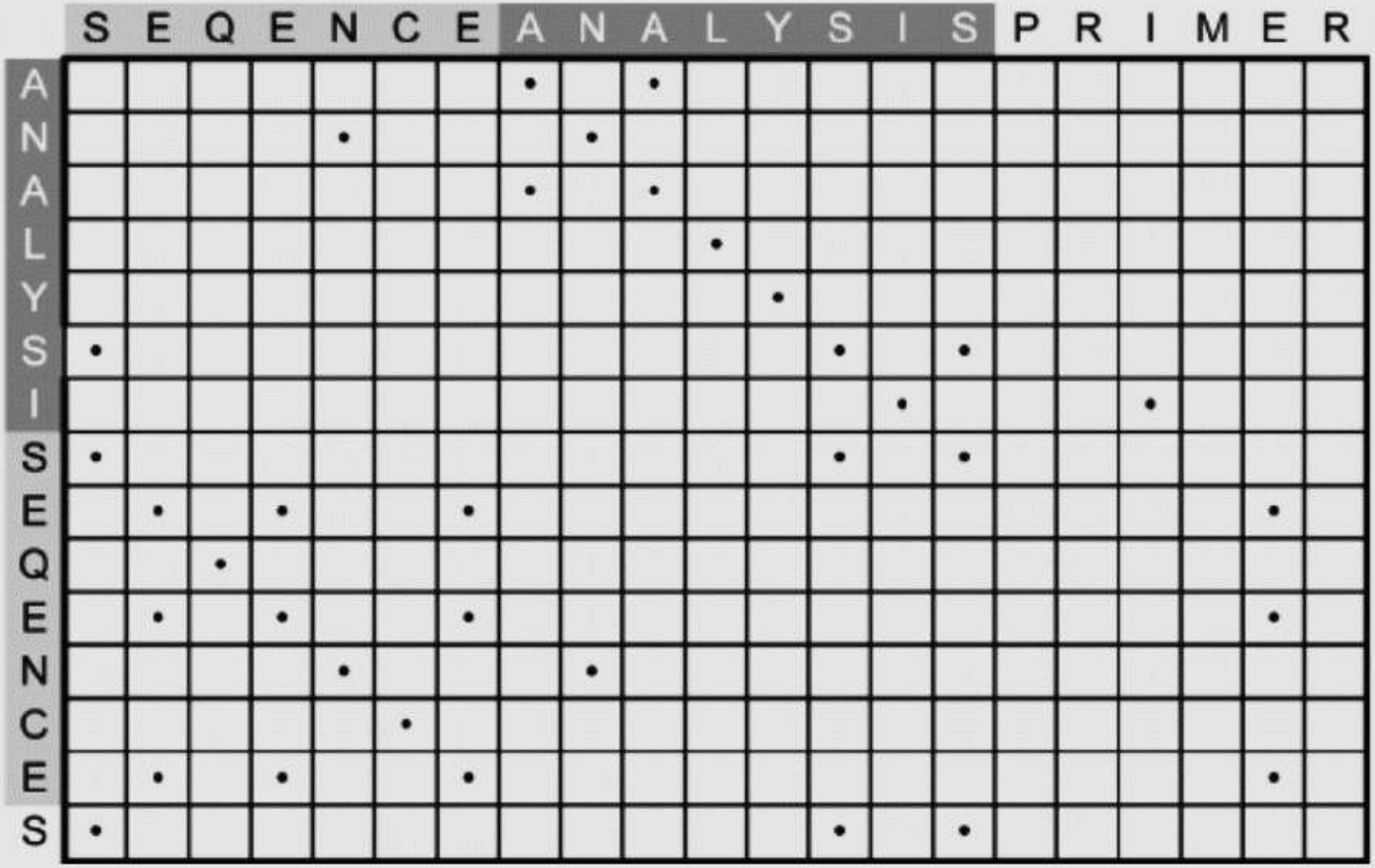


palindromes (ANA, SIS et EQE).

Exemples de phénomènes biologiques détectés par le dot plot entre deux séquences (ici c'est la même séquence)

ALIGNEMENT DE DEUX SÉQUENCES

LE DOT PLOT



transposition (exemple d'inversions des segments « SEQENCE » et « ANALYSIS » dans la séquence)

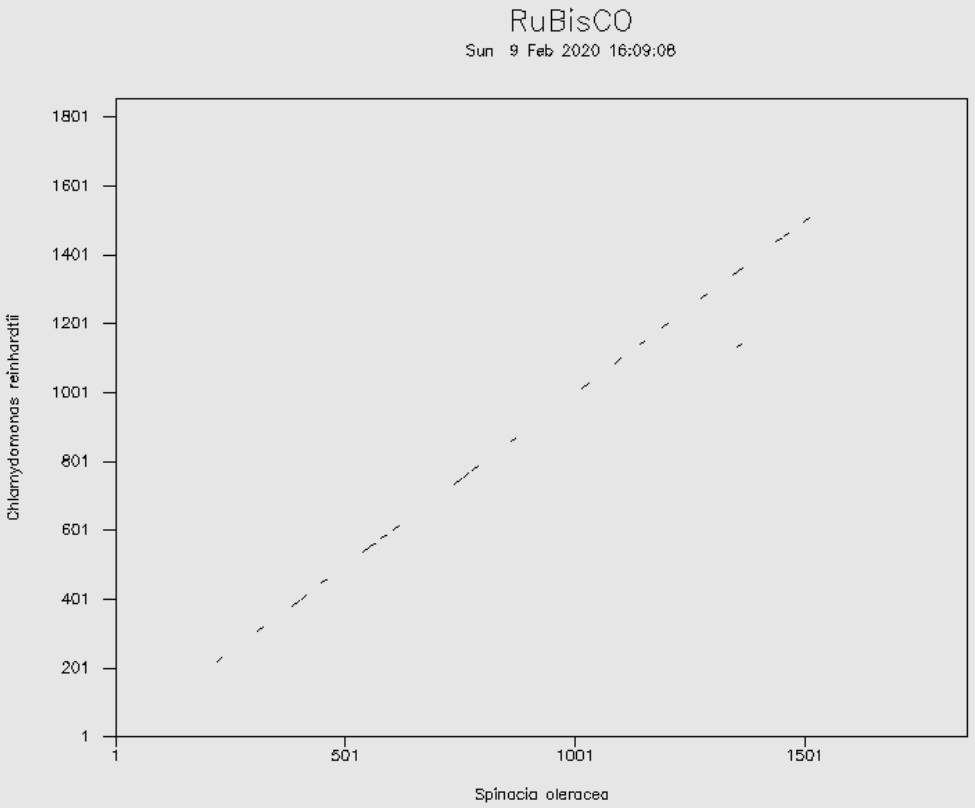
Exemples de phénomènes biologiques
détectés par le dot plot entre deux séquences

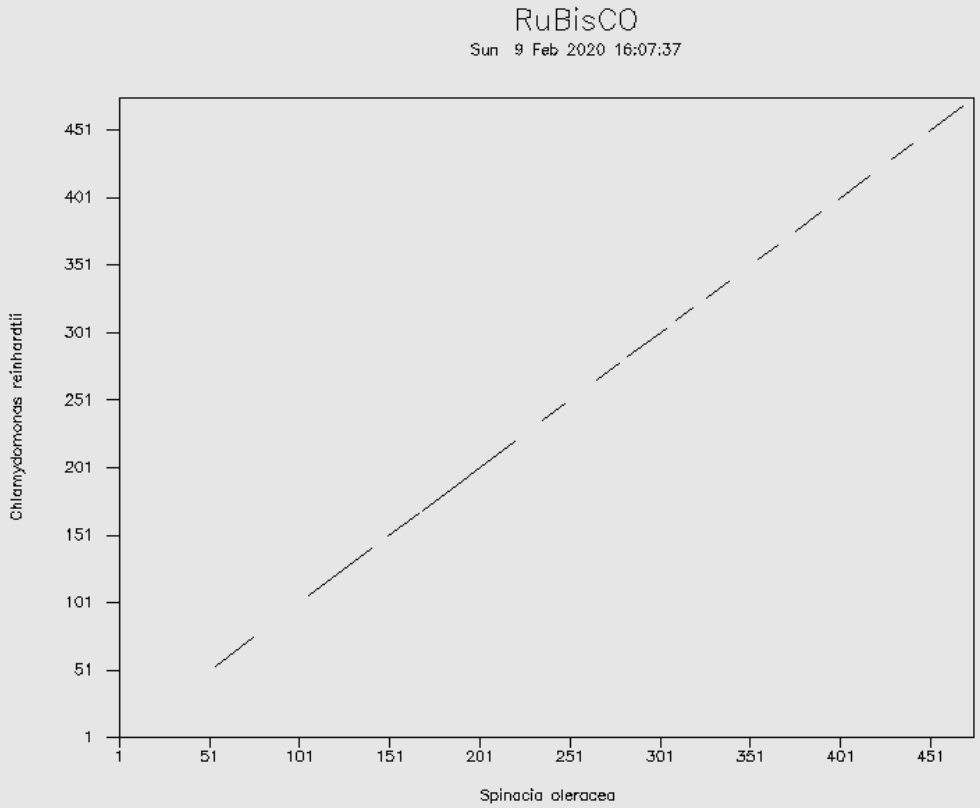
Avantage : permet d'explorer sans à priori toutes les combinaisons possibles de ressemblance entre deux séquences.

Inconvénients : manque de précision, lenteur, limitation à l'alignement par paire uniquement.

ALIGNEMENT DE DEUX SÉQUENCES

LE DOT PLOT

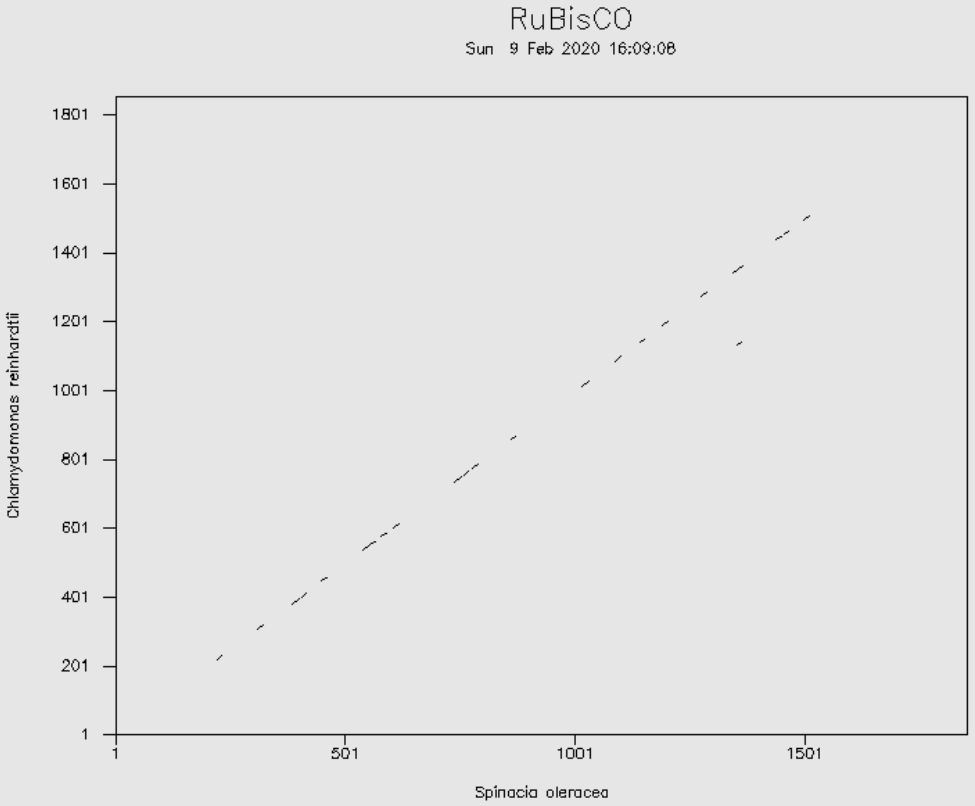


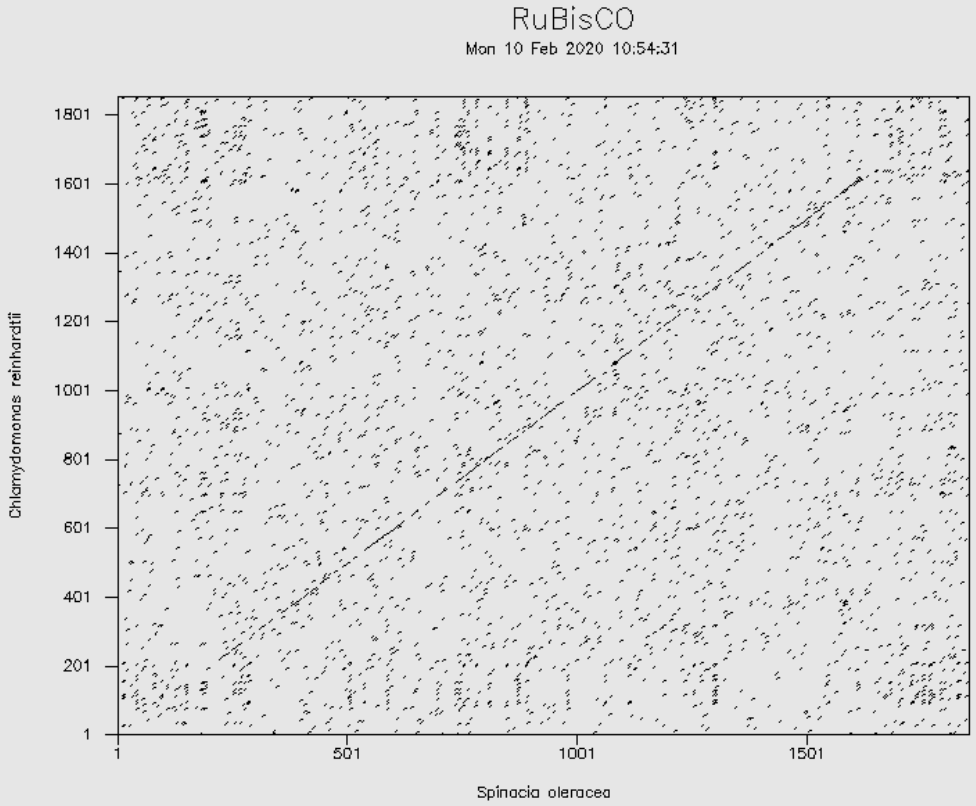


Exemple de comparaison entre deux séquences nucléiques (à gauche) et protéiques (à droite) de l'enzyme RuBisCO (ribulose-1,5-bisphosphate carboxylase/oxygénase) issues de deux groupes de végétaux différents (*Spinacia oleracea* -épinard- et *Chlamydomonas reinhardtii* -algue-). Le dot plot a été réalisé avec le logiciel **Dottup (http://www.bioinformatics.nl/cgi-bin/emboss/dottup)**.

ALIGNEMENT DE DEUX SÉQUENCES

LE DOT PLOT





Pour le même exemple, les zones communes entre les deux séquences nucléiques sont montrées lorsque ces régions sont constituées de mots d'au moins 10 pb de longueur (à gauche) et d'au moins de 3 pb (à droite).

Dot plots réalisés avec le logiciel **Dottup** (http://www.bioinformatics.nl/cgi-bin/emboss/dottup). (Démonstration visionnable ici : https://youtu.be/yKOocOqbHJU)

Soit les deux séquences suivantes :

Séquence 1	**A**	**T**	**T**	**C**	**G**	**A**	**T**		
Séquence 2	**A**	**T**	**T**	**C**	**G**	**A**	**T**	**G**	**C**

Plusieurs possibilités d'alignements :

Si on compare les deux séquences de bout en bout, avec comme début la position 1 sur les 2 séquences, on aura l'alignement suivant :

Séquence 1	**A**	**T**	**T**	**C**	**G**	**A**	**T**	**_**	**_**
	|	|	|	|	|	|	|		
Séquence 2	**A**	**T**	**T**	**C**	**G**	**A**	**T**	**G**	**C**

On définit le **pourcentage d'identité (%id)** comme étant le nombre d'identités ponctuelles **id** (***matchs***) entre deux séquences après alignement des séquences, rapporté à la **longueur** (nombre total de positions) de l'alignement **(T)**.

$$\%id = \#id/T$$

Si on compare les deux séquences de bout en bout, avec comme début la position 1 sur les 2 séquences, on aura l'alignement suivant :

Séquence 1	**A**	**T**	**T**	**C**	**G**	**A**	**T**	**_**	**_**
	|	|	|	|	|	|	|		
Séquence 2	**A**	**T**	**T**	**C**	**G**	**A**	**T**	**G**	**C**

L'écriture des deux séquences donne :

identité (|) = **7** positions ; substitution (*) = **0** ; indels (gap) (_) = **2** ;

longueur de l'alignement = 9 ;

%id = 7/9 x 100 = 77,77 %

Si on décale le début de la comparaison de 2 positions pour la séquence 1, on aura le nouvel alignement :

Séquence 1	_	_	A	T	T	C	G	A	T
			*	*	*	*	*	*	*
Séquence 2	A	T	T	C	G	A	T	G	C

L'écriture des deux séquences donne :

identité (|) = **0** ; substitution (*) = **7** ; indels (gap) (_) = **2** ;

longueur de l'alignement = **9** ;

%id = 0/11 x 100 = 0 %

Si on décale le début de la comparaison de 2 positions pour la séquence 2, on aura le nouvel alignement :

Séquence 1	A	T	T	C	G	A	T	_	_	_	_
			*	*	*	*	*				
Séquence 2	_	_	A	T	T	C	G	A	T	G	C

L'écriture des deux séquences donne :

identité (|) = **0** ; substitution (*) = **5** ; indels (gap) (_) = **6** ;

longueur de l'alignement = 11 ;

%id = 0/11 x 100 = 0 %

Si on suppose qu'une délétion s'est produite avant la dernière nucléotide de la séquence 1, on aura le nouvel alignement :

Séquence 1	**A**	**T**	**T**	**C**	**G**	**A**	**_**	**T**	**_**
	|	**|**	**|**	**|**	**|**	**|**		*****	
Séquence 2	**A**	**T**	**T**	**C**	**G**	**A**	**T**	**G**	**C**

L'écriture des deux séquences donne :

identité (|) = **6** ; substitution (*) = **1** ; indels (gap) (_) = **2** ;

longueur de l'alignement = 9 ;

%id = 6/9 x 100 = 66,66 %

On devra choisir l'alignement qui présentera le **maximum** de positions d'**identité**, le **minimum** de positions présentant des **mutations ponctuelles** sur une **taille d'alignement** la **plus petite** possible :

Alignement 1

Séquence 1	A	T	T	C	G	A	T	_	_
	|	|	|	|	|	|	|		
Séquence 2	A	T	T	C	G	A	T	G	C

%id = 77,77 %

Alignement 2

Séquence 1	_	_	A	T	T	C	G	A	T
			*	*	*	*	*	*	*
Séquence 2	A	T	T	C	G	A	T	G	C

%id = 0 %

Quel(s) alignement choisir?

Alignement 3

Séquence 1	A	T	T	C	G	A	_	T	_
	|	|	|	|	|	|		*	
Séquence 2	A	T	T	C	G	A	T	G	C

%id = 66,66 %

Alignement 4

Séquence 1	A	T	T	C	G	A	T	_	_	_	_
			*	*	*	*	*				
Séquence 2	_	_	A	T	T	C	G	A	T	G	C

%id = 0 %

ALIGNEMENT DE DEUX SÉQUENCES

ALIGNEMENT OPTIMAL DE DEUX SÉQUENCES

On devra choisir l'alignement qui présentera le **maximum** de positions d'**identité**, le **minimum** de positions présentant des **mutations ponctuelles** sur une **taille d'alignement** la **plus petite** possible :

Alignement 1 (optimal)

Séquence 1	A	T	T	C	G	A	T	_	_
	|	|	|	|	|	|	|		
Séquence 2	A	T	T	C	G	A	T	G	C

%id = 77,77 %

Alignement 2

Séquence 1	_	_	A	T	T	C	G	A	T
			*	*	*	*	*	*	*
Séquence 2	A	T	T	C	G	A	T	G	C

%id = 0 %

Alignement 3

Séquence 1	A	T	T	C	G	A	_	T	_
	|	|	|	|	|	|		*	
Séquence 2	A	T	T	C	G	A	T	G	C

%id = 66,66 %

Alignement 4

Séquence 1	A	T	T	C	G	A	T	_	_	_	_
			*	*	*	*	*				
Séquence 2	_	_	A	T	T	C	G	A	T	G	C

%id = 0 %

- ✓ Afin de comparer deux séquences d'une manière objective (indépendante de l'observateur), on doit d'abord les aligner d'une manière optimale. L'alignement **optimal** est obtenu quand la **coïncidence des lettres** composant les deux séquences est **maximale** ;
- ✓ Un alignement optimal est une analyse qui permet de trouver le nombre **minimum de mutations ponctuelles** (insertion-délétion, substitution) qui permettent de transformer une séquence en une autre.

LA DISTANCE

- Elle correspond au nombre d'indels et de substitutions séparant deux séquences A et B ;
- En fonction de la quantité de mutations ponctuelles, la distance entre deux séquences A et B prend la forme suivante :

$$d(A, B) = \Sigma substitutions + \Sigma indels$$

ALIGNEMENT DE DEUX SÉQUENCES

EXEMPLE DE CALCUL DE DISTANCE

De l'exemple précédent :

Alignement 1 (optimal)

Séquence 1	A	T	T	C	G	A	T	_	_
	|	|	|	|	|	|	|		
Séquence 2	A	T	T	C	G	A	T	G	C

d = 2

Alignement 2

Séquence 1	_	_	A	T	T	C	G	A	T
			*	*	*	*	*	*	*
Séquence 2	A	T	T	C	G	A	T	G	C

d = 2 + 7 = 9

Alignement 3

Séquence 1	A	T	T	C	G	A	_	T	_
	|	|	|	|	|	|		*	
Séquence 2	A	T	T	C	G	A	T	G	C

d = 2 + 1 = 3

Alignement 4

Séquence 1	A	T	T	C	G	A	T	_	_	_	_
			*	*	*	*	*				
Séquence 2	_	_	A	T	T	C	G	A	T	G	C

d = 5 + 6 = 11

LE SCORE DE SIMILARITÉ

Le score exprime le degré de similitude entre deux séquences :

$$S(A, B) = \Sigma\text{identité} - (\Sigma\text{substitutions} + \Sigma\text{indels})$$

$$S(A, B) = \Sigma id - d(A, B)$$

Séquence 1	A	T	T	C	G	A	T	_	_
	|	|	|	|	|	|	|		
Séquence 2	A	T	T	C	G	A	T	G	C

S (1, 2) = 7 - 2 = 5

ALIGNEMENT DE DEUX SÉQUENCES

EXEMPLE DE CALCUL DE SCORE DE SIMILARITÉ

De l'exemple précédent :

Alignement 1 (optimal)

Séquence 1	A	T	T	C	G	A	T	_	_
	|	|	|	|	|	|	|		
Séquence 2	A	T	T	C	G	A	T	G	C

S (1, 2) = 7 - 2 = 5

Alignement 2

Séquence 1	_	_	A	T	T	C	G	A	T
			*	*	*	*	*	*	*
Séquence 2	A	T	T	C	G	A	T	G	C

S (1, 2) = 0 – (2 + 7) = - 9

Alignement 3

Séquence 1	A	T	T	C	G	A	_	T	_
	|	|	|	|	|	|		*	
Séquence 2	A	T	T	C	G	A	T	G	C

S (1, 2) = 6 – (2 + 1) = 3

Alignement 4

Séquence 1	A	T	T	C	G	A	T	_	_	_	_
			*	*	*	*	*				
Séquence 2	_	_	A	T	T	C	G	A	T	G	C

S (1, 2) = 0 – (6 + 5) = - 11

LE SCORE DE SIMILARITÉ

Le score exprime le degré de similitude entre deux séquences :

$$S(A, B) = \Sigma\text{identité} - (\Sigma\text{substitutions} + \Sigma\text{indels})$$

$$S(A, B) = \Sigma\text{id} - d(A, B)$$

La construction de l'alignement consiste donc à identifier le meilleur alignement possible entre deux séquences, celui qui **minimise la distance** d'édition d(A, B) ou qui **maximise le score** S(A, B).

LES MATRICES DE SCORES

Le score de similarité peut être calculé de sorte à ce que la différence dans la fréquence et la gravité des mutations ponctuelles soient prise en considération. Dans ce cas **les scores élémentaires (s_e)** peuvent prendre des valeurs autres que (0 ; 1).

Par exemple, les indels provoquent souvent de très graves mutations au niveaux des séquences et leur présence seraient donc moins probables que celles des substitutions. Certaines substitutions se produisent plus souvent que d'autres (pour l'ADN, les transitions apparaissent plus souvent que les transversions), etc.

LES MATRICES DE SCORES

Pour représenter ces informations, nous pouvons utiliser la matrice suivante :

s_e	A	C	G	T
A	2	-2	-1	-2
C	-2	2	-2	-1
G	-1	-2	2	-2
T	-2	-1	-2	2

Score d'identité ponctuelle = 2

Pénalité de transition = -1

Pénalité de transversion = -2

Pénalité d'ouverture de gap = -4

pénalité d'extension de gap = -3

ALIGNEMENT DE DEUX SÉQUENCES

De l'exemple précédent :

Alignement 1 (optimal)

Séquence 1	A	T	T	C	G	A	T	_	_
	|	|	|	|	|	|	|		
Séquence 2	A	T	T	C	G	A	T	G	C
Se	2	2	2	2	2	2	2	-4	-3

$$S(1, 2) = \sum s_e = 2 + 2 + 2 + 2 + 2 + 2 + 2 - 4 - 3 = 7$$

Alignement 3

Séquence 1	A	T	T	C	G	A	_	T	_
	|	|	|	|	|	|		*	
Séquence 2	A	T	T	C	G	A	T	G	C
Se	2	2	2	2	2	2	-4	-2	-4

$$S(1, 2) = \sum s_e = 2$$

EXEMPLE DE CALCUL DE SCORE DE SIMILARITÉ

s_e	A	C	G	T
A	2	-2	-1	-2
C	-2	2	-2	-1
G	-1	-2	2	-2
T	-2	-1	-2	2

Pénalité d'ouverture de gap = -4

pénalité d'extension de gap = -3

ALIGNEMENT DE DEUX SÉQUENCES

EXEMPLE DE CALCUL DE SCORE DE SIMILARITÉ

De l'exemple précédent :

s_e	A	C	G	T
A	2	-2	-1	-2
C	-2	2	-2	-1
G	-1	-2	2	-2
T	-2	-1	-2	2

Alignement 2

Séquence 1	_	_	A	T	T	C	G	A	T
			*	*	*	*	*	*	*
Séquence 2	A	T	T	C	G	A	T	G	C
Se	-4	-3	-2	-1	-2	-2	-2	-1	-1

$S(1, 2) = \sum s_e = -18$

Pénalité d'ouverture de gap = -4

pénalité d'extension de gap = -3

Alignement 4

Séquence 1	A	T	T	C	G	A	T	_	_	_	_
			*	*	*	*	*				
Séquence 2	_	_	A	T	T	C	G	A	T	G	C
Se	-4	-3	-2	-1	-2	-2	-2	-4	-3	-3	-3

$S(1, 2) = \sum s_e = -29$

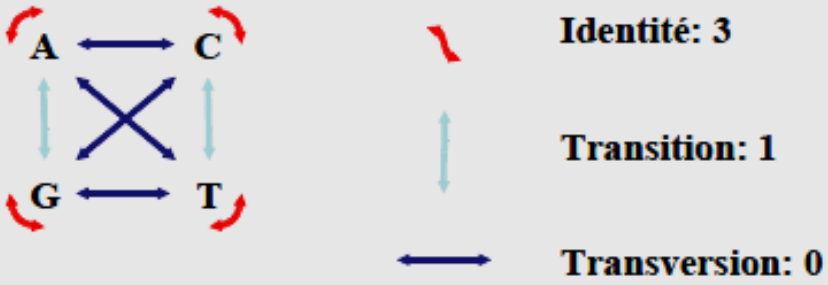
Matrices de scores pour l'ADN

- La matrice identité

match ⟶ 1
mismatch ⟶ 0

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

- La matrice de transition/transversion (substitutions)



Identité: 3

Transition: 1

Transversion: 0

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

- La matrice identité dans BLAST

	A	C	G	T
A	5	-4	-4	-4
C	-4	5	-4	-4
G	-4	-4	5	-4
T	-4	-4	-4	5

Matrices pour les protéines

➢ Matrice de substitution PAM250

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
A	2																			
R	-2	6																		
N	0	0	2																	
D	0	-1	2	4																
C	-2	-4	-4	-5	4															
Q	0	1	1	2	-5	4														
E	0	-1	1	3	-5	2	4													
G	1	-3	0	1	-3	-1	0	5												
H	-1	2	2	1	-3	3	1	-2	6											
I	-1	-2	-2	-2	-2	-2	-2	-3	-2	5										
L	-2	-3	-3	-4	-6	-2	-3	-4	-2	2	6									
K	-1	3	1	0	-5	1	0	-2	0	-2	-3	5								
M	-1	0	-2	-3	-5	-1	-2	-3	-2	2	4	0	6							
F	-4	-4	-4	-6	-4	-5	-5	-5	-2	1	2	-5	0	9						
P	1	0	-1	-1	-3	0	-1	-1	0	-2	-3	-1	-2	-5	6					
S	1	0	1	0	0	-1	0	1	-1	-1	-3	0	-2	-3	1	3				
T	1	-1	0	0	-2	-1	0	0	-1	0	-2	0	-1	-2	0	1	3			
W	-6	2	-4	-7	-8	-5	-7	-7	-3	-5	-2	-3	-4	0	-6	-2	-5	17		
Y	-3	-4	-2	-4	0	-4	-4	-5	0	-1	-1	-4	-2	7	-5	-3	-3	0	10	
V	0	-2	-2	-2	-2	-2	-2	-1	-2	4	2	-2	2	-1	-1	-1	0	-6	2	4

- ✓ Il n'est pas simple de trouver le meilleur alignement de deux séquences. Par exemple, pour deux séquences de 140 résidus chacune, il y a 10^{82} alignements possibles. La **programmation dynamique** permet de résoudre efficacement ce problème.
- ✓ La programmation dynamique est une **méthode algorithmique** utilisée pour déterminer l'alignement optimal entre deux séquences.
- ✓ La programmation dynamique permet de trouver l'**alignement optimal global** ou **local** entre **deux** séquences **nucléiques** ou **protéiques**.

- ✓ Un algorithme est la description **précise**, sous forme de concepts simples ou d'**instructions**, de la manière dont on peut **résoudre** un problème ;
- ✓ L'algorithme de Needleman et Wunsch permet de réaliser un alignement **global** entre deux séquences nucléiques. Son expression est de la forme :

$$S(i,j) = Max \begin{cases} S(i-1,j-1) + s(i,j) \\ S(i-1,j) + p \\ S(i,j-1) + p \end{cases}$$

p étant une pénalité attribuée aux gaps apparaissant sur les positions i-1 ou j-1

ALIGNEMENT DE DEUX SÉQUENCES

NEEDLEMAN ET WUNCH : EXEMPLE PRATIQUE

- ✓ Réalisons un alignement global des deux séquences **(S_1, S_2)** suivantes de taille m et n respectivement (n et m peuvent être inégales) :

S_1 = TAAGTCCG m = 8 et S_2 = TAAGTACG n = 8

- ✓ Pour construire l'alignement entre les deux séquences S_1 et S_2, quatre étapes sont nécessaires :

ETAPE 1: CALCUL DE LA MATRICE INITIALE

- Il s'agit d'insérer les deux séquences S_1 et S_2 dans une matrice de sorte que S_1 soit à l'horizontal et S_2 à la verticale du tableau, puis remplir les cases par 1 (identité des deux nucléotides de S_1 et de S_2) ou 0 (sinon) :

(matrice d'identité)
Match = 1
Mismatch = 0
p = 0
(gaps ignorées)

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

ETAPE 1: CALCUL DE LA MATRICE INITIALE

- On peut l'écrire plus simplement de cette manière :

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

Match = 1 ; Mismatch = 0 ; p = 0

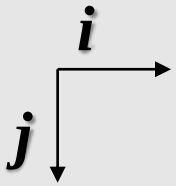
ETAPE 2: CALCUL DE LA MATRICE TRANSFORMÉE: INITIALISATION DE LA MATRICE

- Nouvelle matrice (m+2, n+2) dans laquelle la 1ère ligne et la 1ère colonne sont initialisées à zéro :

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

ETAPE 2: CALCUL DE LA MATRICE TRANSFORMÉE

✓ L'application de l'algorithme de Needleman et Wunsh permet de remplir les cases de cette matrice. Le résultat est le suivant :



$$S(i,j) = Max \begin{cases} S(i-1,j-1) + s(i,j) \\ S(i-1,j) + (p = 0) \\ S(i,j-1) + (p = 0) \end{cases}$$

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

ETAPE 2: CALCUL DE LA MATRICE TRANSFORMÉE

✓ L'application de l'algorithme de Needleman et Wunsh permet de remplir case par case cette matrice :

$$S(i,j) = Max \begin{cases} S(i-1,j-1) + s(i,j) \\ S(i-1,j) + (p=0) \\ S(i,j-1) + (p=0) \end{cases}$$

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

$S(i,j-1)$: case au-dessus de $S(i,j)$

$S(i-1,j-1)$: case en diagonale gauche de $S(i,j)$

$S(i-1,j)$: case à gauche de $S(i,j)$

Matrice initiale

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

$s(i,j)$

ETAPE 2: CALCUL DE LA MATRICE TRANSFORMÉE

✓ L'application de l'algorithme de Needleman et Wunsh permet de remplir case par case cette matrice :

$$S(i,j) = Max \begin{cases} 0+1 \\ 0+0 \\ 0+0 \end{cases}$$

			A	A	G	T	C	C	G
		T	A	A	G	T	C	C	G
	$S(i,j-1)$								
	$S(i-1,j-1)$ →	0	0	0	0	0	0	0	0
T	0	$S(i,j)$							
A	0	$S(i-1,j)$							
A	0								
G	0								
T	0								
A	0								
C	0								
G	0								

i : →

j : ↓

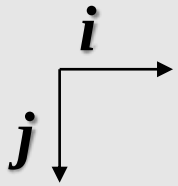
Matrice initiale

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

$s(i,j)$

ETAPE 2: CALCUL DE LA MATRICE TRANSFORMÉE

✓ L'application de l'algorithme de Needleman et Wunsh permet de remplir case par case cette matrice :

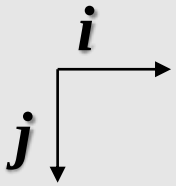


$$S(i,j) = 1$$

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

ETAPE 2: CALCUL DE LA MATRICE TRANSFORMÉE

✓ L'application de l'algorithme de Needleman et Wunsh permet de remplir les cases de cette matrice. Le résultat est le suivant :



$$S(i,j) = Max \begin{cases} S(i-1,j-1) + s(i,j) \\ S(i-1,j) + (p = 0) \\ S(i,j-1) + (p = 0) \end{cases}$$

		T	A	A	G	T	C	C	G
	0	0	0	0	0	0	0	0	0
T	0	1	1	1	1	1	1	1	1
A	0	1	2	2	2	2	2	2	2
A	0	1	2	3	3	3	3	3	3
G	0	1	2	3	4	4	4	4	4
T	0	1	2	3	4	5	5	5	5
A	0	1	2	3	4	5	5	5	5
C	0	1	2	3	4	5	6	6	6
G	0	1	2	3	4	5	6	6	7

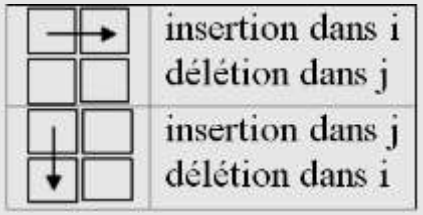
ETAPE 3: TRAÇAGE DU PARCOURS DE LA MATRICE TRANSFORMÉE

✓ Parcourir la matrice transformée depuis le plus haut score calculé **en bas à droite** (ici S=7) jusqu'au score le plus petit (ici S=1) :

j \ *i*		T	A	A	G	T	C	C	G
	0	0	0	0	0	0	0	0	0
T	0	**1**	1	1	1	1	1	1	1
A	0	1	**2**	2	2	2	2	2	2
A	0	1	2	**3**	3	3	3	3	3
G	0	1	2	3	**4**	4	4	4	4
T	0	1	2	3	4	**5**	5	5	5
A	0	1	2	3	4	**5**	5	5	5
C	0	1	2	3	4	5	**6**	**6**	6
G	0	1	2	3	4	5	6	6	**7**

Fin

Début



ETAPE 4: CONSTRUCTION DE L'ALIGNEMENT DES DEUX SÉQUENCES ET CALCUL DU SCORE DE SIMILARITÉ

Séquence S_1	**T**	**A**	**A**	**G**	**T**	**—**	**C**	**C**	**G**
	|	|	|	|	|		|		|
Séquence S_2	**T**	**A**	**A**	**G**	**T**	**A**	**C**	**—**	**G**

✓ Le score global de cet alignement est égal à la somme des matchs moins la distance (nous avions choisi : match = 1; mismatch = 0 ; p = 0) :

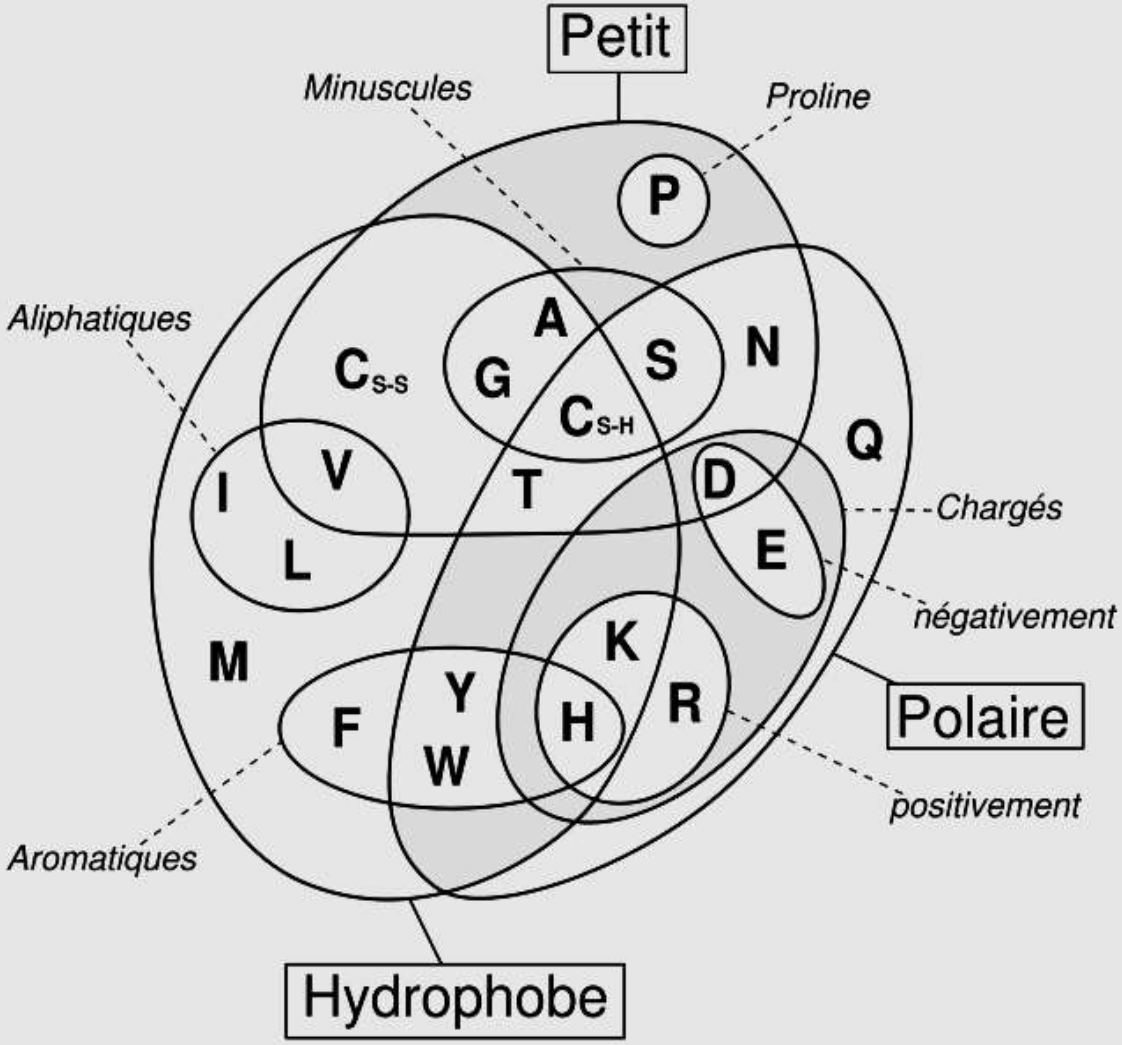
$$S(S_1, S_2) = (7*1) - (2*0) = 7$$

✓ Le pourcentage d'identité entre S_1 et S_2 est égal au rapport entre le nombre d'identités ponctuelles (7) entre les deux séquences et la longueur de l'alignement (9) :

- ✓ Les matrices protéiques utilisées pour réaliser l'alignement des séquences protéiques sont basées sur la nature physico-chimiques des acides aminés ;
- ✓ Certains acides aminés peuvent être remplacés par d'autres sans altérer le rôle et la fonction biologique de la protéine ;
- ✓ Les acides aminés sont classés en familles selon leurs propriétés ;
- ✓ Le système de scores utilisé prend en compte l'affinité des résidus protéiques entre eux qui permet à un acide aminé d'être substitué par un autre ;
- ✓ Deux structures protéiques peuvent avoir des séquences similaires (non-identiques) et une fonction conservée.

Code international des acides aminés selon l'IUPAC (*INTERNATIONAL UNION OF PURE AND APPLIED CHEMISTRY*)

G - Glycine (Gly)|P - Proline (Pro)|A - Alanine (Ala)|V - Valine (Val)|L - Leucine (Leu)|I - Isoleucine (Ile)|M - Methionine (Met)|C - Cysteine (Cys)|F - Phenylalanine (Phe)|Y - Tyrosine (Tyr)|W – Tryptophan (Trp)|H - Histidine (His)|K - Lysine (Lys)|R - Arginine (Arg)|Q - Glutamine (Gln)|N- Asparagine (Asn)|E - Glutamic Acid (Glu)|D - Aspartic Acid (Asp)|S - Serine (Ser)|T - Threonine (Thr)



Les matrices protéiques sont des matrices de substitution, elles peuvent être classées en deux catégories :

1. **Matrices évolutives**. montrant le caractère de substitution des acides aminées au cours de l'évolution (matrices liées à l'évolution). Les plus utilisées ;
2. **Matrices physicochimiques**. basées sur les caractéristiques physico-chimiques des acides aminés : caractère hydrophile ou hydrophobe des protéines, structure secondaire ou tertiaire des protéines, etc.

ALIGNEMENT DE DEUX SÉQUENCES

EXEMPLES DE MATRICES DE SUBSTITUTIONS DES ACIDES AMINÉS

Auteurs (année)	Matrice - principe de construction
Fitch & Margoliash (1967)	*minimum base change matrix for amino acid exchange converted to similarity measure*
Dayhoff et al. (1978)	matrices PAM
McLachlan (1971)	matrice dérivée de 16 familles de protéines
Grantham (1974)	matrice dérivée de trois propriétés physico-chimiques des acides aminés
Doolittle (1979)	*intuitive structural-genetic matrix*
Miyata et al. (1979)	matrice dérivée de la polarité et du volume moléculaire des acides aminés (Grantham, 1974)
Levin et al. (1986)	matrice empirique & structures secondaires
Rao (1987)	matrice dérivée des paramètres de Chou & Fasman (1974)
Risler et al. (1988)	matrice dérivée de la comparaison des structures 3D de 11 familles de protéines homologues
Gonnet et al. (1992)	matrices Gonnet
Henikoff & Henikoff (1992)	matrices BLOSUM
Jones et al. (1992)	matrices MS dérivées de 23 000 séquences de protéines
Johnson & Overington (1993)	matrices JOHM - substitutions dans des parties similaires des structures de protéines
Jones et al. (1994)	matrices JTT
Ng et al. (2000)	matrices PHDhtm
Kann et al. (2000)	matrice OPTIMA
Muller et al. (2002)	matrices VTML
Midic et al. (2009)	matrices MidicMat - régions désordonnées
Yamada & Tomii (2014)	matrices MIQS
Keul et al. (2017)	matrices PFASUM
Jia & Jernigan (2018)	matrices SeqStruct - corrélations [séquences / contacts au sein des structures de protéines]
Trivedi & Nagarajaram (2019)	matrices EDSSMat - acides aminés des régions désordonnées des protéines eucaryotes

EXEMPLES DE MATRICES ÉVOLUTIVES :

1. LES MATRICES DE SUBSTITUTION PAM (*Point Accepted Mutation*)

- ✓ Développées à partir de l'alignement global d'environ 16130 séquences très semblables (> 85% d'identité) appartenant à 2621 familles de protéines ;
- ✓ Représentent les échanges **possibles** et **acceptables** d'un acide aminé par un autre lors de l'évolution des protéines.

EXEMPLE DE LA MATRICE DE SUBSTITUTION PAM250

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
A	2																			
R	-2	6																		
N	0	0	2																	
D	0	-1	2	4																
C	-2	-4	-4	-5	4															
Q	0	1	1	2	-5	4														
E	0	-1	1	3	-5	2	4													
G	1	-3	0	1	-3	-1	0	5												
H	-1	2	2	1	-3	3	1	-2	6											
I	-1	-2	-2	-2	-2	-2	-2	-3	-2	5										
L	-2	-3	-3	-4	-6	-2	-3	-4	-2	2	6									
K	-1	3	1	0	-5	1	0	-2	0	-2	-3	5								
M	-1	0	-2	-3	-5	-1	-2	-3	-2	2	4	0	6							
F	-4	-4	-4	-6	-4	-5	-5	-5	-2	1	2	-5	0	9						
P	1	0	-1	-1	-3	0	-1	-1	0	-2	-3	-1	-2	-5	6					
S	1	0	1	0	0	-1	0	1	-1	-1	-3	0	-2	-3	1	3				
T	1	-1	0	0	-2	-1	0	0	-1	0	-2	0	-1	-2	0	1	3			
W	-6	2	-4	-7	-8	-5	-7	-7	-3	-5	-2	-3	-4	0	-6	-2	-5	17		
Y	-3	-4	-2	-4	0	-4	-4	-5	0	-1	-1	-4	-2	7	-5	-3	-3	0	10	
V	0	-2	-2	-2	-2	-2	-2	-1	-2	4	2	-2	2	-1	-1	-1	0	-6	2	4

- ✓ **Score Positif** : les résidus sont similaires, les mutations entre eux arrivent plus souvent qu'attendues par hasard ;
- ✓ **Score Négatif** : les résidus sont dissimilaires, les mutations entre eux arrivent moins souvent qu'attendus par hasard ;

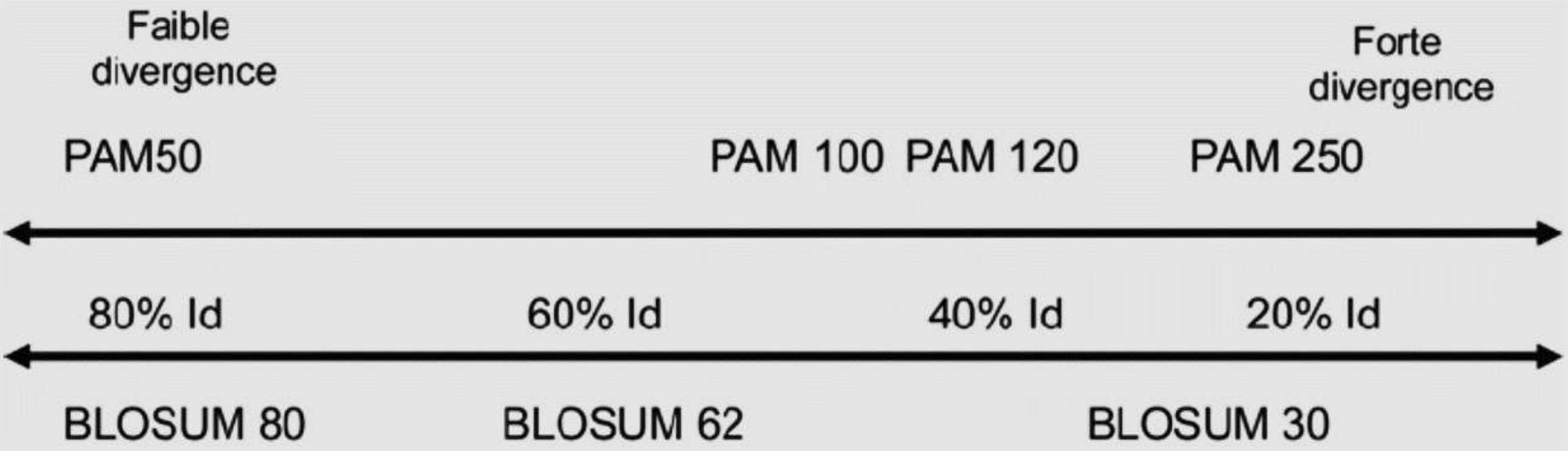
2. LES MATRICES DE SUBSTITUTION BLOSUM (*BLOcks Substitution Matrix*)

- ✓ Observation de blocs d'acides aminés issus des régions conservées de protéines relativement éloignées mais apparentées (alignement local) ;
- ✓ Pour mesurer les fréquences des acides aminés, 2 000 blocs sans gaps de 500 groupes de protéines ont été examinés en comptant le nombre de correspondances et le nombre de mésappariements de chacun des 20 acides aminés.

LA MATRICE DE SUBSTITUTION BLOSUM62

	C	S	T	P	A	G	N	D	E	Q	H	R	K	M	I	L	V	F	Y	W
C	9																			
S	-1	4																		
T	-1	1	5																	
P	-3	-1	-1	7																
A	0	1	0	-1	4															
G	-3	0	-2	-2	0	6														
N	-3	1	0	-2	-2	0	6													
D	-3	0	-1	-1	-2	-1	1	6												
E	-4	0	-1	-1	-1	-2	0	2	5											
Q	-3	0	-1	-1	-1	-2	0	0	2	5										
H	-3	-1	-2	-2	-2	-2	1	-1	0	0	8									
R	-3	-1	-1	-2	-1	-2	0	-2	0	1	0	5								
K	-3	0	-1	-1	-1	-2	0	-1	1	1	-1	2	5							
M	-1	-1	-1	-2	-1	-3	-2	-3	-2	0	-2	-1	-1	5						
I	-1	-2	-1	-3	-1	-4	-3	-3	-3	-3	-3	-3	-3	1	4					
L	-1	-2	-1	-3	-1	-4	-3	-4	-3	-2	-3	-2	-2	2	2	4				
V	-1	-2	0	-2	0	-3	-3	-3	-2	-2	-3	-3	-2	1	3	1	4			
F	-2	-2	-2	-4	-2	-3	-3	-3	-3	-3	-1	-3	-3	0	0	0	-1	6		
Y	-2	-2	-2	-3	-2	-3	-2	-3	-2	-1	2	-2	-2	-1	-1	-1	-1	3	7	
W	-2	-3	-2	-4	-3	-2	-4	-4	-3	-2	-2	-3	-3	-1	-3	-2	-3	1	2	11

- ✓ BLOSUM62 a été obtenue à partir d'un jeu de séquences ayant un pourcentage d'identité d'environ 62%. Matrice très utilisée ;
- ✓ Les scores entre résidus correspondent à des probabilités calculées à partir des fréquences de correspondances et de mésappariements des résidus puis converties en échelle logarithmique.



GAMME D'UTILISATION DES MATRICES PAM ET BLOSUM

L'ALGORITHME DE NEEDLEMAN ET WUNSCH POUR LE CAS DES PROTÉINES

- ✓ Une matrice est d'abord constituée avec une séquence disposée verticalement et l'autre horizontalement.
- ✓ On commence par attribuer à chaque cellule de la matrice (i,j) la valeur correspondant au maximum du score dans la ligne (i+1) et la colonne (j+1) à laquelle on additionne la valeur d'échange des acides aminés appariés en (i,j) ;
- ✓ À la fin de ce processus, chaque cellule (i,j) contient donc le score maximal pour toutes les sous-séquences jusqu'au point (i,j) ;
- ✓ La dernière étape consiste à retracer l'alignement à partir des cellules contenant les scores les plus élevés ;
- ✓ L'équation suivante résume le principe de calcul d'une case de la matrice transformée :

L'ALGORITHME DE NEEDLEMAN ET WUNSCH POUR LE CAS DES PROTÉINES

S (i,j) = se (i,j) + max (S(x,y))

avec :

$i < x \leq m$ et $y = j+1$

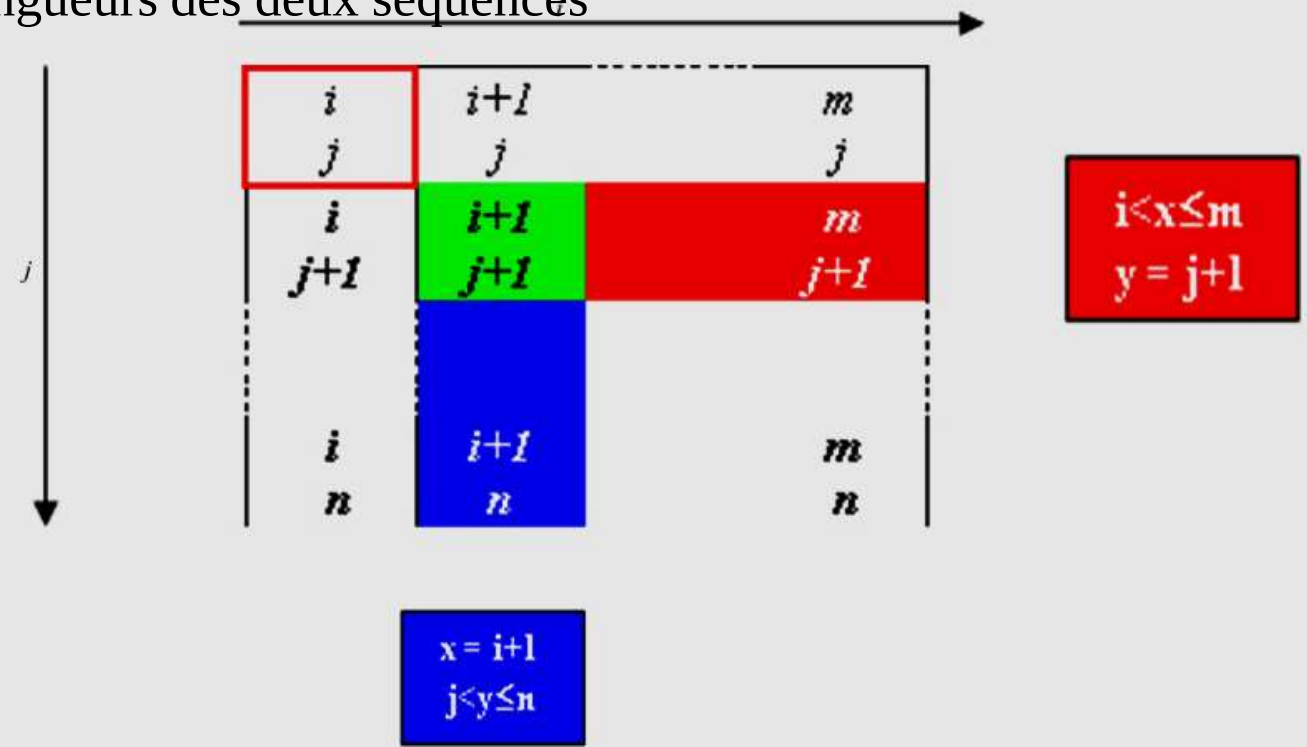
ou

$x = i+1$ et $j < y \leq n$

S(i,j) est le score somme de la case d'indice i et j ;

se le score élémentaire de la case d'indice i et j de la matrice initiale (score issu de la matrice de substitution)

m et n sont les longueurs des deux séquences



Exemple d'alignement avec utilisation de la matrice de substitution PAM250 :

On considère les deux séquences suivantes:

S_1 = VTEERDAF **m = 8** et S_2 = LTSHEAL **n = 7**

ETAPE 1: CALCUL DE LA MATRICE INITIALE À PARTIR DE PAM250

(matrice de substitution)

	V	T	E	E	R	D	A	F
L	2	-2	-3	-3	-3	-4	-2	2
T	0	3	0	0	-1	0	1	-2
S	-1	1	0	0	0	0	1	-3
H	-2	-1	1	1	2	1	-1	-2
E	-2	0	4	4	-1	3	0	-5
A	0	1	0	0	-2	0	2	-4
L	2	-2	-3	-3	-3	-4	-2	2

ETAPE 2: CALCUL DE LA MATRICE TRANSFORMÉE

INITIALISATION DE LA MATRICE

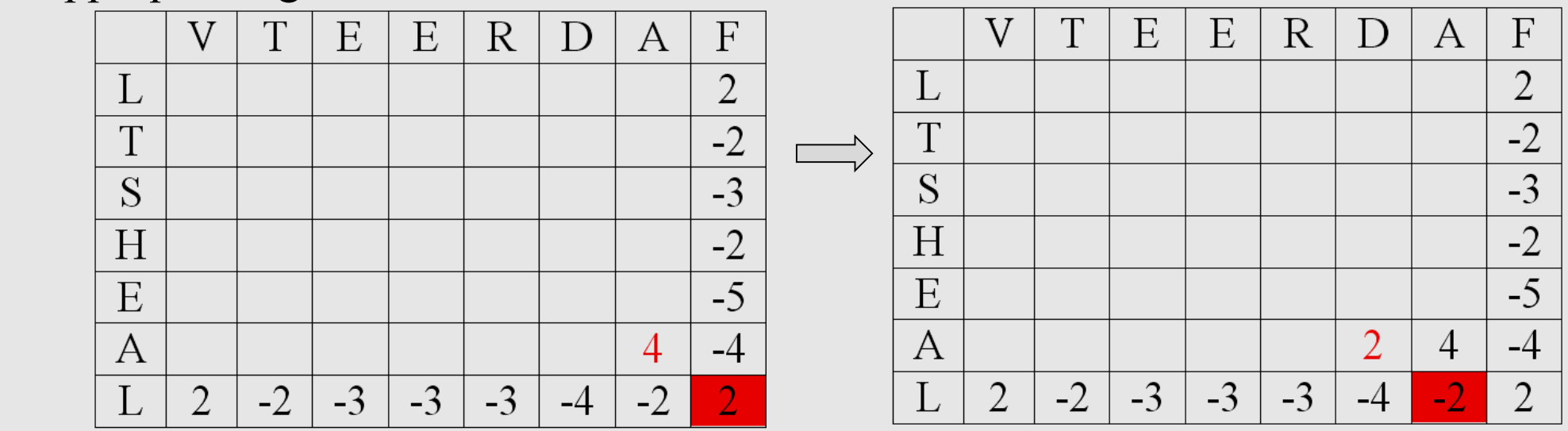
On commence par noter les valeurs de la dernière colonne et de la dernière ligne :

	V	T	E	E	R	D	A	F
L								2
T								-2
S								-3
H								-2
E								-5
A								-4
L	2	-2	-3	-3	-3	-4	-2	2

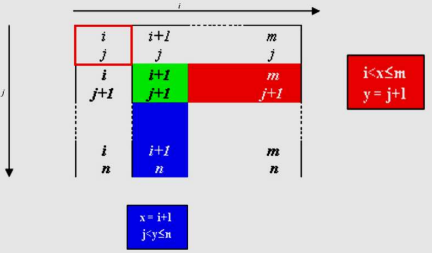
ETAPE 2: CALCUL DE LA MATRICE TRANSFORMÉE

On applique l'algorithme de Needleman et Wunch :

	V	T	E	E	R	D	A	F
L								2
T								-2
S								-3
H								-2
E								-5
A							4	-4
L	2	-2	-3	-3	-3	-4	-2	2



	V	T	E	E	R	D	A	F
L								2
T								-2
S								-3
H								-2
E								-5
A						2	4	-4
L	2	-2	-3	-3	-3	-4	-2	2



	V	T	E	E	R	D	A	F
L				6	4	0	0	2
T	10	12	9	9	6	4	3	-2
S	8	10	9	9	7	4	3	-3
H	6	7	9	8	$9_{(max)}$	5	1	-2
E	2	4	8	8	3	7	2	-5
A	2	3	2	2	0	2	4	-4
L	2	-2	-3	-3	-3	-4	-2	2

ETAPE 2: CALCUL DE LA MATRICE TRANSFORMÉE

On applique l'algorithme de Needleman et Wunch :

	V	T	E	E	R	D	A	F
L	14	7	6	6	4	0	0	2
T	10	12	9	9	6	4	3	-2
S	8	10	9	9	7	4	3	-3
H	6	7	9	8	9	5	1	-2
E	2	4	8	8	3	7	2	-5
A	2	3	2	2	0	2	4	-4
L	2	-2	-3	-3	-3	-4	-2	2

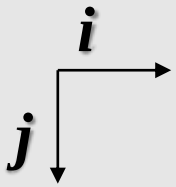
ETAPE 3: TRAÇAGE DU PARCOURS DE LA MATRICE TRANSFORMÉE

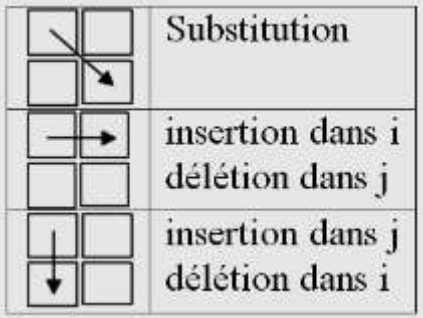
✓ Le parcours s'effectue du plus élevé **en haut à gauche** (ici S = 14) score vers le plus petit. Si les trois cases ont des valeurs de scores égales, alors le chemin vers la diagonale est favorisé :

Début →

	V	T	E	E	R	D	A	F
L	**14**	7	6	6	4	0	0	2
T	10	**12**	9	9	6	4	3	-2
S	8	**10**	9	9	7	4	3	-3
H	6	7	**9**	8	9	5	1	-2
E	2	4	8	**8**	**3**	**7**	2	-5
A	2	3	2	2	0	2	**4**	-4
L	2	-2	-3	-3	-3	-4	-2	**2**

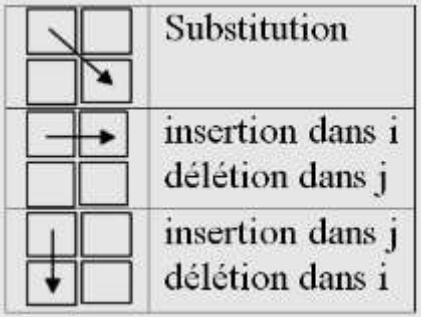
← Fin





ETAPE 4: CONSTRUCTION DE L'ALIGNEMENT ET CALCUL DU SCORE DE SIMILARITÉ

	V	T	E	E	R	D	A	F
L	14	7	6	6	4	0	0	2
T	10	12	9	9	6	4	3	-2
S	8	10	9	9	7	4	3	-3
H	6	7	9	8	9	5	1	-2
E	2	4	8	8	3	7	2	-5
A	2	3	2	2	0	2	4	-4
L	2	-2	-3	-3	-3	-4	-2	2

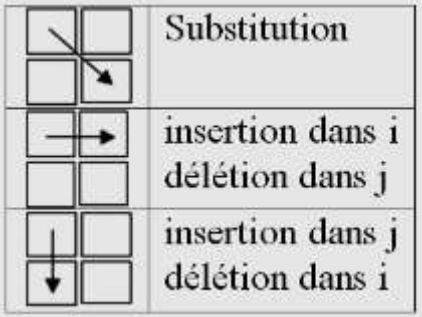


Séquence S_1	**V**	**T**	**—**	**E**	**E**	**R**	**D**	**A**	**F**
	*	|		*	|			|	*
Séquence S_2	**L**	**T**	**S**	**H**	**E**	**—**	**—**	**A**	**L**

- ✓ Il y a trois identités : T-T, E-E et A-A et trois similarités (substitutions): V-L, E-H et F-L
- ✓ On peut supposer que la valine a été substituée en leucine dans la 2ème séquence (ou *vis versa*) par besoin d'adaptation de l'organisme à partir duquel a été isolée cette séquence. Le même raisonnement concernera les substitutions E-H et F-L.

ETAPE 4: CONSTRUCTION DE L'ALIGNEMENT ET CALCUL DU SCORE DE SIMILARITÉ

	V	T	E	E	R	D	A	F
L	14	7	6	6	4	0	0	2
T	10	12	9	9	6	4	3	-2
S	8	10	9	9	7	4	3	-3
H	6	7	9	8	9	5	1	-2
E	2	4	8	8	3	7	2	-5
A	2	3	2	2	0	2	4	-4
L	2	-2	-3	-3	-3	-4	-2	2



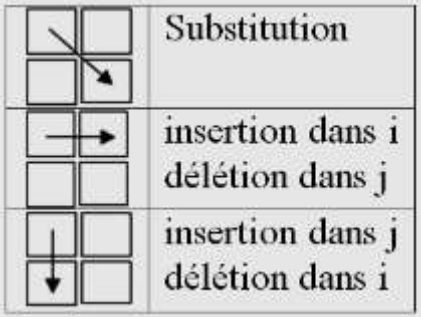
Séquence S_1	**V**	**T**	**—**	**E**	**E**	**R**	**D**	**A**	**F**
	*	|		*	|			|	*
Séquence S_2	**L**	**T**	**S**	**H**	**E**	**—**	**—**	**A**	**L**

✓ Le score global de cet alignement est la somme des scores élémentaires d'identité (sur la matrice transformée) moins la distance.

$$S(S_1, S_2) = (12 + 8 + 4) - (14 + 10 + 9 + 3 + 7 + 2) = -21$$

ETAPE 4: CONSTRUCTION DE L'ALIGNEMENT ET CALCUL DU SCORE DE SIMILARITÉ

	V	T	E	E	R	D	A	F
L	14	7	6	6	4	0	0	2
T	10	12	9	9	6	4	3	-2
S	8	10	9	9	7	4	3	-3
H	6	7	9	8	9	5	1	-2
E	2	4	8	8	3	7	2	-5
A	2	3	2	2	0	2	4	-4
L	2	-2	-3	-3	-3	-4	-2	2



Séquence S_1	**V**	**T**	**—**	**E**	**E**	**R**	**D**	**A**	**F**
	*	|		*	|			|	*
Séquence S_2	**L**	**T**	**S**	**H**	**E**	**—**	**—**	**A**	**L**

✓ Le pourcentage d'identité entre S_1 et S_2 est égal au rapport entre le nombre d'identités ponctuelles entre les deux séquences (3) et la longueur de l'alignement (9) :

$$\%id = (3/9)*100 = 33,33\%$$

ALIGNEMENT DE DEUX SÉQUENCES

ACTIVITÉ 2

Réalisez un alignement optimal des deux séquences protéiques précédentes (S1 = VTEERDAF et S2 = LTSHEAL) en i) appliquant les étapes de l'algorithme de Needleman et Wunch, ii) en utilisant la matrice de substitution BLOSUM62 et iii) en calculant le pourcentage d'identité et le score de similarité de l'alignement puis iv) comparez le résultat obtenu avec celui de la PAM250.

L'activité est à réaliser en groupes de 2 à 4 étudiants.

A remettre sur un support papier ou en fichier numérique au plus tard le mardi 17/03/2020.

- ✓ L'algorithme de **Smith et Waterman** est une méthode d'alignement local ;
- ✓ Plus sensible que celles directement inspirées de Needleman et Wunsch surtout lorsque les séquences à comparer sont **inconnues** ou de **longueurs différentes**.
- ✓ Si les régions trouvées entre les deux séquences recouvrent la totalité de celles-ci, alors on peut considérer l'alignement local comme étant un alignement global.
- ✓ Similaire à l'algorithme de Needleman et Wunsch, à la différence que n'importe quelle case de la matrice de comparaison peut être considérée comme point de départ pour le calcul des scores.

$$S(i,j) = Max \begin{cases} S(i-1,j-1) + s(i,j) \\ S(i-1,j) + p \\ S(i,j-1) + p \\ 0 \end{cases}$$

- ✓ Pour un alignement local entre les deux séquences **(S_1, S_2)** :

S_1 = GGTTGACTA m = 9 et S_2 = TGTTACGG n = 8

- ✓ Les mêmes étapes citées pour la méthode de Needleman et wunsch sont appliquées dans cet algorithme :

ETAPE 1: MATRICE INITIALE

	A	T	G	C
A	3	-3	-3	-3
T	-3	3	-3	-3
G	-3	-3	3	-3
C	-3	-3	-3	3

match = 3 ; mismatch = -3 p = -2

ETAPE 2: INITIALISATION DE LA MATRICE

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

ETAPE 2: CALCUL DE LA MATRICE TRANSFORMÉE

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

		G
	0	0
T	0	0

$$S = 0 = Max \begin{cases} 0 + (-3) \\ 0 + (-2) \\ 0 + (-2) \\ 0 \leftarrow \text{Score maximal} \end{cases}$$

$$S(i,j) = Max \begin{cases} S(i-1, j-1) + s(i,j) \\ S(i-1, j) + p \\ S(i, j-1) + p \\ 0 \end{cases}$$

	A	T	G	C
A	3	-3	-3	-3
T	-3	3	-3	-3
G	-3	-3	3	-3
C	-3	-3	-3	3

$p = -2$

ETAPE 2: CALCUL DE LA MATRICE TRANSFORMÉE

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

		G
	0	0
T	0	0

$$S = 0 = Max \begin{cases} 0 + (-3) \\ 0 + (-2) \\ 0 + (-2) \\ 0 \leftarrow \text{Score maximal} \end{cases}$$

$$S(i,j) = Max \begin{cases} S(i-1, j-1) + s(i,j) \\ S(i-1, j) + p \\ S(i, j-1) + p \\ 0 \end{cases}$$

	A	T	G	C
A	3	-3	-3	-3
T	-3	3	-3	-3
G	-3	-3	3	-3
C	-3	-3	-3	3

$p = -2$

ETAPE 2: CALCUL DE LA MATRICE TRANSFORMÉE

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

		T
	0	0
T	0	3

$$S = 3 = Max \begin{cases} 0 + (3) \leftarrow \text{Score maximal} \\ 0 + (-2) \\ 0 + (-2) \\ 0 \end{cases}$$

$$S(i,j) = Max \begin{cases} S(i-1, j-1) + s(i,j) \\ S(i-1, j) + p \\ S(i, j-1) + p \\ 0 \end{cases}$$

	A	T	G	C
A	3	-3	-3	-3
T	-3	3	-3	-3
G	-3	-3	3	-3
C	-3	-3	-3	3

$p = -2$

ETAPE 2: CALCUL DE LA MATRICE TRANSFORMÉE

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

		T
	0	0
T	3	1

$$S = 1 = Max \begin{cases} 0 + (-3) \\ 0 + (-2) \\ 3 + (-2) \leftarrow \text{Score maximal} \\ 0 \end{cases}$$

$$S(i,j) = Max \begin{cases} S(i-1, j-1) + s(i,j) \\ S(i-1, j) + p \\ S(i, j-1) + p \\ 0 \end{cases}$$

	A	T	G	C
A	3	-3	-3	-3
T	-3	3	-3	-3
G	-3	-3	3	-3
C	-3	-3	-3	3

$p = -2$

ETAPE 2: CALCUL DE LA MATRICE TRANSFORMÉE

		G	G	T	T	G	A	C	T	A
	0	0	0	0	0	0	0	0	0	0
T	0	0	0	3	3	1	0	0	3	1
G	0	3	3	1	1	6	4	2	1	0
T	0	1	1	6	4	4	3	1	5	3
T	0	0	0	4	9	7	5	3	4	2
A	0	0	0	2	7	6	10	8	6	7
C	0	0	0	0	5	4	8	13	11	9
G	0	3	3	1	3	8	6	11	10	8
G	0	3	6	4	2	6	5	9	8	7

		A
	10	8
G	8	7

$$S = 7 = Max \begin{cases} 10 + (-3) \\ 8 + (-2) \\ 8 + (-2) \leftarrow \text{Score maximal} \\ 0 \end{cases}$$

$$S(i,j) = Max \begin{cases} S(i-1, j-1) + s(i,j) \\ S(i-1, j) + p \\ S(i, j-1) + p \\ 0 \end{cases}$$

	A	T	G	C
A	3	-3	-3	-3
T	-3	3	-3	-3
G	-3	-3	3	-3
C	-3	-3	-3	3

$p = -2$

ETAPE 3 : TRAÇAGE DU PARCOURS DE LA MATRICE TRANSFORMÉE

Pour tracer le(s) parcours, on recherche la (les) case(s) contenant **le score maximal** de la matrice (dans ce cas il y a un seul $S_{max} = 13$), **quelque soit** la position de la case. Ensuite, on trace le parcours **jusqu'à** une case contenant la valeur **0**.

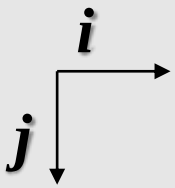
		G	G	T	T	G	A	C	T	A
	0	0	0	0	0	0	0	0	0	0
T	0	0	0	3	3	1	0	0	3	1
G	0	3	3	1	1	6	4	2	1	0
T	0	1	1	6	4	4	3	1	5	3
T	0	0	0	4	9	7	5	3	4	2
A	0	0	0	2	7	6	10	8	6	7
C	0	0	0	0	5	4	8	13	11	9
G	0	3	3	1	3	8	6	11	10	8
G	0	3	6	4	2	6	5	9	8	7

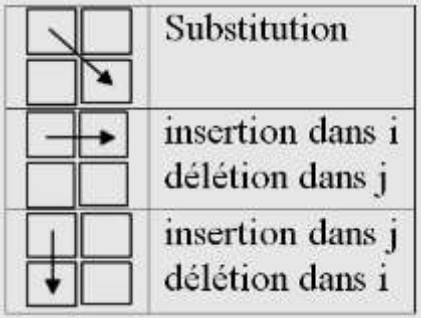
Fin

Début

ETAPE 4: CONSTRUCTION DE L'ALIGNEMENT ET CALCUL DU SCORE DE SIMILARITÉ

	2	3	4	5	6	7
S_1	G	T	T	G	A	C
	|	|	|		|	|
S_2	G	T	T	—	A	C
	2	3	4	-	5	6





Position			1	2	3	4	5	6	7	8	9
Position			G	G	T	T	G	A	C	T	A
		0	0	0	0	0	0	0	0	0	0
1	T	0	**0**	0	3	3	1	0	0	3	1
2	G	0	3	**3**	1	1	6	4	2	1	0
3	T	0	1	1	**6**	4	4	3	1	5	3
4	T	0	0	0	4	**9**	**7**	5	3	4	2
5	A	0	0	0	2	7	6	**10**	8	6	7
6	C	0	0	0	0	5	4	8	**13**	11	9
7	G	0	3	3	1	3	8	6	11	10	8
8	G	0	3	6	4	2	6	5	9	8	7

ETAPE 4: CONSTRUCTION DE L'ALIGNEMENT ET CALCUL DU SCORE DE SIMILARITÉ

	2	3	4	5	6	7
S_1	G	T	T	G	A	C
	|	|	|		|	|
S_2	G	T	T	—	A	C
	2	3	4	-	5	6

- ✓ Le score local de cet alignement est égal à la somme des matchs moins la distance (nous avions choisi : match = 3 ; mismatch = -3 ; p = -2) :

$$S(S_1, S_2) = (5*3) - (1*-2) = 13$$

- ✓ Le pourcentage d'identité locale entre S_1 et S_2 est égal au rapport entre le nombre d'identités ponctuelles entre les deux séquences (5) et la longueur de l'alignement local (6) :

$$\%id = (5/6)*100 = 83{,}33\%$$

ALGORITHME DE SMITH ET WATERMAN : EXEMPLE PRATIQUE

		G	G	T	T	G	A	C	T	A
	0	0	0	0	0	0	0	0	0	0
T	0	0	0	3	3	1	0	0	3	1
G	0	3	3	1	1	6	4	2	1	0
T	0	1	1	6	4	4	3	1	5	3
T	0	0	0	4	9	7	5	3	4	2
A	0	0	0	2	7	6	13	2	6	7
G	0	0	0	0	5	4	1	0	11	9
T	0	3	3	1	3	8	6	11	10	8
A	0	3	6	4	2	6	5	9	8	13

G	T	T	G	A
|	|	|		|
G	T	T	—	A

Alignement local 1

T	A
|	|
T	A

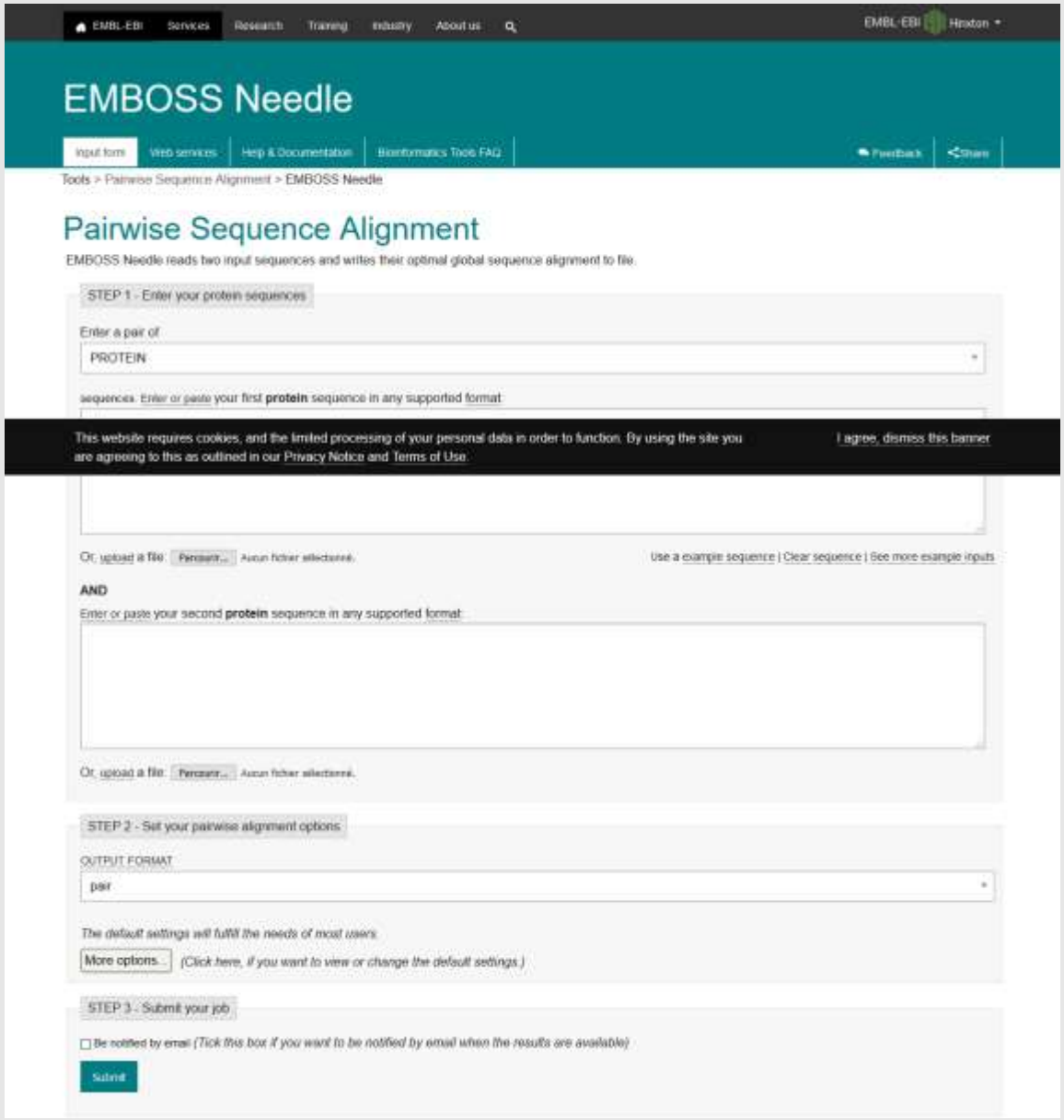
Alignement local 2

EXEMPLE DE L'APPLICATION DE L'ALGORITHME DE SMITH ET WATERMAN QUI DONNE DEUX ALIGNEMENTS LOCAUX EN MÊME

- ✓ L'alignement optimal (au sens programmation dynamique) n'est pas obligatoirement celui qui est pertinent au niveau biologique ;
- ✓ Un alignement est sensible aux changements de paramètres ;
- ✓ D'un point de vue biologique, l'alignement le plus pertinent est celui qui retrace le déroulement évolutif le plus probable. En effet, on sait que des variations importantes dans les probabilités de mutations des positions et des résidus se produisent.

ALIGNEMENT DE DEUX SÉQUENCES

EXEMPLES DE LOGICIELS D'ALIGNEMENT PAR PAIRE

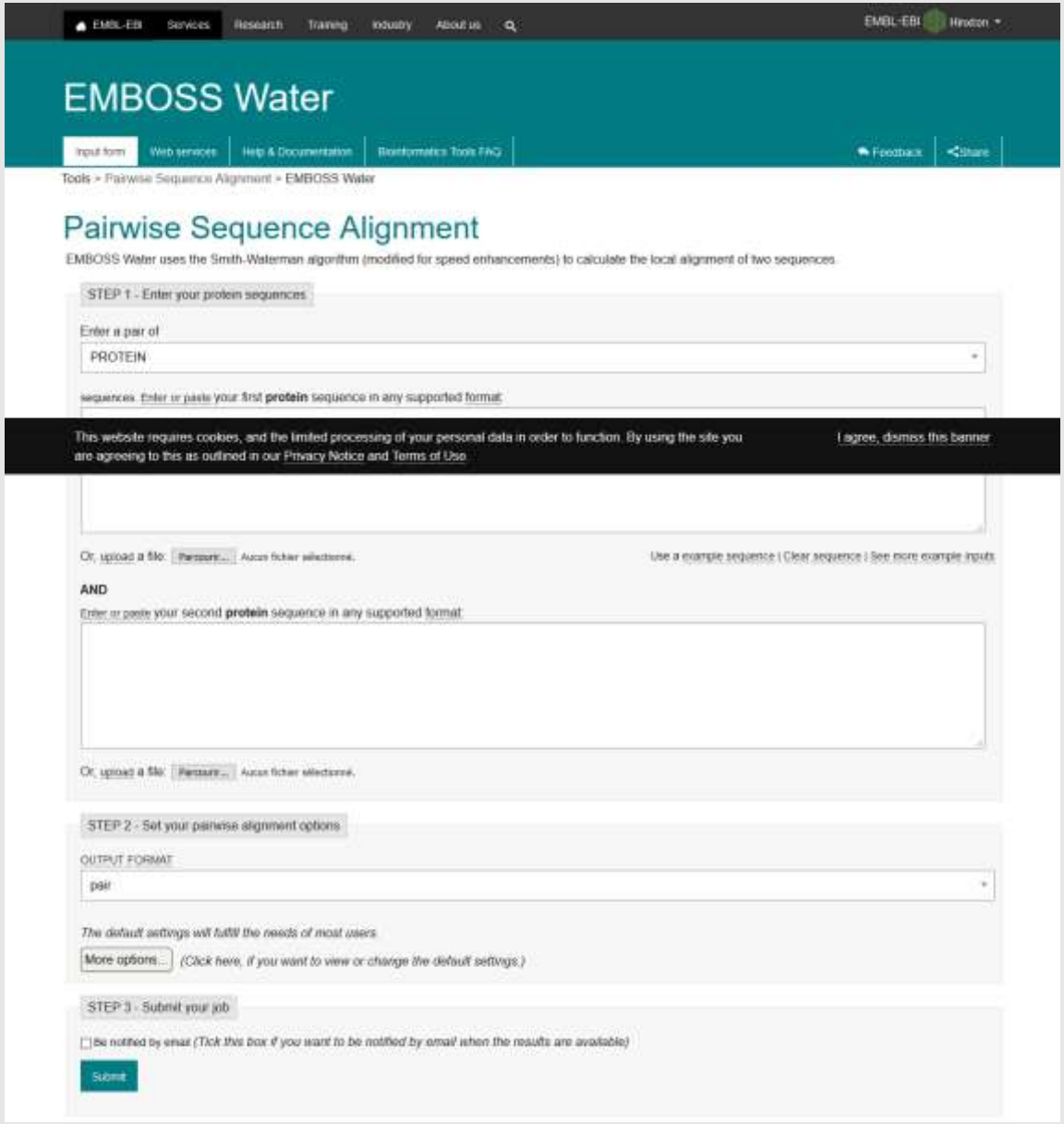


EMBOSS NEEDLE (https://www.ebi.ac.uk/Tools/psa/emboss_needle/)

POUR RÉALISER UN ALIGNEMENT GLOBAL

ALIGNEMENT DE DEUX SÉQUENCES

EXEMPLES DE LOGICIELS D'ALIGNEMENT PAR PAIRE

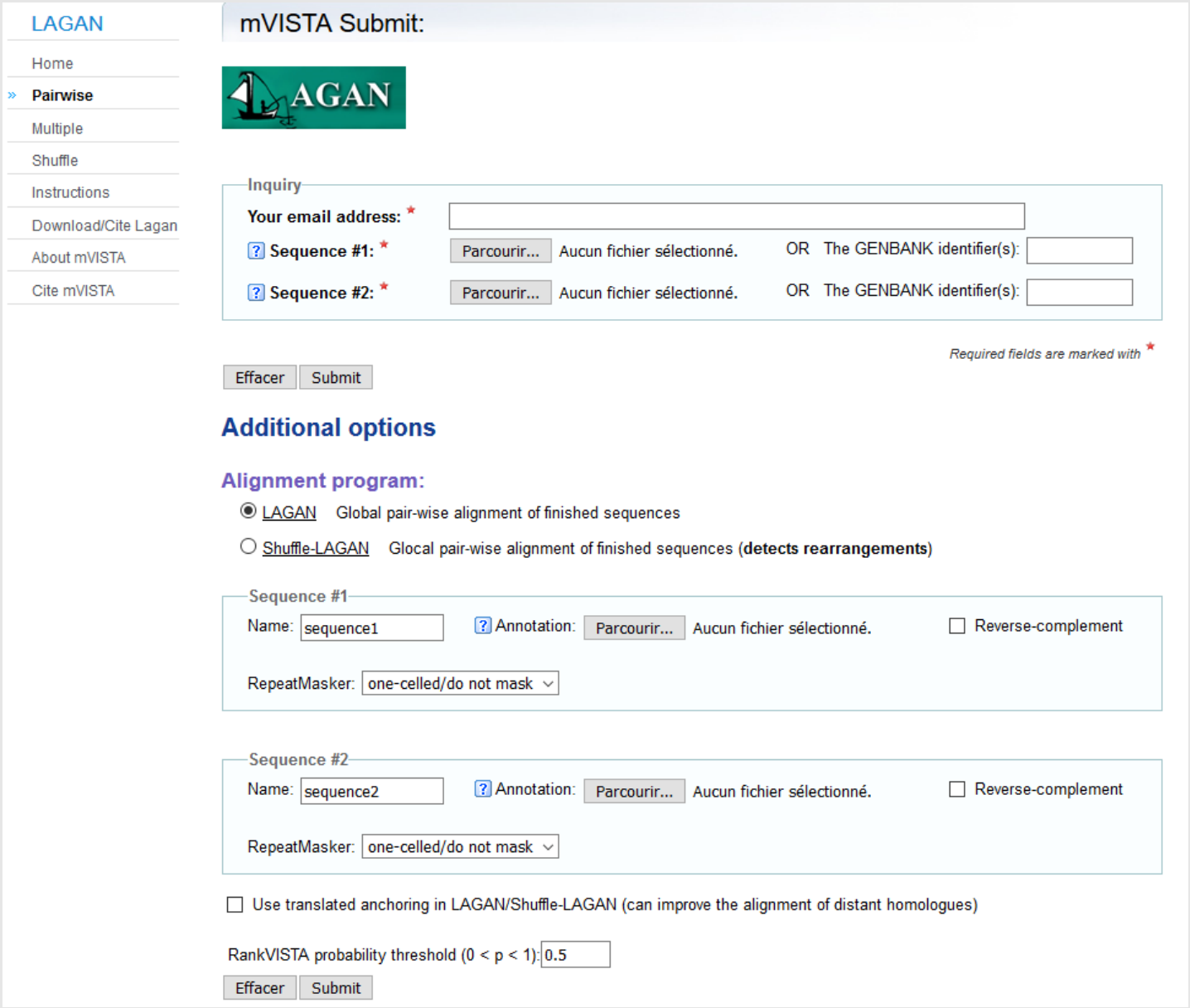


EMBOSS WATER (https://www.ebi.ac.uk/Tools/psa/emboss_water/)

POUR RÉALISER UN ALIGNEMENT LOCAL

ALIGNEMENT DE DEUX SÉQUENCES

EXEMPLES DE LOGICIELS D'ALIGNEMENT PAR PAIRE



LAGAN (http://genome.lbl.gov/cgi-bin/VistaInput?align_pgm=lagan&num_seqs=2)

POUR RÉALISER UN ALIGNEMENT GLOBAL DE SÉQUENCES GÉNOMIQUES

Les outils d'alignement par paire **EMBOSS (https://www.ebi.ac.uk/Tools/emboss/)**. (Démonstration visionnable ici : **https://youtu.be/rJGMK9Iwo8A**)