Alignement optimal ?

```
AACT--GGTAACCG
AGCTACGGT--CCG
```

➡ Calcul d'un score

Le score de l'alignement doit prendre en compte toutes les positions alignées : identités, substitutions et indels. Chacun de ces évènements va recevoir un poids, appelé score élémentaire s_e. Le score de l'alignement correspondra à la somme des scores élémentaires correspondant aux positions alignées.

$$S = \sum_{i=1}^{l} Se(i)$$

Où *l* est le nombre de positions alignées

exemple: *l* = 14

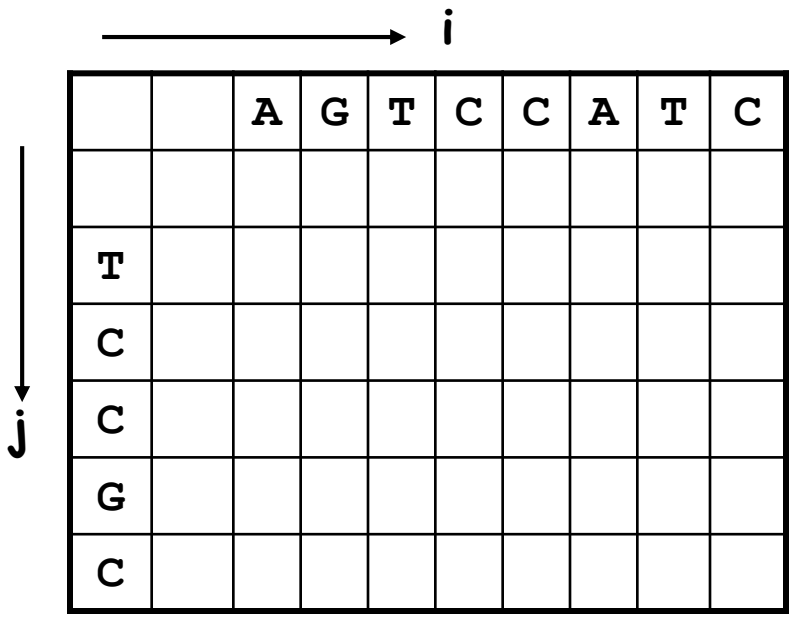
s_e identité = +2
s_e substitution = -1
s_e indels = -2

➡ S = 9

Prenons comme exemple deux séquences X et Y de longueur M et N :

X = AGTCCATC M=8
Y = TCCGC N=5

Matrice de programmation dynamique :



Le score optimal sera calculé récursivement. Le score calculé pour la cellule (i,j) correspondra au meilleur alignement des résidus $x_1 \ldots\ldots x_i$ avec les résidus $y_1 \ldots\ldots y_j$

Score d'homologie:

$$s(i,j) = \max \begin{cases} s(i-1, j-1) + w(x_i, y_j) \\ s(i-1, j) + \delta \\ s(i, j-1) + \delta \end{cases}$$

Où :

- $w(x_i, y_j)$ est la valeur dans le système de score correspondant soit à l'identité soit à la substitution (mismatch)
- d est le poids de l'insertion/délétion (indel)

matrice PAM250

	C	S	T	P	A	G	N	D	E	Q	H	R	K	M	I	L	V	F	Y	W
C	12																			
S	0	2																		
T	-2	1	3																	
P	-3	1	0	6																
A	-2	1	1	1	2															
G	-3	1	0	-1	1	5														
N	-4	1	0	-1	0	0	2													
D	-5	0	0	-1	0	1	2	4												
E	-5	0	0	-1	0	0	1	3	4											
Q	-5	-1	-1	0	0	-1	1	2	2	4										
H	-3	-1	-1	0	-1	-2	2	1	1	3	6									
R	-4	0	-1	0	-2	-3	0	-1	-1	1	2	6								
K	-5	0	0	-1	-1	-2	1	0	0	1	0	3	5							
M	-5	-2	-1	-2	-1	-3	-2	-3	-2	-1	-2	0	0	6						
I	-2	-1	0	-2	-1	-3	-2	-2	-2	-2	-2	-2	-2	2	5					
L	-6	-3	-2	-3	-2	-4	-3	-4	-3	-2	-2	-3	-3	4	2	6				
V	-2	-1	0	-1	0	-1	-2	-2	-2	-2	-2	-2	-2	2	4	2	4			
F	-4	-3	-3	-5	-4	-5	-3	-6	-5	-5	-2	-4	-5	0	1	2	-1	9		
Y	0	-3	-3	-5	-3	-5	-2	-4	-4	-4	0	-4	-4	-2	-1	-1	-2	7	10	
W	-8	-2	-5	-6	-6	-7	-4	-7	-7	-5	-3	2	-3	-4	-5	-2	-6	0	0	17

S – small hydrophilic

N- acid, acid amide, hydrophylic

H - basic

V – small hydrophobic

F- aromatic

Matrices PAM

PAM : Point/Percent Accepted Mutation (Dayhoff, 1978)

Construction :

- 71 familles de protéines
- alignements multiples (global)
- reconstruction des arbres phylogénétiques et inférences des séquences ancêtres (1 572 mutations)
- pour chaque a.a. on compte le nombre de fois où il a été substitué

  Exemple : pour *Val* combien de fois il est resté inchangé, et combien de fois il a été substitué par *Ala*, par *Arg*, …

- Cumul des mutations « acceptées » par la sélection naturelle au sein des différentes familles
- Calcul de la mutabilité des a.a. (propension d'un a.a. à être remplacé, f(Val->X))
- Calcul de la matrice de probabilités de mutation à partir des 2 étapes précédentes
  - Pour chaque a.a., ex: Val -> Ala = mutabilité(Val) * cumul(Val->Ala) / nb(Val)
- Calcul de la matrice Lods (Log odd ratios) : PAM1

```
                                  1                   2                   3
                 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9…
KAPPA
 1 HUMAN EU      T - V A A P S V F I F P P S D E Q - L K S - G T A S V V C L L N N F Y P - R E…
 2 MOUSE MOPC 21 A - D A A P T V S I F P P S S E Q - L T S - G G A S V V C F L N N F Y P - K D
 3 QAT S211      A - N A A P T V S I F P P S T Z Z - L A T - G G A S V V C L M N K.F Y P – R D
 4 84 RA881T 4135 D - P V A P T V L I F P P A A D Q - V A T - G T V T I V C V A N K Y F P - - D
 B9 RA881T       D P P I A P T V L L F P P S A D Q - L T T - Z T V T I V C V A N K F R P - D D

LAMBDA
 6 HUMAN SH      Q P K A A P S V T L F P P S S E E - L Q A - N K A T L V C L I S D F Y P - G A
 7 PIG           Q P K A A P T V N L F P P S S E E - L G T - N K A T L V C L I S D F Y P - G A
 8 MOUSE MOPC 104E Q P K S S P S V T L F P P S S E E - L T E - N K A T L V C T I T O F Y P - G V
 9 MOUSE MOPC 315 Q P K S T P T L T V F P P S S E E - L K E - N K.A T L V C L I S N F S P - G S
 …
```

Matrices PAM

PAM : Point/Percent Accepted Mutation (Dayhoff, 1978)

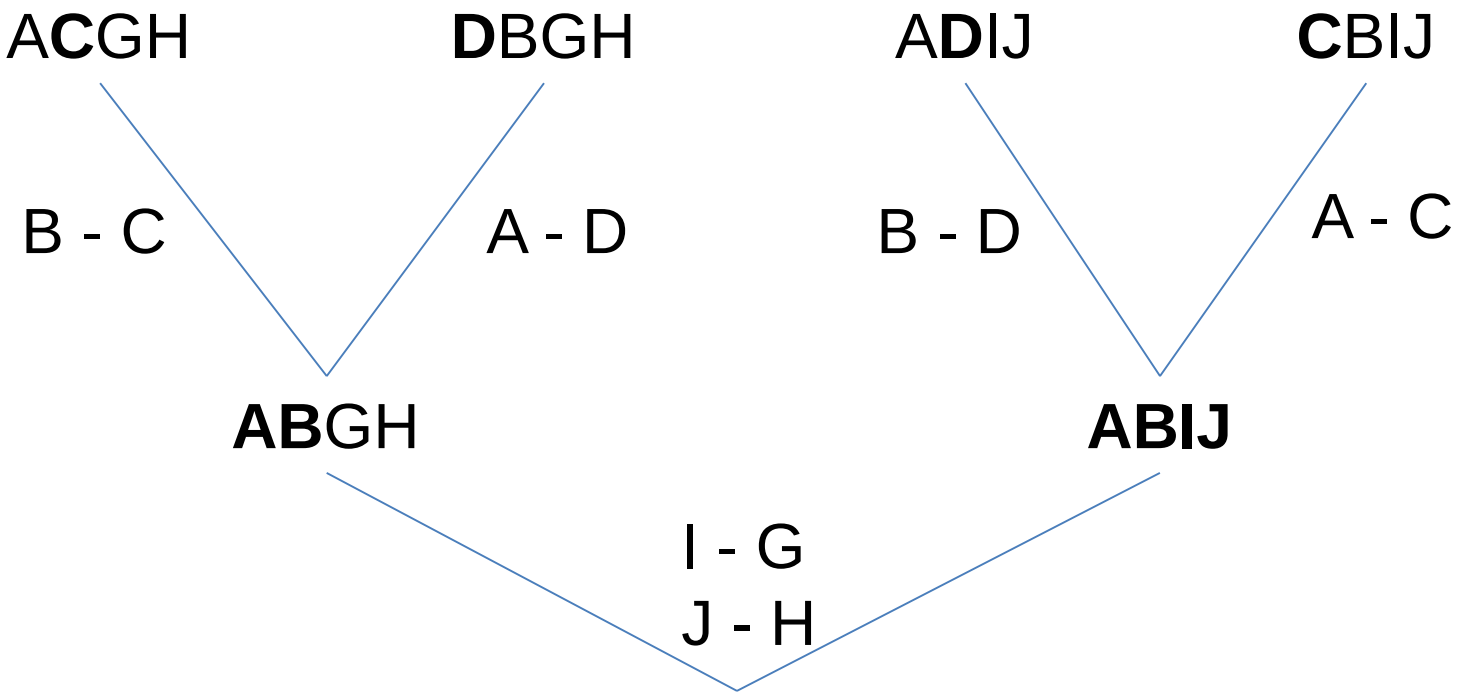
Construction :

- 71 familles de protéines
- alignements multiples (global)
- reconstruction des arbres phylogénétiques et inférences des séquences ancêtres (1 572 mutations)
- pour chaque a.a. on compte le nombre de fois où il a été substitué

  Exemple : pour *Val* combien de fois il est resté inchangé, et combien de fois il a été substitué par *Ala*, par *Arg*, …

- Cumul des mutations « acceptées » par la sélection naturelle au sein des différentes familles
- Calcul de la mutabilité des a.a. (propension d'un a.a. à être remplacé, f(Val->X))
- Calcul de la matrice de probabilités de mutation à partir des 2 étapes précédentes
  - Pour chaque a.a., ex: Val -> Ala = mutabilité(Val) * cumul(Val->Ala) / nb(Val)
- Calcul de la matrice Lods (Log odd ratios) : PAM1

Arbre phylogénétique



Matrice des mutations acceptées

	A	B	C	D	G	H	I	J
A			1	1				
B			1	1				
C	1	1						
D	1	1						
G							1	
H								1
I					1			
J						1		

Matrices PAM

PAM : Point/Percent Accepted Mutation (Dayhoff, 1978)

Construction :

- 71 familles de protéines
- alignements multiples (global)
- reconstruction des arbres phylogénétiques et inférences des séquences ancêtres (1 572 mutations)
- pour chaque a.a. on compte le nombre de fois où il a été substitué

  Exemple : pour *Val* combien de fois il est resté inchangé, et combien de fois il a été substitué par *Ala*, par *Arg*, …

- Cumul des mutations « acceptées » par la sélection naturelle au sein des différentes familles
- Calcul de la mutabilité des a.a. (propension d'un a.a. à être remplacé, f(Val->X))
- Calcul de la matrice de probabilités de mutation à partir des 2 étapes précédentes
  - Pour chaque a.a., ex: Val -> Ala = mutabilité(Val) * cumul(Val->Ala) / nb(Val)
- Calcul de la matrice Lods (Log odd ratios) : PAM1

Matrice de cumul des mutations acceptées (x10)

	ala	arg	asn	asp	cys	gln	glu	gly	his	ile	leu	lys	met	phe	pro	ser	thr	trp	tyr	val
A																				
R	30																			
N	109	17																		
D	154	0	532																	
C	33	10	0	0																
Q	93	120	50	76	0															
E	266	0	94	831	0	422														
G	579	10	156	162	10	30	112													
H	21	103	226	43	10	243	23	10												
I	66	30	36	13	17	8	35	0	3											
L	95	17	37	0	0	75	15	17	40	253										
K	57	477	322	85	0	147	104	60	23	43	39									
M	29	17	0	0	0	20	7	7	0	57	207	90								
F	20	7	7	0	0	0	0	17	20	90	167	0	17							
P	345	67	27	10	10	93	40	49	50	7	43	43	4	7						
S	772	137	432	98	117	47	86	450	26	20	32	168	20	40	269					
T	590	20	169	57	10	37	31	50	14	129	52	200	28	10	73	696				
W	0	27	3	0	0	0	0	0	3	0	13	0	0	10	0	17	0			
Y	20	3	36	0	30	0	10	0	40	13	23	10	0	260	0	22	23	6		
V	365	20	13	17	33	27	37	97	30	661	303	17	77	10	50	43	186	0	17	

Matrices PAM

PAM : Point/Percent Accepted Mutation (Dayhoff, 1978)

Construction :

- 71 familles de protéines
- alignements multiples (global)
- reconstruction des arbres phylogénétiques et inférences des séquences ancêtres (1 572 mutations)
- pour chaque a.a. on compte le nombre de fois où il a été substitué

  Exemple : pour *Val* combien de fois il est resté inchangé, et combien de fois il a été substitué par *Ala*, par *Arg*, …

- Cumul des mutations « acceptées » par la sélection naturelle au sein des différentes familles
- Calcul de la mutabilité des a.a. (propension d'un a.a. à être remplacé, f(Val->X))
- Calcul de la matrice de probabilités de mutation à partir des 2 étapes précédentes
  - Pour chaque a.a., ex: Val -> Ala = mutabilité(Val) * cumul(Val->Ala) / nb(Val)
- Calcul de la matrice Lods (Log odd ratios) : PAM1

Séquences alignées

A D A

A D B

Acides aminés	A	B	D
Changements observés	1	1	0
Occurrences	3	1	2
Mutabilité	.33	1	0

Mutabilité (Dayhoff, 1978)

```
Ser 149
Met 122
Asn 111
Ile 110
Glu 102
Ala 100
Gln 98
Asp 90
Thr 90
Gap 84
Val 80
Lys 57
Pro 56
His 50
Gly 48
Phe 45
Arg 44
Leu 38
Tyr 34
Cys 27
Trp 22
```

Positionnée à 100 arbitrairement

Matrices PAM

PAM : Point/Percent Accepted Mutation (Dayhoff, 1978)

Construction :

- 71 familles de protéines
- alignements multiples (global)
- reconstruction des arbres phylogénétiques et inférences des séquences ancêtres (1 572 mutations)
- pour chaque a.a. on compte le nombre de fois où il a été substitué

  Exemple : pour *Val* combien de fois il est resté inchangé, et combien de fois il a été substitué par *Ala*, par *Arg*, …

- Cumul des mutations « acceptées » par la sélection naturelle au sein des différentes familles
- Calcul de la mutabilité des a.a. (propension d'un a.a. à être remplacé, f(Val->X))
- Calcul de la matrice de probabilités de mutation à partir des 2 étapes précédentes
  - Pour chaque a.a., ex: Val -> Ala = mutabilité(Val) * cumul(Val->Ala) / nb(Val)

  m_j : mutabilité de l'a.a. j

  A_{ij} : nombre de fois que l'a.a. j a été remplacé par l'a.a. i

  λ : paramètre d'ajustement pour avoir 1 mutation acceptée pour 100 résidus

$$M_{i,j} = \lambda \frac{m_j A_{ij}}{\sum_i A_{ij}}$$

- Calcul de la matrice Lods (Log odd ratios) : $PAM1_{i,j} = \log \frac{M_{ij}}{f_i}$

Matrices PAM

Calcul de la matrice Lods (Log odd ratios) :

Permet de faire la somme des scores élémentaires pour un alignement plutôt que le produit des probabilités : log (a*b) = log a + log b

$$w_{i,j} = \log \frac{q_{ij}}{p_{ij}}$$

où:
q_{ij} est la fréquence observée de substitution de l'acide aminé *i* en *j*
p_{ij} est la fréquence théorique de substitution de l'acide aminé *i* en *j*

PAM1 : Normalisée pour avoir 1 mutation acceptée pour 100 a.a.

→ Temps qu'il faut pour qu'une mutation se fixe dans la population
= Distance évolutive conceptuelle : 1 PAM

Hypothèse : les probabilités de mutations sont indépendantes

PAM2 = PAM1 x PAM1
Matrice pour une distance évolutive de 2 PAM

De même, PAM40 = $PAM1^{40}$, PAM120 = $PAM1^{120}$, PAM250 = $PAM1^{250}$

Alignement de deux séquences protéiques
Matrices de substitution

La matrice PAM250

	C	S	T	P	A	G	N	D	E	Q	H	R	K	M	I	L	V	F	Y	W
C	12																			
S	0	2																		
T	-2	1	3																	
P	-3	1	0	6																
A	-2	1	1	1	2															
G	-3	1	0	-1	1	5														
N	-4	1	0	-1	0	0	2													
D	-5	0	0	-1	0	1	2	4												
E	-5	0	0	-1	0	0	1	3	4											
Q	-5	-1	-1	0	0	-1	1	2	2	4										
H	-3	-1	-1	0	-1	-2	2	1	1	3	6									
R	-4	0	-1	0	-2	-3	0	-1	-1	1	2	6								
K	-5	0	0	-1	-1	-2	1	0	0	1	0	3	5							
M	-5	-2	-1	-2	-1	-3	-2	-3	-2	-1	-2	0	0	6						
I	-2	-1	0	-2	-1	-3	-2	-2	-2	-2	-2	-2	-2	2	5					
L	-6	-3	-2	-3	-2	-4	-3	-4	-3	-2	-2	-3	-3	4	2	6				
V	-2	-1	0	-1	0	-1	-2	-2	-2	-2	-2	-2	-2	2	4	2	4			
F	-4	-3	-3	-5	-4	-5	-3	-6	-5	-5	-2	-4	-5	0	1	2	-1	9		
Y	0	-3	-3	-5	-3	-5	-2	-4	-4	-4	0	-4	-4	-2	-1	-1	-2	7	10	
W	-8	-2	-5	-6	-6	-7	-4	-7	-7	-5	-3	2	-3	-4	-5	-2	-6	0	0	17

S – small hydrophilic

N- acid, acid amide, hydrophylic

H - basic

V – small hydrophobic

F- aromatic

Matrices PAM

Remarques :

- Matrice calculée à partir de séquences ayant moins de 15% de divergence
- Biais dans la sélection des protéines (petites protéines globulaires)
- Actualisées : 16 130 séquences appartenant à 2 621 familles de protéines