

# Introduction

**Alignement multiple** = Alignement simultané de plusieurs séquences (Nt ou Prot)

Outil essentiel pour :

- Signatures protéiques
- Homologie avec une famille de protéines particulière
- Structure secondaire ou tertiaire des protéines
- Choix d'amorces pour PCR
- Étape pré-requise pour les analyses d'évolution moléculaire

Alignement Multiple : -> car alignement de 2 séquences **n'est pas transitif**

Programmation dynamique : Solution mathématique optimale -> alignement optimal

- Alignement de 3 séquences maximum
- jusqu'à 8 séquences protéiques d'environ 200aa (en diminuant la demande en mémoire)



Insuffisant

# comparaison de séquences multiples

Pour algorithmes exacts beaucoup de mémoire et de temps de calcul requis.

Algorithme de Needleman-Wunsch

Algorithme de Smith-Waterman

2 Globines => 1 sec

3 Globines => 2 min

4 Globines => 5 hr

5 Globines => 3 semaines

6 Globines => 9 ans

7 Globines => 1000 ans



Heuristiques

# Introduction

Méthodes approchées = Heuristiques

2 méthodes principales :

- détection de zones d'ancrage
- alignement progressif

Au cours de l'évolution : Accumulation **non homogène** de mutations le long d'une séquence (zones fonctionnelles plus contraintes)

Certaines positions de la protéine vont donc être **conservées**, mais cette conservation peut être **diffuse** (qq aa séparés par de grandes zones non similaires)

But : détecter les zones conservées entre plusieurs séquences homologues

# Alignement multiple : ClustalW

Thompson *et al.*, Nucleic Acids Res., 22, 4673-80 (1994)

## Plusieurs étapes :

### Alignement deux à deux de toutes les séquences

- soit algorithme de programmation dynamique (alignement global)  
(matrice de substitution, pondération affine des indels)
- soit méthode d'alignement rapide  
(score =  $\sum$  des mots de longueur k identiques (k-tuples) - pénalité fixe pour chaque indel)
  - k = 1 ou 2 résidus pour des protéines
  - k = 2 à 4 résidus pour des séquences ADN



Construction d'une matrice de distances entre les séquences, calculée à partir des scores des alignements obtenus

# Exemple d'une matrice de distances

Extrait de Nucleic Acids Res., 22, 4673-80 (1994)

Hbb-Human	1	-						
Hbb_Horse	2	0.17	-					
Hba_Human	3	0.59	0.60	-				
Hba_Horse	4	0.59	0.59	0.13	-			
Myg_Phyca	5	0.77	0.77	0.75	0.75	-		
Glb5_Petma	6	0.81	0.82	0.73	0.74	0.80	-	
Lgb2_Luplu	7	0.87	0.86	0.86	0.88	0.93	0.90	-
		1	2	3	4	5	6	

Hbb\_Human : Globine  $\beta$  humaine

Hbb\_Horse : Globine  $\beta$  de cheval

Hba\_Human : Globine  $\alpha$  humaine

Hba\_Horse : Globine  $\alpha$  de cheval

Myg\_Phyca : myoglobine de sperme de baleine

Glb5\_Petma : cyanohémoglobine de lamproie

Lgb2\_Luplu : Leghémoglobine de lupin

# Alignement multiple : ClustalW

Thompson *et al.*, Nucleic Acids Res., 22, 4673-80 (1994)

## Plusieurs étapes :

Alignement deux à deux de toutes les séquences

- soit algorithme de programmation dynamique (alignement global)
- soit méthode d'alignement rapide ( $\Sigma$  des mots de longueur k identiques - pénalité indel)



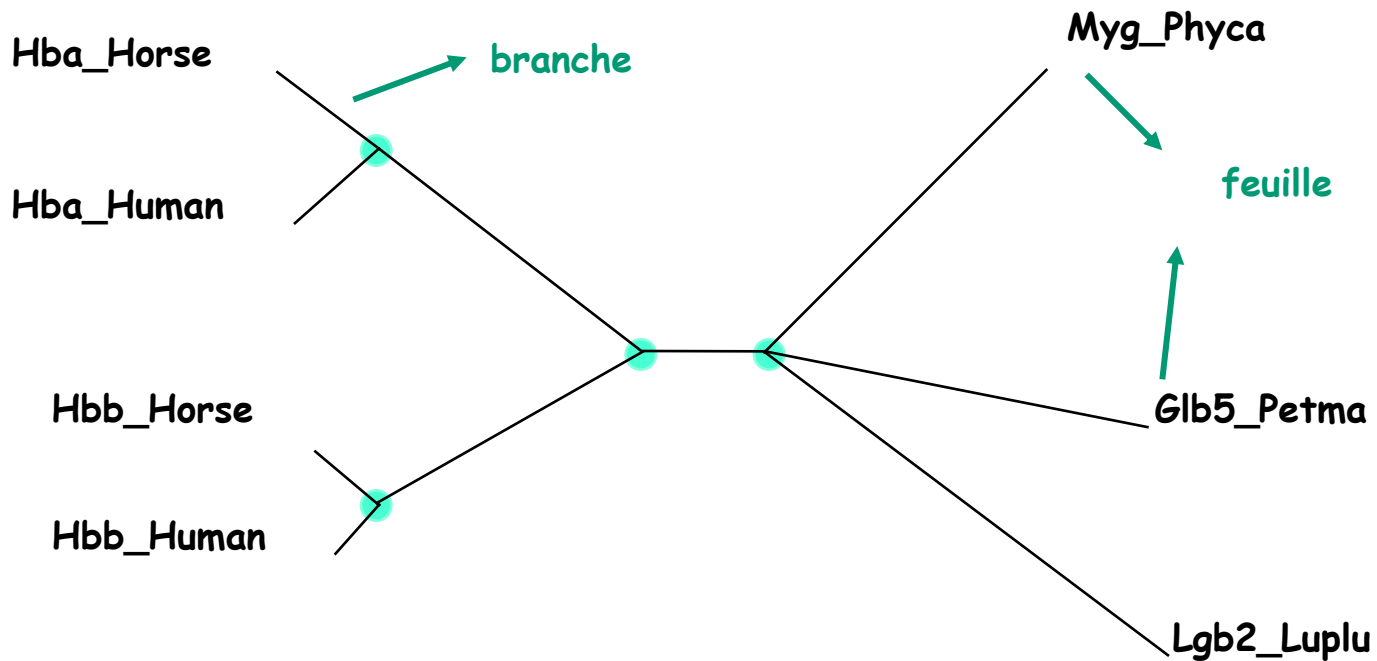
Construction d'une matrice de distances entre les séquences, calculée à partir des scores des alignements obtenus



Construction d'un arbre de parenté entre ces séquences en utilisant la méthode des plus proches voisins (Neighbor Joining Method)

# Arbre sans racine déduit de la matrice de distance

Extrait de Nucleic Acids Res., 22, 4673-80 (1994)



● = nœud interne : ancêtre commun hypothétique

# Alignement multiple : ClustalW

Thompson *et al.*, Nucleic Acids Res., 22, 4673-80 (1994)

## Plusieurs étapes :

Alignement deux à deux de toutes les séquences

- soit algorithme de programmation dynamique (alignement global)
- soit méthode d'alignement rapide ( $\Sigma$  des mots de longueur  $k$  identiques - pénalité indel)



Construction d'une matrice de distances entre les séquences, calculée à partir des scores des alignements obtenus



Construction d'un arbre de parenté entre ces séquences en utilisant la méthode des plus proches voisins (Neighbor Joining Method)

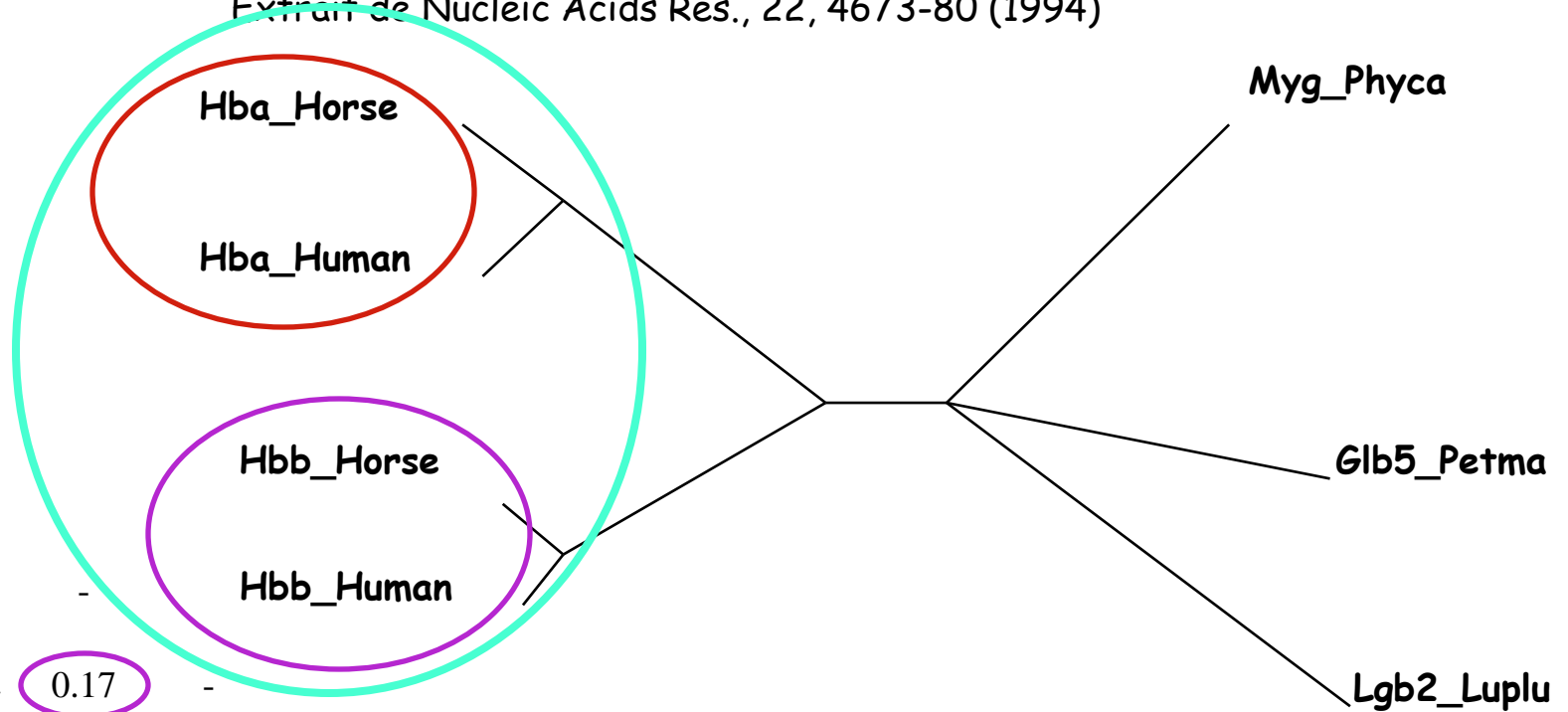


Alignement multiple progressif : ajout des séquences en fonction de leurs distances dans l'arbre (distance  $\nearrow$ , on commence par aligner les deux plus proches)  
 $\Rightarrow$  Alignement par profil



# Arbre sans racine déduit de la matrice de distance

Extrait de Nucleic Acids Res., 22, 4673-80 (1994)



Hbb-Human	1	-						
Hbb_Horse	2	0.17	-					
Hba_Human	3	0.59	0.60	-				
Hba_Horse	4	0.59	0.59	0.13	-			
Myg_Phyca	5	0.77	0.77	0.75	0.75	-		
Glb5_Petma	6	0.81	0.82	0.73	0.74	0.80	-	
Lgb2_Luplu	7	0.87	0.86	0.86	0.88	0.93	0.90	-
		1	2	3	4	5	6	

# Plusieurs problèmes pour cette approche

## Le problème du minimum local

Lié à la nature progressive de la stratégie d'alignement

Pas de correction des erreurs faites dans les premiers alignements, pas de remise en cause des indels :

- Mauvaise structure de l'arbre initial, ordre de branchement incorrect (car arbre établi à partir d'une matrice de distances réalisée à partir d'alignements de séquences 2 à 2 moins fiable que les arbres provenant d'un alignement multiple)
- Même si topologie de l'arbre correcte : un certain pourcentage de résidus mal alignés à chaque étape

# Plusieurs problèmes pour cette approche

## Le problème du choix des paramètres

### Choix de la matrice de substitution :

Une même matrice tout au cours de l'alignement

- fonctionne pour des séquences proches
- problème avec des séquences divergentes

### Choix des pénalités de gaps :

Pénalité de gaps affine (2 valeurs : ouverture d'un nouveau gap et extension d'un gap)

Les indels ne sont pas localisés aléatoirement dans les séquences  
Gaps plus fréquents entre les éléments de structure secondaire (hélice  $\alpha$  et feuillet  $\beta$ ) qu'à l'intérieur de ces structures.

## Pas de pondérations des séquences

Engendre un biais si les distances évolutives entre séquences ne sont pas également représentées

# Améliorations apportées à la dernière version de ClustalW

## Le problème des paramètres

### Matrices de substitution :

4 matrices différentes utilisées au cours de l'alignement multiple en fonction de la distance séparant les séquences à aligner  
(les distances sont mesurées directement à partir de l'arbre initial)

2 séries de matrices proposées : PAM et BLOSUM  
BLOSUM par défaut

% de similarités des séquences :

80%-100% :	PAM20	BLOSUM80
60%-80% :	PAM60	BLOSUM62
40%-60% :	PAM120	BLOSUM45 (30%-60%)
0 - 40% :	PAM350	BLOSUM30 (0 - 30%)

# Améliorations apportées à la dernière version de ClustalW

## Le problème des paramètres

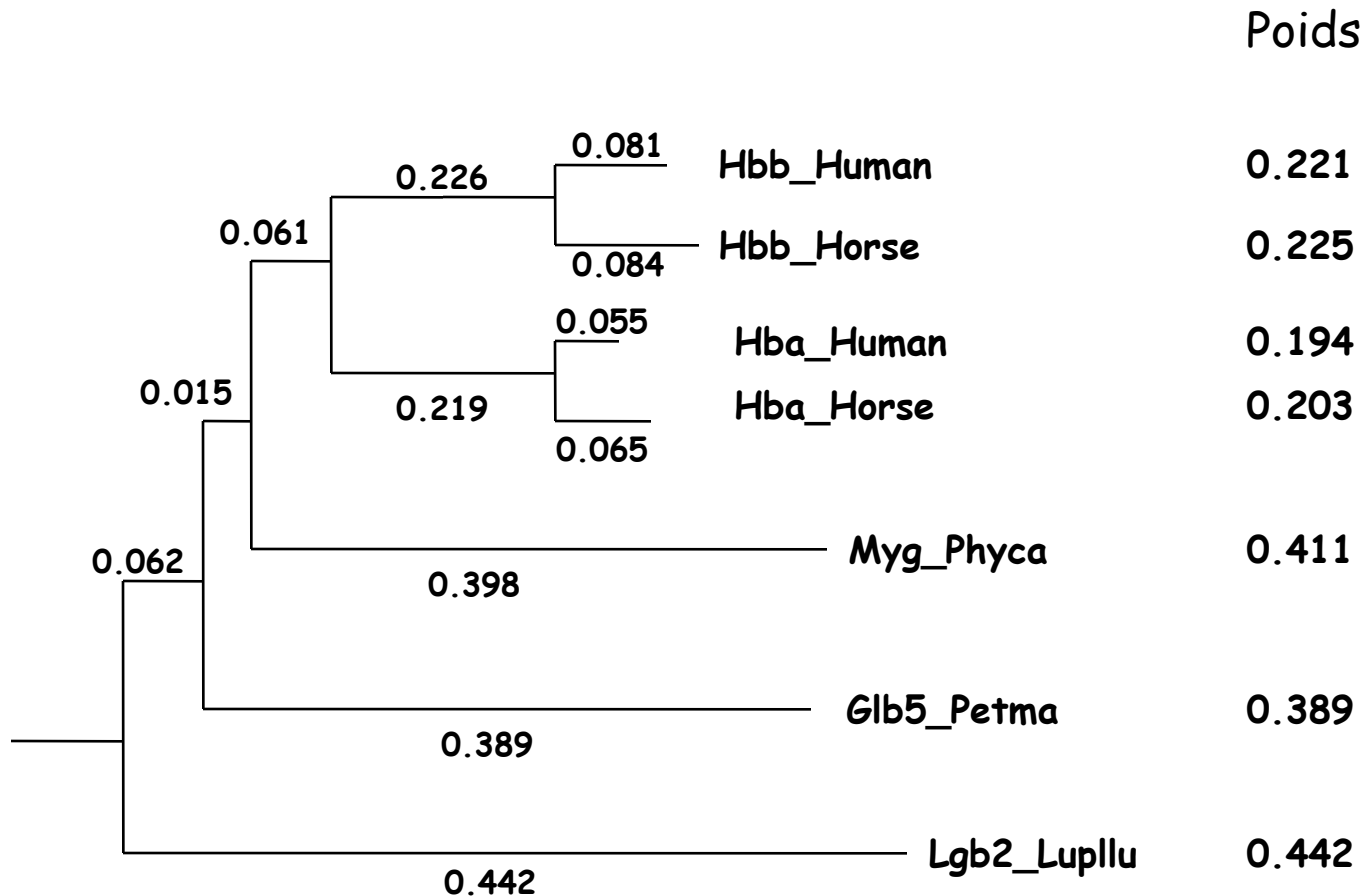
### Pondération des indels :

- Si un indel doit être inséré dans une région contenant déjà un indel, la pénalité d'ouverture est réduite en fonction du nombre de séquences possédant cet indel, et la pénalité d'extension est divisée par 2.
- Insertion d'un indel dans une région sans indel mais dont la distance à un indel déjà présent est  $< 8$  alors la pénalité d'ouverture est augmentée
- Insertion d'un indel dans une suite de 5aa ou plus hydrophyles alors la pénalité d'ouverture est diminuée. Suite de 5aa ou plus (D,E,G,K,N,Q,P,R ou S) = boucle, random coil, donc région non impliquée dans un élément de structure secondaire

# Améliorations apportées à la dernière version de ClustalW

Calcul du poids des séquences du premier exemple :

Extrait de Nucleic Acids Res., 22, 4673-80 (1994)



# Exemple de l'alignement multiple des protéines homologues à ComD

ClustalX (1.85)

File Edit Alignment Trees Colors Quality Help

Multiple Alignment Mode Font Size: 8

1	SagaC_R6FC	HPKNDIDLDVNDPAINLSLLAAKLSQALANIMVSIKVPDSIKPQGMELDITIVSVLIDNAREASLETDDPKIAMSVTTLKSHKQILTIENSTKEMALDSTIPNTEISSKGNRGGIGLNNVKGW
2	SequiCOMD1	KQKIDVGRLLIDVTDTALKSLLAAKNSQALESGISVSIKVPEDISFQGMELDITIVSILCDNAREARVEASSKIIITISVTLIDNMQILMIENNTKKEEIDLSELYDQVSTKGDNRGVLGNVAREI
3	SpneA_HK13	DNKIDLGRLLVNDPAINLSLLAGKFLMAREKNIVNVEVPEEIQVEGMSLDDITIVSILCDNAREASREASQPHVSLAFKNGAQETFTIENSINKEEGIDISEIPSSGASSKGEERGVGLYTVMKI
4	SmitA_BlpH	DNKIDLGRLLVNDPAINLSLLAGKFLMAREKNIVNVEVPEEIQVEGMSLDDITIVSILCDNAREASREASQPHVSLAFKNGAQETFTIENSINKEEGIDISEIPSSGASSKGEERGVGLYTVMKI
5	SpyoE_0539	SEKFNLRALVNDPAINLPLSLLVFLSEARQLSIPVSIETGEPISLENELDITIVSILCDNAREASREASQ---IVLCFPEQKNGKQILVQNDTMEQDNLSTIPERGVSSEKGNRGGIGLNNVREI
6	StheC_HK09	TEFNDLARLVNDPAINLSLLSAKVIQAEEEGETAQLSIPFPIHLIGMEILDITIVSILCDNAREASREASQKIIITISFQEDHRSQLLSISNTISEEQETETKTIIPERGVSSEKGNRGGIGLNNVTEI
7	SmutA_COMD	DFRDNHGLANVQNDPAINVGLISAKILEARQNKKIAVNVVSSKIQLPPEMELDITIVSILCDNAREARAEISLNPEIQLAFFKNGSIVTIIQNSTKEMQIDVSKIPKENTSTKGSNRGGIGLAKVNHII
8	SgorA_BlpH	DRKIDIANLSKIENTALKSVLSAKLLEAANKGVEIAIEVDGRQDLSPPELDDITIVSILCDNAREATLISKDPEINLAIINSHNDRFYFIVENKTKKEEIVNENIPKNGFSTKGEERGGIGLKYVQTI
9	SpyoE_0224	DKKIDNISKLSKISVSSINSLLSSKIIIAEAKGALSIEVDPDIRDSVISELDLILLVSIICDNAREARLETAQPAVALAFVLSDDQIIFVITNTTNDKRVMSRIPPEEGSSKGSERGGIGLSNAKRI
10	SequiCOMD2	DKKIDNISKLSKISVSSINSLLSSKIIIAEAKGALSIEVDPDIRDSVISELDLILLVSIICDNAREARLETAQPAVALAFVLSDDQIIFVITNTTNDKRVMSRIPPEEGSSKGSERGGIGLSNAKRI
11	SpyoE_0226	STNTEISSLDNTEAVIRSIINSKLLEAQVGLIELVTEIPDVIDHLPKLLDVLVLTGLVNDNAREATANGSRNPLSLAFVLSDDQIIFVITNTTNDKRVMSRIPPEEGSSKGSERGGIGLSNAKRI
12	SequiCOMD3	QITPELSALDNDVQDSVIRSLIARAKLLEAKLQGLAVYIDIPETLASSVSLDDVIVLKVVIDNALLSAKGSSEPLSLAFVLSDDQIIFVITNTTNDKRVMSRIPPEEGSSKGSERGGIGLSNAKRI
13	SepiA_1637	TRSIKMGTEKLVKREINGLITTKIIQAQEKRIPISTIEVPEIDRIDNNTVELSKRIIGIIVNDNAREASENLEEPLINLAFIDNEE-SVTFIVMVKCSDHDKPKIDHLEFQEGSTKGDNRGLGLSTLKEI
14	ShaeA_A6RC	MRSIKINGTENLKVRAINGLWTTKIIQAQEKNIPISTIEVPELWHEIDNNTIDLSRIIGIIVNDNAREASETLEDALRIAFINTDS-SVTFIVMVKCSDHDKPKIDHLEFQEGSTKGDNRGLGLSTLKEI
15	SaurC_A6RC	QNALKINGTENLKVREINGLITTKIIQAQEKNIPISTIEVPEWTRINDNNDLSRSIGIIVNDNAREASSEIDDPLIRVAFISEN-SVTFIVMVKCADDIPRIDEHFQEGSTKGEGRGLGLSTLKEI
16	SsapA17985	MNAIKINGTENLKVREINGLITTKIIQAQEKSRISISTIEVDPPIEKIEHPTENLSRIIGIIVNDNAREASEKIEDDPLIRVAFISEN-SVTFIVMVKCEPMPKVTILFQENSTKGNRGLGLSTLKEI
17	SpneA_COMD	SDKNTVEDLNNVEDSALRSIWAQSIIVVARNNGVEFTLEVKDTITKLPFELLDLVRIMSVLNNNAVEGSAESVYKQNEVAVIKMET-ETVIVIQNSCKMTMTPSGDLFALGFSTKGNRGGIGLNNVKEI
18	Sora_COMD	SDKNTVEDLNNVEDSALRSIWAQSIIVVARNNGVEFTLEVKDTITKLPFELLDLVRIMSVLNNNAVEGSAESVYKQNEVAVIKMET-ETVIVIQNSCKMTMTPSGDLFALGFSTKGNRGGIGLNNVKEI
19	SmitA_COMD	SDKNTVEDLNNVEDSALRSIWAQSIIVVARNNGVEFTLEVKDKITKLPFELLDLVRIMSVLNNNAVEGSAESVYKQNEVAVIKMET-ETVIVIQNSCKMTMTPSGDLFALGFSTKGNRGGIGLNNVKEI
20	SinfA_COMD	SDKNTVEDLNNVEDSALRSIWAQSIIVVARNNGVEFTLEVKDTITKLSIDLLDVRIMSVLNNNAVEGSAESVYKQNEVAVIKMET-ETVIVIQNSCKMTMTPSGDLFALGFSTKGNRGGIGLNNVKEI
21	Sang_COMD	SDKNTVEDLNNVEDSALRSVMTQTLFARDNNEILTFEVKDFVWPLPIKLLDVRMTSVLNNNAIEGAAESVQKTMNVSIVLDLDT-ETILVIQNSRKKRPLDLEEIQQDFSTKGEGRGLGLSNLKEI
22	Sinte_COMD	SDKNTVEDLNNVEDSALRSVMTQTLFARDNNEILTFEVKDFVWPLPIKLLDVRMTSVLNNNAIEGAAESVQKTMNVSIVLDLDT-ETILVIQNSRKKRPLDLEEIQQDFSTKGEGRGLGLSNLKEI
23	Scons_COMD	SDKNTVEDLNNVEDSALRSVMTQTLFARDNNEILTFEVKDFVWPLPIKLLDVRMTSVLNNNAIEGAAESVQKTMNVSIVLDLDT-ETILVIQNSRKKRPLDLEEIQQDFSTKGEGRGLGLSNLKEI
24	Smill_COMD	SDKNTVEDLNNVEDSALRSVMTQTLFARDNNEILTFEVKDFVWPLPIKLLDVRMTSVLNNNAIEGAAESVQKTMNVSIVLDLDT-ETILVIQNSRKKRPLDLEEIQQDFSTKGEGRGLGLSNLKEI
25	SgorA_COMD	SDDVTFEELNNVQDIALRSVLIQTIPARECGVEIYPEKMDVETLPMKLLDVRVAVSVLNNNAIEGAAESVQKTMNVSIVLDLDT-ETIVVQNSRQSRINLEEIVYVSTKGNRGGIGLNNVKEI
26	SsuiA_COMD	KNEVSAENLKNEDLAINMLAQSILEAQDIDQNCVSGHISIQWVPMEMIQILSIMLINSIEAALISKKPLIEVALRDKS-LVTIVINSRNPISIDKRTIEVGVSSKGEGRGGIGLNNVRL

ruler ... 320 ..... 330 ..... 340 ..... 350 ..... 360 ..... 370 ..... 380 ..... 390 ..... 400 ..... 410 ..... 420 ..... 430 ..... 440 ...

# Exemples de logiciels d'alignement multiple

- Clustal W : (Thompson, Higgins, Gibson, 1994) [www.clustal.org](http://www.clustal.org)  
(progressive)
- T-Coffee : (Notredame, Higgins, Heringa, 2000) [tcoffee.crg.cat](http://tcoffee.crg.cat)  
(consistency)
- Mafft : (Katoh ea, Miyata, 2002) [mafft.cbrc.jp/alignment/server/](http://mafft.cbrc.jp/alignment/server/)  
(progressive/consistency, iterative)
- Muscle : (Edgar, 2004) [www.ebi.ac.uk/Tools/msa/muscle/](http://www.ebi.ac.uk/Tools/msa/muscle/)  
(progressive, iterative)
- Espresso/3DCoffee (Poirot, Notredame, 2004) [tcoffee.crg.cat](http://tcoffee.crg.cat)  
(structures)



# Sélection de séquences

Séquences identiques ou trop proches n'ajoutent pas d'information.

**Diversité** : propice pour l'alignement



**Attention** aux séquences répétées

Utilisation de données expérimentales (Swiss-Prot, PDB..)

## Analyse de séquences

Comparer les alignements obtenus avec différentes méthodes permet d'analyser les conservations et la variabilité.

# Editeurs d'alignements

- . Jalview, [multiplatform \(JAVA\)](#) (Dundee, UK)  
[www.jalview.org](http://www.jalview.org)
- . Seaview, [multiplatform](#) (+ phylogenetic trees) (Lyon, Fr)  
[pbil.univ-lyon1.fr/software/seaview.html](http://pbil.univ-lyon1.fr/software/seaview.html)
- . BioEdit, [Windows](#) (CA, USA)  
[www.mbio.ncsu.edu/bioedit/bioedit.html](http://www.mbio.ncsu.edu/bioedit/bioedit.html)