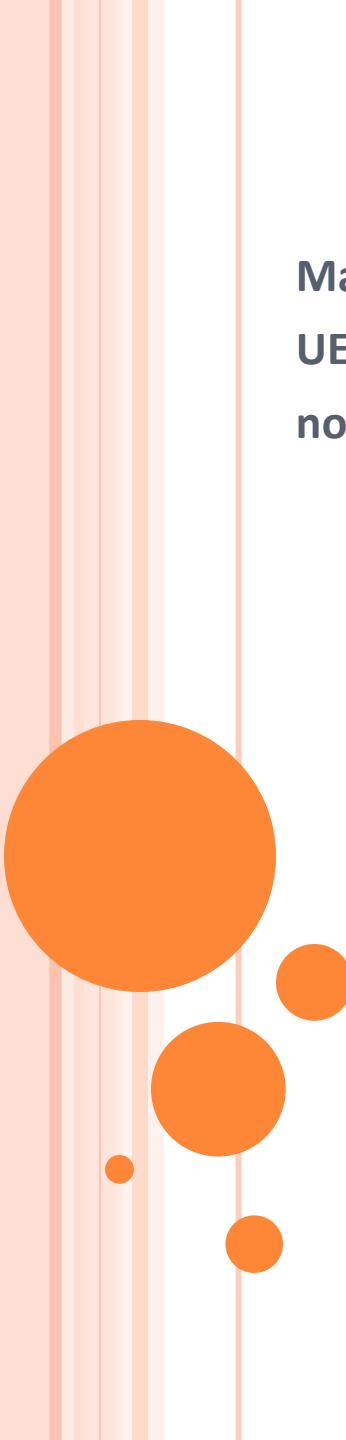


- <http://www.elixir-europe.org/news/how-fast-life-science-data-growing>



Master 1 MABS

UE Bioinformatique des séquences

novembre - décembre 2013

ALIGNEMENTS MULTIPLES DE SÉQUENCES BIOLOGIQUES

POURQUOI ANALYSER DES ENSEMBLES DE SÉQUENCES?

Un alignement multiple de séquences révèle des aspects que l'on ne peut pas visualiser en comparant 2 séquences

- Identifier des **acides aminés essentiels**;
- Identifier des **domaines**;
- Établir des **signatures** de famille de protéines;
- Comme une aide à la **modélisation structurale**.

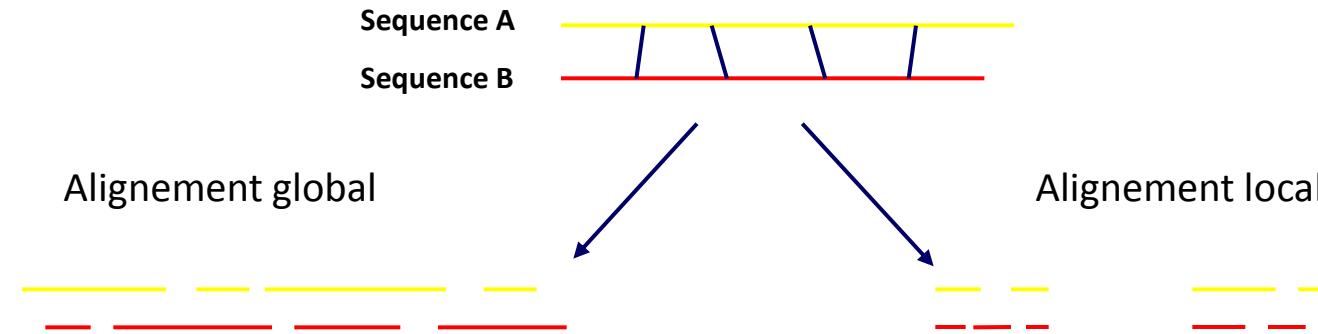
Les algorithmes de prédition de structures secondaires exploitent beaucoup mieux les alignements multiples.

Connaître les acides aminés permis à telle ou telle position facilite l'inférence 3D;

- Domaines transmembranaires;
- A partir d'un alignement, on détermine les résidus les plus fréquents à chaque position. Si la fréquence dépasse un certain seuil : séquence inclue dans le consensus. Par exemple **consensus** de 90%
- Établir la **phylogénie** des séquences, des organismes.



ALIGNEMENT PAR PAIRES



Alignement des séquences sur la totalité de leurs longueurs

G G C T G A C C A C C - T T
| | | | | | | |
G A - T C A C T T C C A T G

Alignement global par paires optimal :
[Needleman and Wunsch](#), 1970
Premier algorithme d'alignement

Alignement de régions hautement similaires

G A C C A C C T T
| | | | | | | |
G A T C A C - T T

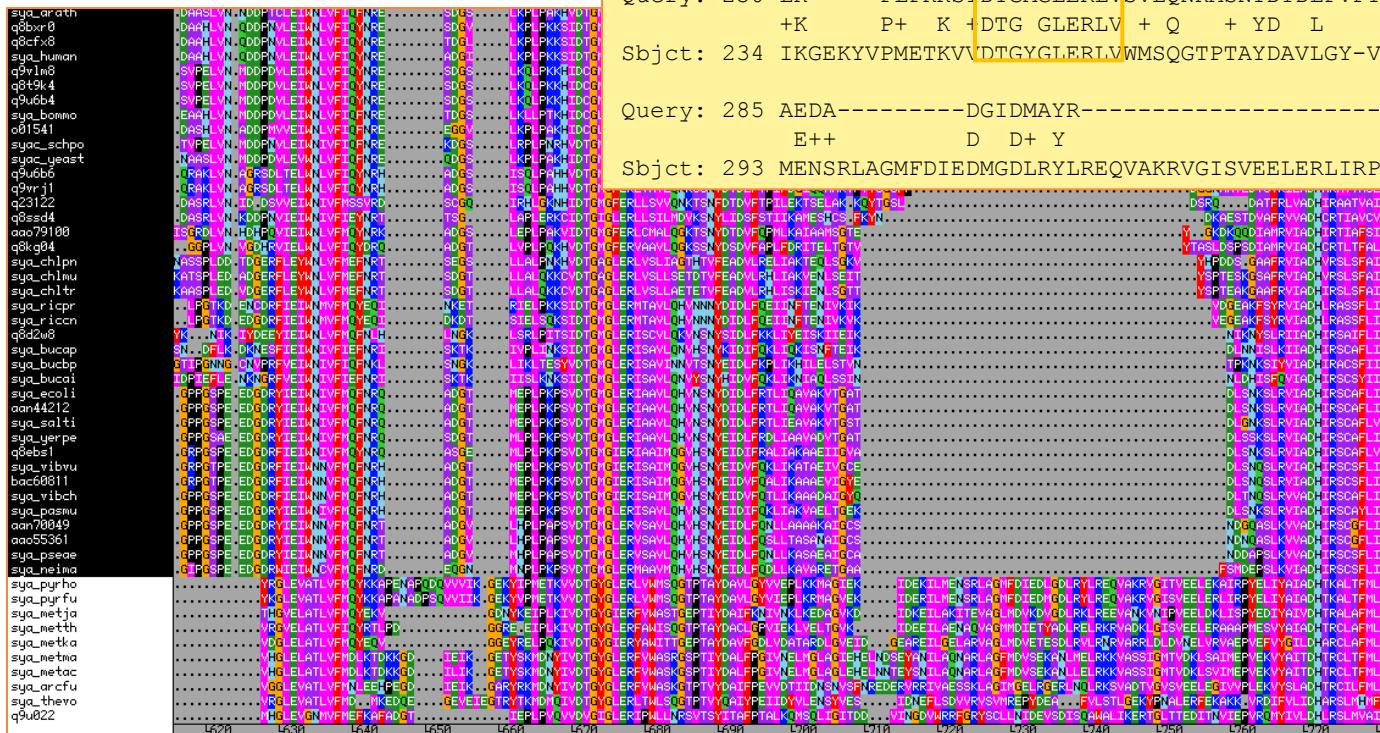
Alignement local par paires optimal :
[Smith and Waterman](#), 1981 (modification de N&W)
Recherche de la région de plus forte similarité entre les séquences

Ce sont des algorithmes exacts (programmation dynamique) qui garantissent de construire le meilleur alignement



ALIGNEMENT GLOBAUX PAR PAIRES / ALIGNEMENT MULTIPLE

Alignement par paires



Alignement multiple

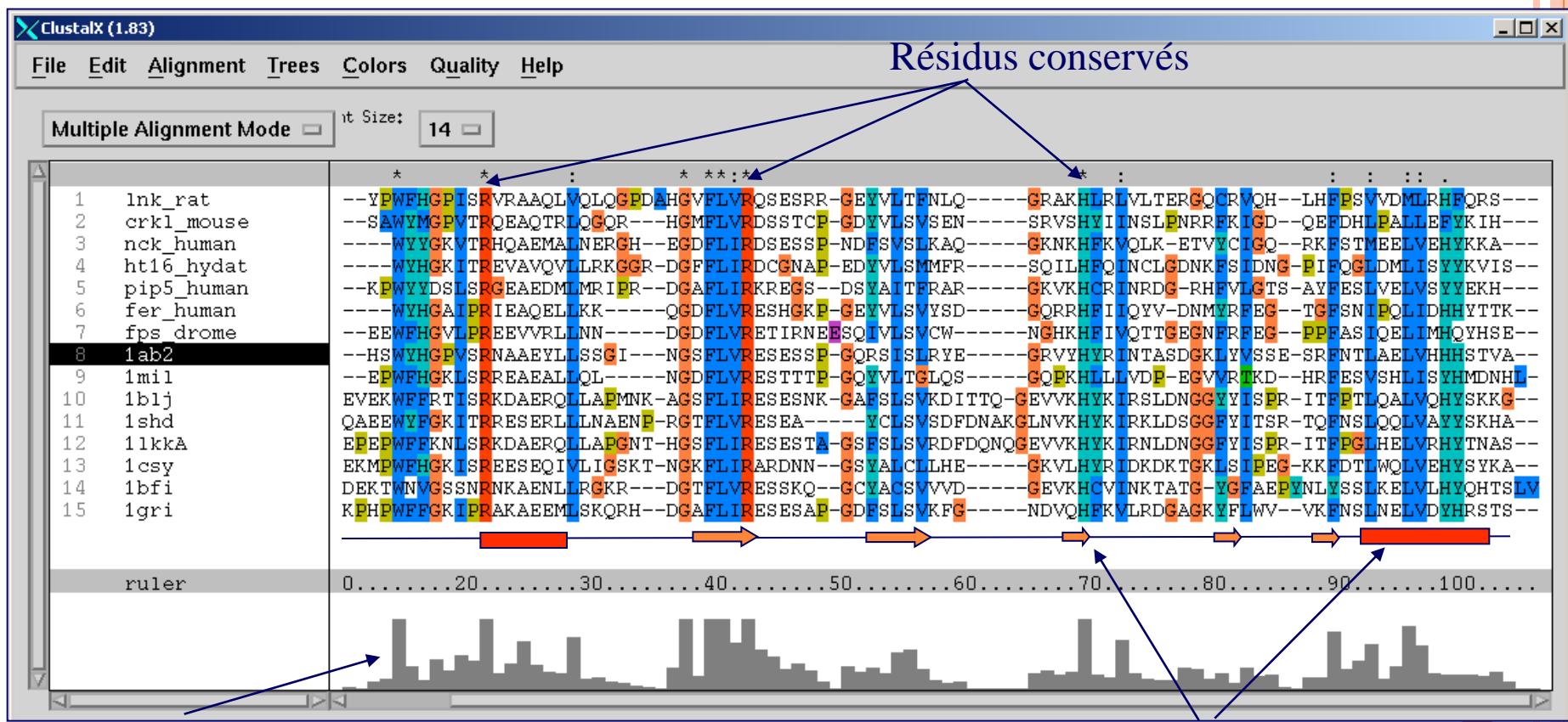
Les algorithmes utilisés pour la comparaison de 2 séquences sont trop gourmands en calculs.

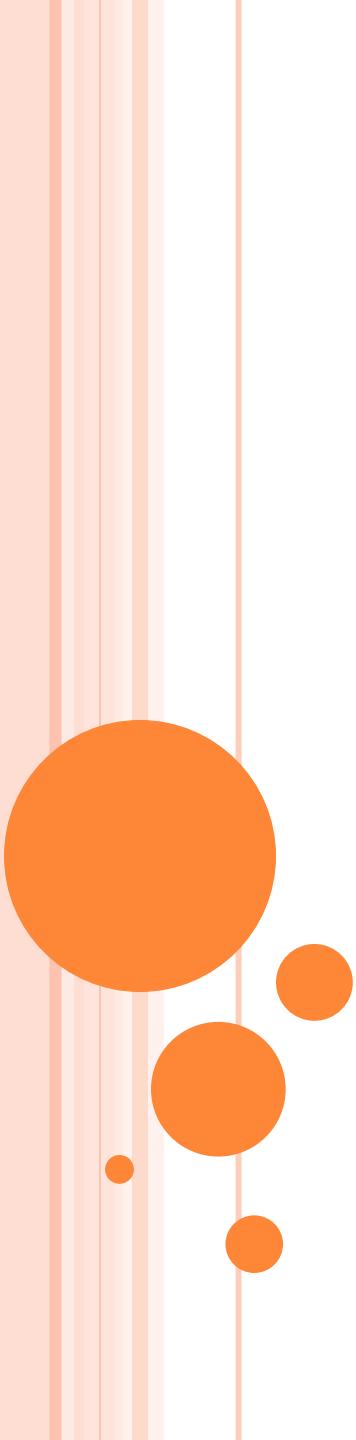
Besoin de simplifier les calculs au détriment de la qualité des résultats.



QU'EST-CE QU'UN ALIGNEMENT MULTIPLE ?

Une représentation d'un ensemble de séquences, dans lesquelles les résidus équivalents (d'un point de vue fonctionnel ou structural) sont alignés en colonnes.





CONSTRUIRE UN ALIGNEMENT MULTIPLE

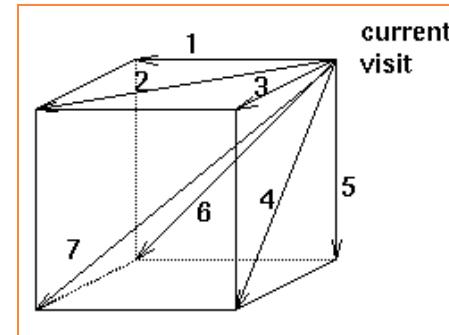
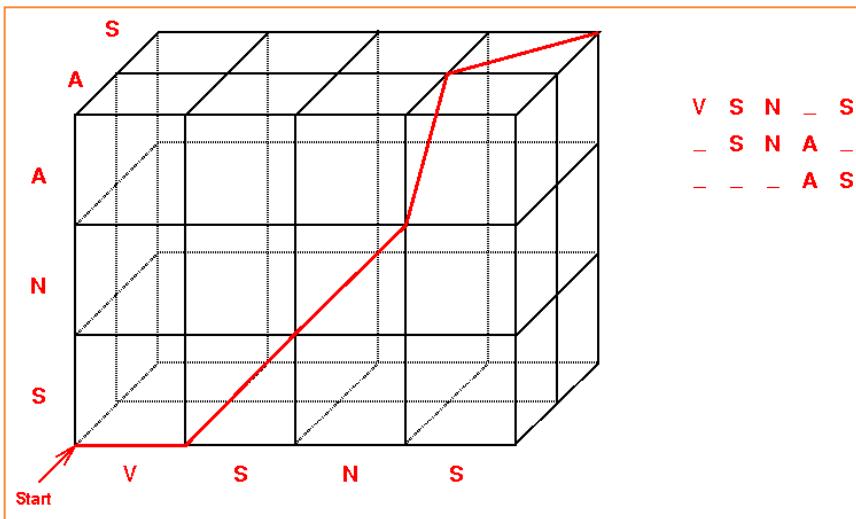
APPROCHES TRADITIONNELLES

- A. alignement multiple optimal
- B. alignement multiple progressif
- C. alignement multiple itératif



A. ALIGNEMENT MULTIPLE OPTIMAL

- Extension directe des programmes dynamiques* des alignements de séquences par paires à N dimensions (Sankoff, 1975).
- Examine l'ensemble des alignements possibles afin de trouver l'alignement optimal
- Exemple: alignement de 3 séquences

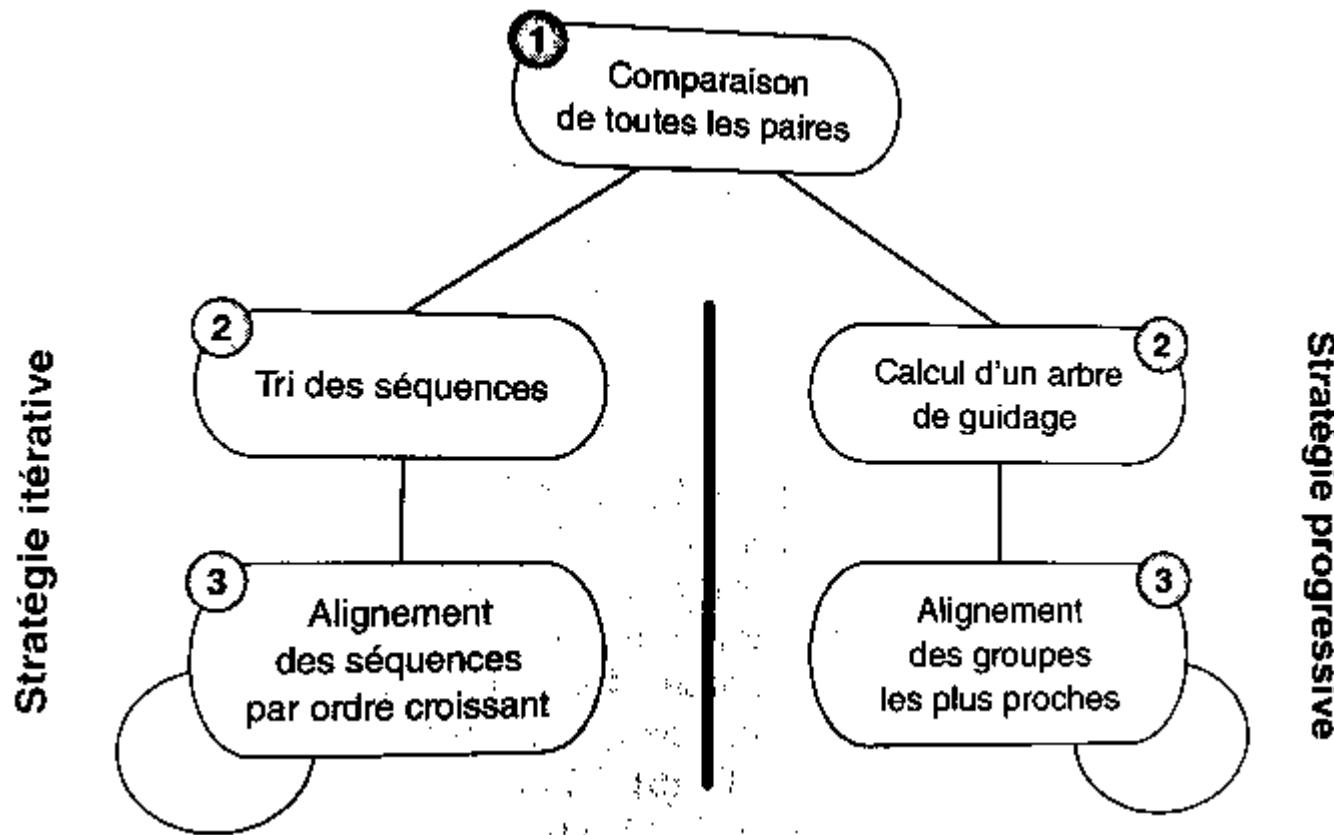


Problème

- L'alignement mathématique optimisé n'est pas nécessairement l'alignement biologique optimal.
- Le temps CPU (temps de calcul sur ordinateur) et la mémoire requise sont prohibitifs pour un usage classique (temps requis est proportionnel à N^k avec k séquences de longueur N).
- En pratique, moins de 10 séquences peuvent être alignées.

*La programmation dynamique est une technique algorithmique qui permet de résoudre une catégorie particulière des problèmes d'optimisation sous contrainte. Elle a été désignée par ce terme pour la première fois dans les années 1940 par Richard Bellman. Elle s'applique à des problèmes d'optimisation dont la fonction objectif se décrit comme « la somme de fonctions monotones non-décroissantes des ressources ».

STRATÉGIES D'ALIGNEMENTS NON-OPTIMALES



B. ALIGNEMENT MULTIPLE PROGRESSIF

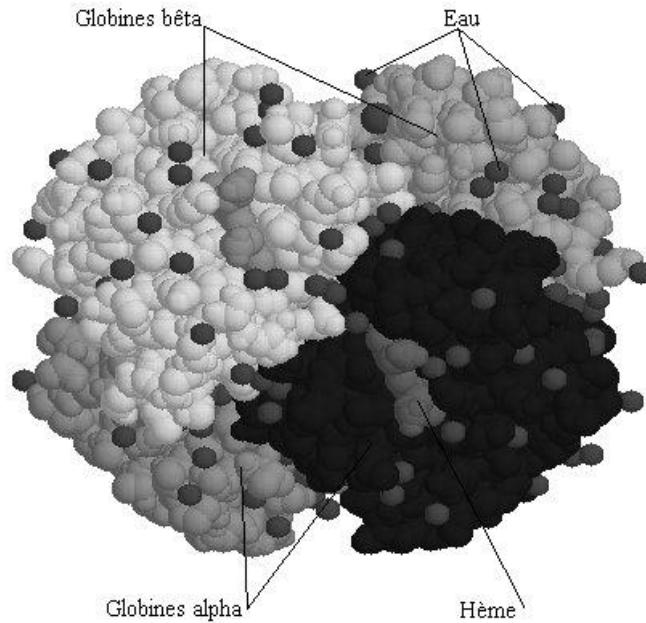
- Un algorithme heuristique qui évite le calcul de l'ensemble des alignements possibles, mais il ne garantie pas l'alignement optimal
- Principe :
 - Les séquences (ou groupe de séquences) sont alignées progressivement par paires
- Problème :
 - Quelles sont les deux premières séquences à aligner? Dans quel ordre aligner les séquences ?
 - On aligne en premier les deux séquences les plus proches
 - Comment estimer la distance entre deux séquences ?
 - Aligner toutes les paires de séquences
 - Calculer la matrice de distance à partir des alignements par paires
 - Construire un arbre guide à partir de la matrice de distance
 - Alignement multiple progressif selon l'ordre des branches de l'arbre



ALIGNEMENT MULTIPLE PROGRESSIF

Exemple :

Alignment de 7 globines (Hbb_human, Hbb_horse, Hba_human, Hba_horse, Myg_phyca, Glb5_petma et Lgb2_lupla)



ALIGNEMENT MULTIPLE PROGRESSIF

○ Étape 1: alignement par paire de toutes les séquences

```

Hbb_human    1 LTPEEKSAVTALWGKV..NVDEVGGEALGRLLVVYPWTQRFFESFGDLST ...
                  | . | : . | | | ||| . | | | ||| : . : | . : | | | |
Hba_human    3 LSPADKTNVKAAGKVGAAHAGEYGAELERMFLSFPTTAKTYFPHF.DLS. ...

```

```
Hba_human    3 LSPADKTNVKAAGKVGAGHEGYGAEALERMFLSFPTTKTYFPFH . DLSH ...
                  || :| | | | | || | | | | |: . :| | . :| | | | .
Hbb_horse    2 LSGEEKAAVLALWDKVNEE..EVGGEALGRILLVVYPWTQRFFDSFGDLSN ...
```

- Les alignement peuvent être obtenus avec:

- méthodes globales ou locales
 - Programmation dynamiques ou méthodes heuristiques

*Exemple : Clustalw (ou Clustalx, idem mais avec une interface graphique), Muscle...
=> alignements globaux*



ALIGNEMENT MULTIPLE PROGRESSIF

○ Étape 2: construction de la matrice de distance

Dans Clustalw:

$$\text{Distance entre deux séquences} = 1 - \frac{\text{Nb de résidus identiques}}{\text{Nb de résidus comparés}}$$

Ex : Hbb_human vs Hbb_horse = 83% identité = distance de 17%

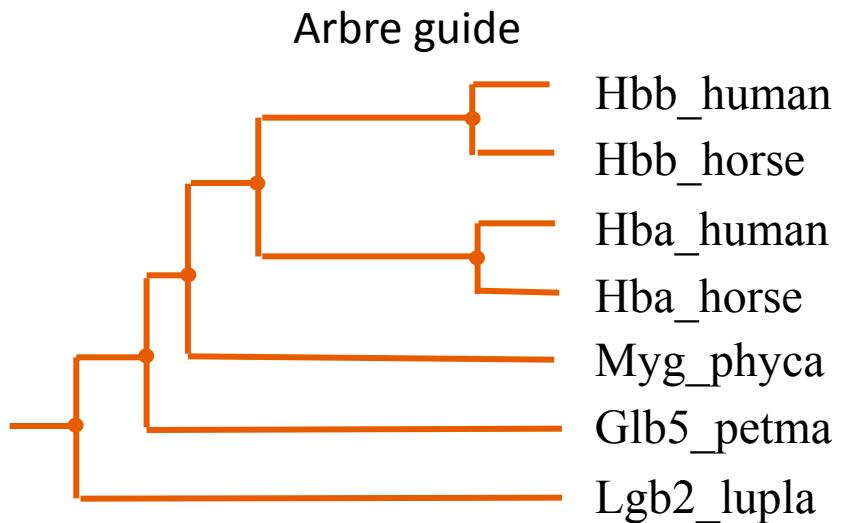
Hbb_human	1	-					
Hbb_horse	2	.17	-				
Hba_human	3	.59	.60	-			
Hba_horse	4	.59	.59	.13	-		
Myg_phyca	5	.77	.77	.75	.75	-	
Glb5_petma	6	.81	.82	.73	.74	.80	-
Lgb2_lupla	7	.87	.86	.86	.88	.93	.90
	1	2	3	4	5	6	7



ALIGNEMENT MULTIPLE PROGRESSIF

- Étape 3: construction de l'arbre guide

	1	2	3	4	5	6	7
Hbb_human	-						
Hbb_horse	.17	-					
Hba_human	.59	.60	-				
Hba_horse	.59	.59	.13	-			
Myg_phyca	.77	.77	.75	.75	-		
Glb5_petma	.81	.82	.73	.74	.80	-	
Lgb2_lupla	.87	.86	.86	.88	.93	.90	-
	1	2	3	4	5	6	7

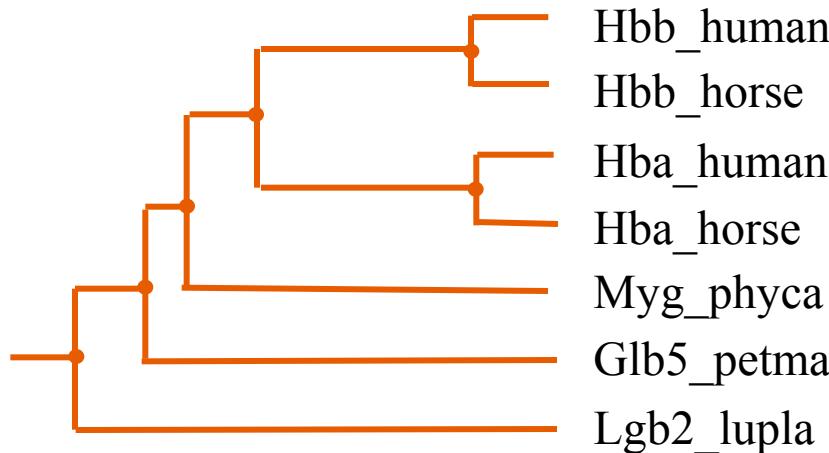
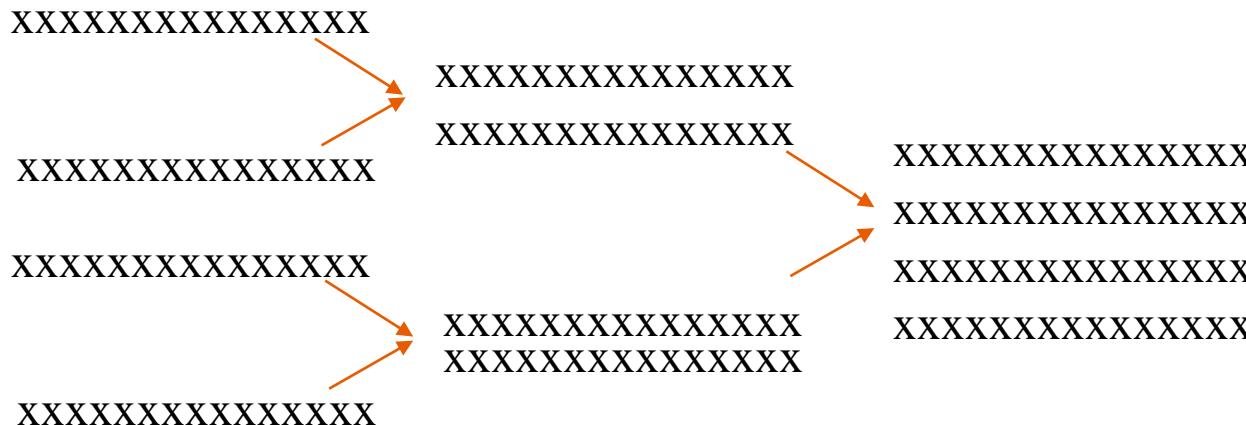


1. Joint les deux séquences les plus proches
2. Calcul à nouveau les distances et joint les deux séquences les plus proches ou les noue
3. Répétition de l'étape 3 jusqu'à ce que toutes les séquences soient jointes.



ALIGNEMENT MULTIPLE PROGRESSIF

- Étape 4: Alignement progressif selon l'ordre des branches de l'arbre guide



ALIGNEMENT MULTIPLE PROGRESSIF

Sequence alignment of HBB, HBB, HBA, HBA, MYG, GLB5, and LGB2 proteins across seven domains (H1-H7). Domains H1-H4 are shown in the top panel, and H5-H7 in the bottom panel. Conserved residues are highlighted in orange.

	H1	H2	H3	H4	H5	H6	H7	
HBB_HUMAN	-----VHL PEEKSAVTALWGKV	N-----VDEVGGEALGRLLVV	-----PWTQRPFESFGDLSTPDAVMGNP	PKVKAHGKKVLGAFSDGLAHLDN	-----LKGTFATLSELHC	-----LKGTFATLSELHC	-----LKGTFATLSELHC	-----LKGTFATLSELHC
HBB_HORSE	-----VQLSGEEKAAVLALWDKV	N-----EEEVGGEALGRLLVV	-----PWTQRPFDSFGDLSNPGAVMGNP	PKVKAHGKKVLHSFEGEVHLDN	-----LKGTFPAALSELHC	-----LKGTFPAALSELHC	-----LKGTFPAALSELHC	-----LKGTFPAALSELHC
HBA_HUMAN	-----VLSPADKTNVKAAWGKV	G-----AGEYGAEALERMFLS	-----PPTTKTIFPHF-DLS-----	HGSAQVKGHGKVKADALTNAVAHVDD	-----MPNALSALSDLHA	-----MPNALSALSDLHA	-----MPNALSALSDLHA	-----MPNALSALSDLHA
HBA_HORSE	-----VLSAADKTNVKAAWSKV	GG-----AGEYGAEALERMFLG	-----PPTTKTIFPHF-DLS-----	HGSAQVKAHGKKVGDALTLAVGHLDD	-----HKLRVDPVNF	-----HKLRVDPVNF	-----HKLRVDPVNF	-----HKLRVDPVNF
MYG_PHYCA	-----VLSEGEWQLVLHVWAKV	E-----VAGHGQDILIRLFKS	-----HPETLEKFDRFKHLKTEAEMKASE	DLKKHGVTVLTALGAILRKKGH	-----HEAELKPLAQSHA	-----HEAELKPLAQSHA	-----HEAELKPLAQSHA	-----HEAELKPLAQSHA
GLB5_PETMA	PIVDTGSVAPLSAAEKTKIRSAWAPV	YST-----YETSGVDILVKFTS	-----TPAAQEFPPFKFKLTTADQLKKSADVRWAHERIINNAVNDAVA	SMDD	T-----EKMSMKLRLDLSGKHA	T-----EKMSMKLRLDLSGKHA	T-----EKMSMKLRLDLSGKHA	T-----EKMSMKLRLDLSGKHA
LGB2_LUPLU	-----GALTE SQAALVKSSWEER	NA-----IPKHTHRFFILVLEI	-----TPAAAKDFLFSFLKGTSEVP--QNNPELQAHAGKVF	KLVYEAAIQLQV	TG-----TGAATLKNLGSVHV	TG-----TGAATLKNLGSVHV	TG-----TGAATLKNLGSVHV	TG-----TGAATLKNLGSVHV

ALIGNEMENT MULTIPLE PROGRESSIF: AUTRE EXEMPLE

Soient les 4 séquences :

S1 cgatgagtcattgtgactg
S2 cgagccattgttagctactg
S3 cgaccattgttagctacacctg
S4 cgatgagtcactgtgactg

Alignements 2 à 2

```

S1 cgatgagtcattgt-g--actg
      ||| |    ||||| |    |||||
S2 cga-g--ccatttgttagctactg

```

S2 cgagccattgttagcta-ctg
| | | | | | | | | | | | | | |
S3 cqa-ccattgttagctaccta

S1 cgatgagtcattg-tgactg
| | | | | | | | | |
S3 cgacca-**ttgt**tagctaccctg

S2 cga-g--ccattgttagctactg
| | | | | | | | | | | | | |
S4 cgatgagtcactgt-q--actg

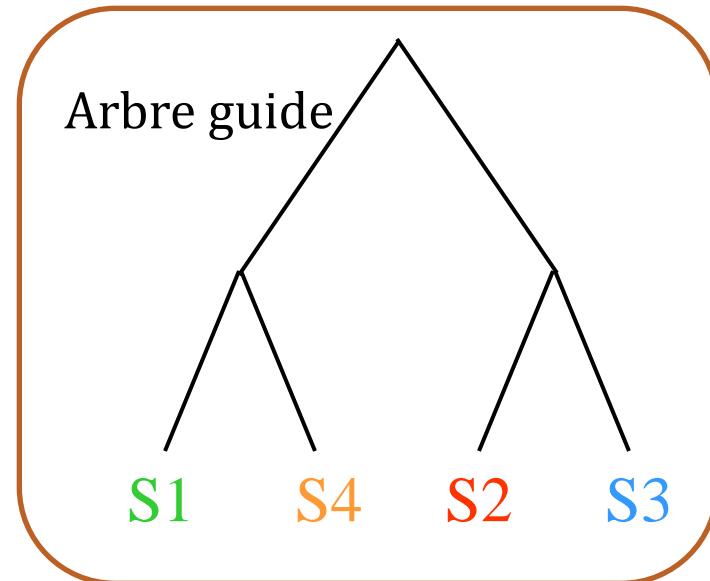
S1 cgatgagtcattgtgactg
| | | | | | | | | | | | | |
S4 cqatqaqtcaactqtqactq

S3 cgaccattgttagctacctg
||| | | | | | |||
S4 cgatgagtcactgtqactg

ALIGNEMENT MULTIPLE PROGRESSIF: AUTRE EXEMPLE

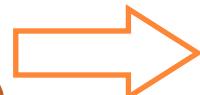
Tableau des scores d'alignement

	S1	S2	S3	S4
S1	-	2	0	17
S2		-	14	0
S3			-	-1
S4				-



Construction de l'alignement multiple final

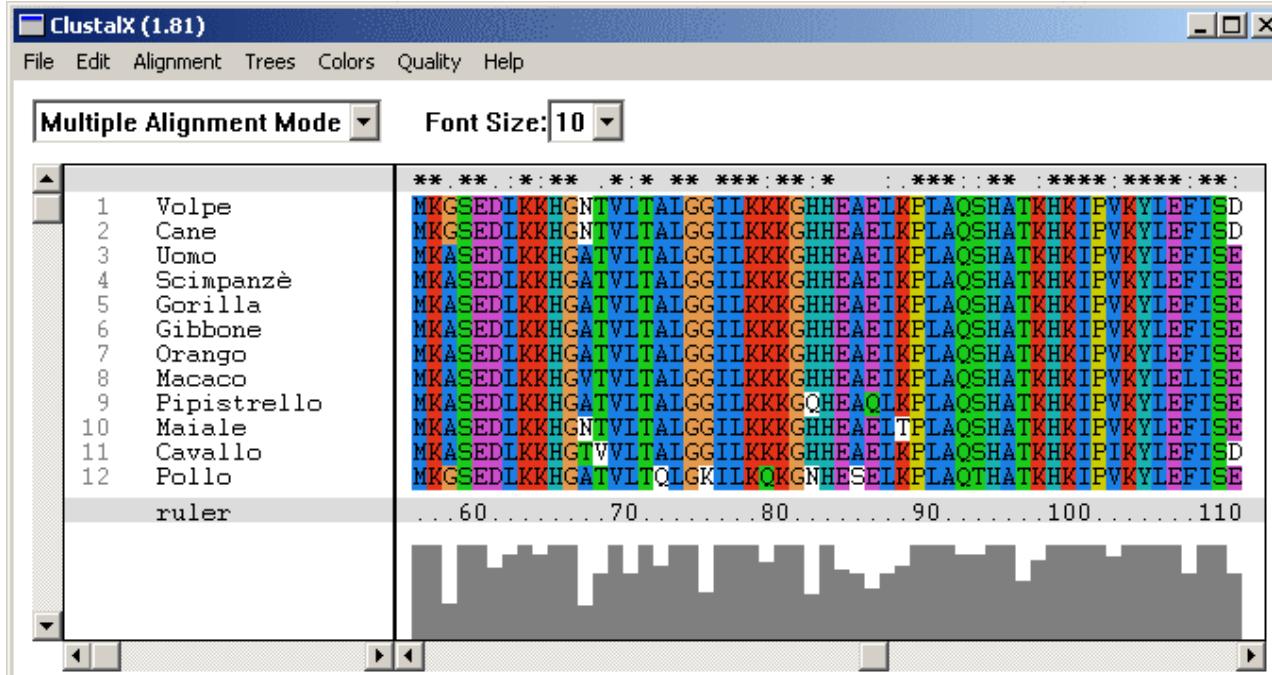
S2 cgagccattgttagcta-ctg
| | | | | | | | | | | | | | |
S3 cga-ccattgttagctaccta



S1 cgatgagtcatgt-g--ac-tg
S4 cgatgagtcaactgt-g--ac-tg
S2 cga---gccattgttagctac-tg
S3 cga----ccattgttagctacctg

CLUSTALW

- ClustalW (de Des Higgins) est le programme d'alignement multiple le plus employé.
- Version actuelle: CLUSTAL Omega - Clustal Omega is a new multiple sequence alignment program that uses seeded guide trees and HMM profile-profile techniques to generate alignments.
- ClustalX c'est ClustalW avec interface graphique

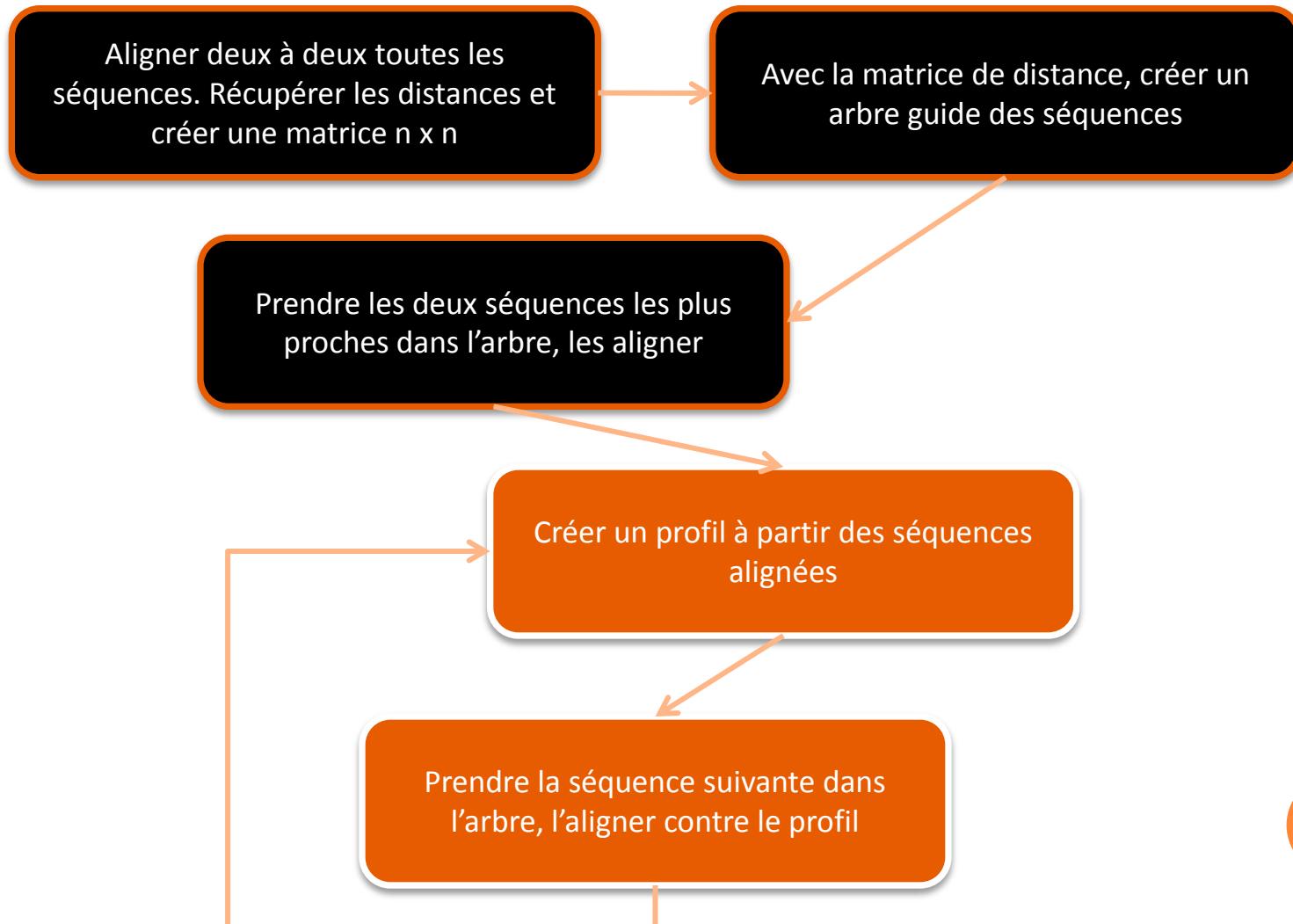


CLUSTALW

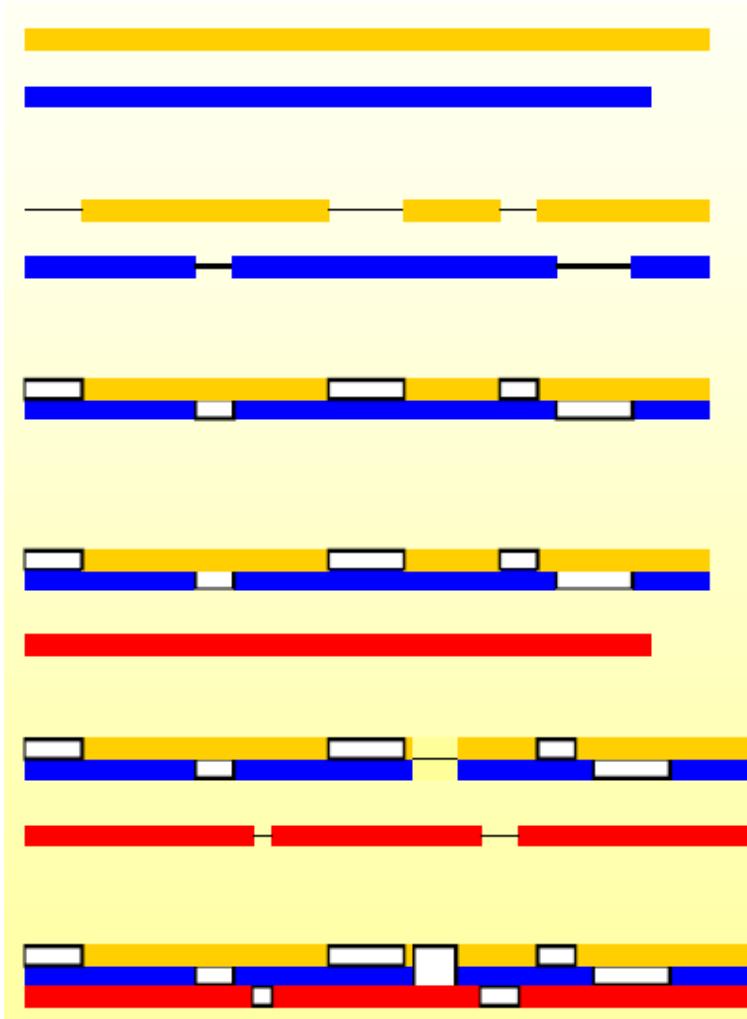
- ClustalW utilise les profils.
 - Les séquences déjà alignées servent de profil pour diriger la suite de l'alignement.
 - Les profils sont représentés sous forme de tableau dans lequel sont données pour chaque position la fréquence observée de chaque lettre.
- Chaque nouvelle séquence est alignée contre le profil des séquences déjà alignées.



ALGORITHME DE CLUSTALW, UN ALGORITHME PROGRESSIF



CLUSTALW: ALIGNEMENT ET CRÉATION DES PROFILS



- 2 séquences à aligner globalement
- Alignement 2 à 2 avec insertion/délétion
- Remplissage des trous et fusion par formation d'un profil
- Profil à aligner avec une autre séquence
- Alignement profil et séquence avec insertion/délétion
- Remplissage des trous et fusion par formation d'un profil



LES PARTICULARITÉS DE CLUSTALW

- Pondération des séquences en fonction de leur sur/sous représentation (Sequence weighting).
 - Lorsque l'alignement contient plusieurs séquences très proches, celles-ci vont prendre plus d'importance dans le profil qu'une séquence isolée éloignée de celles-ci.
Pour éviter que de tels groupes de séquences proches biaissent l'alignement, l'importance de chaque séquence dans le profil est pondérée en fonction du nombre de séquences proches, telles que déterminés par un arbre.
- Adaptation des matrices de similitudes au fil de l'algorithme en fonction de la divergence des séquences à aligner
 - Blosum 80 pour aligner des séquences proches
 - Blosum 50 pour aligner des séquences distantes



LES PARTICULARITÉS DE CLUSTALW

- Pénalités de gaps spécifiques à chaque résidu.
 - Par exemple, les Glycines sont davantage susceptibles d'avoisiner un gap que les Valines.
- Pénalités de gaps réduites dans les régions hydrophiles
 - Encourage la formation de gaps dans des boucles plutôt que dans des régions structurées.
- Pénalités de gaps augmentées dans le voisinage d'autres gaps
 - Evite la formation de petits gaps voisins, au profit de longs gaps.
- Attention l'arbre guide généré dans ClustalW n'est pas un arbre phylogénétique



ALGORITHME DE CLUSTALW

- Les scores d'alignement calculés par clustalW utilisent 2 méthodes:
 - la **programmation dynamique** qui a pour avantage d'être **optimale mais lente**. La programmation dynamique sera plutôt utilisée pour des jeux de courtes séquences mais deviendra extrêmement lente pour des jeux de données supérieures à 100 séquences de 1000 acides aminés chacune.
 - N.B. : Bien qu'implémentant de la programmation dynamique (algorithme de *Needleman et Wunsch*) pour l'alignement par paires, clustalw n'utilise pas la programmation dynamique pour l'alignement multiple global.
 - ou **l'algorithme de Wilbur et Lipman**, qui est **très rapide mais plus approximatif**.
 - Le programme n'utilise que les meilleures diagonales, c'est à dire celles présentant le plus de fragments d'appariements exacts.
- Il faut également noter que **l'ordre des séquences** dans le fichier d'entrée joue un **rôle important sur les résultats finaux**.



CONCLUSION SUR CLUSTALW

- Grand succès de ClustalW
- Plus récemment, d'autres programmes ont vu le jour, fondés sur d'autre algorithmes ou heuristiques et donnent souvent de meilleurs résultats dans les cas difficiles. Si vos séquences sont difficiles à aligner (peu de similarités, longueurs différentes), il est impératif d'essayer d'autres programmes comme DIALIGN, MAFFT ou MUSCLE (méthode itérative).
- N'oubliez pas que clustalW – et il n'est pas le seul – souffre d'un défaut congénital gravissime: certes il effectue un alignement multiple, mais il le fait à partir de l'alignement des paires de séquences. Autrement dit, quand il aligne la j -ième séquence avec la n -ième séquence pour calculer leur score d'alignement et construire l'arbre de guidage, il ignore « royalement » toute l'information contenue dans les autres séquences. Un court motif commun à deux séquences peut ne pas être repéré, même s'il est commun à toutes les séquences...

PRANK

- PRANK is a probabilistic multiple alignment program for DNA, codon and amino-acid sequences.
- It's based on a novel algorithm that treats insertions correctly and avoids over-estimation of the number of deletion events.
- In addition, PRANK borrows ideas from maximum likelihood methods used in phylogenetics and correctly takes into account the evolutionary distances between sequences.
- Lastly, PRANK allows for defining a potential structure for sequences to be aligned and then, simultaneously with the alignment, predicts the locations of structural units in the sequences.



PRANK

- The reconstruction of evolutionary homology -- including the correct placement of insertion and deletion events -- is only feasible for rather **closely-related sequences**.
- **PRANK is not meant for the alignment of very diverged protein sequences.**
- If sequences are very different, the correct homology cannot be reconstructed with confidence and PRANK may simply refuse to match them.



PRANK

- **Differences to other alignment methods**
- PRANK aims at an evolutionarily correct sequence alignment.
- It uses **evolutionary information** for the placement of gaps and modelling of the substitution process.
- When this information is correct, PRANK makes superior alignments compared to other progressive methods.
- However, when its assumptions are violated, the program's performance may be significantly affected.
- Below is a brief description of potential problems caused by an incorrect alignment guidetree.



WEB-PRANK:

HTTP://WWW.EBI.AC.UK/GOLDMAN-SRV/WEBPRANK/



Submit alignment task

Sequence input and submission

Sequence data (required):

Paste sequences in Fasta format or choose a file to upload

Alignment title (optional):

You can start the alignment by clicking the button above. The tabs below allow you to change the alignment options and use the advanced features of the PRANK algorithm. [More information.](#)

Basic alignment options

Guide tree (optional)*:

Paste your tree in Newick format or choose a file to upload

 trust insertions (+F) compute reliabilityInference of insertions and deletions:
Alignment reliability:

Alignment of DNA sequences:

 default Use structure model Fast/Slow Use structure model Genomic align translated codons align translated proteins align translated mt proteinsChange the default alignment options. [More information.](#)

Advanced alignment options

Extra options for structure models (DNA)

Retrieve finished job

PRANK / WEB-PRANK

- Référence: Loyačnoja and Goldman, 2005; Loyačnoja and Goldman, 2008
- Une étude portant sur la partie codante de 12 génomes de drosophiles a montré que PRANK surpassait les autres méthodes d'alignement testées : T-Coffee (Notredame, et al., 2000), ProbCons (Do, et al., 2005), AMAP (Schwartz and Pachter, 2007), ClustalW et Muscle lors de calculs de sélection positive avec des modèles sur site.
- Néanmoins, il semble que cette méthode produise toujours un grand nombre de faux positifs (Markova-Raina and Petrov, 2011).



C. PRINCIPE DE LA MÉTHODE ITÉRATIVE

1. Dans une première phase, on calcule un score de similarité entre toutes les paires de séquences par comparaison des séquences deux à deux; on obtient un ensemble de scores d'alignement qui sont regroupés dans une matrice dites de similarités
2. Cette matrice est utilisée pour trier les séquences, généralement de plus proches ou similaires aux plus éloignées
3. Cette liste est parcourue itérativement pour construire l'alignement multiple final, c'est-à-dire que les deux plus proches séquences sont alignées (itération 1). A partir de cet alignement, on calcule un « profil », qui est en quelque sorte une séquence consensus, puis on aligne la troisième séquence profil (itération 2). Un nouveau profil est calculé avec ces trois séquences, et la quatrième séquence est alignées.

DIALIGN

EXTRAIT DU LIVRE « BIOINFORMATIQUE PRINCIPES D'UTILISATION DES OUTILS »

Comme nous l'avons indiqué dans la fiche 20, DIALIGN est un programme d'alignement multiple qui repose sur une méthode très différente de celle employée par ClustalW. Il s'agit ici d'un algorithme itératif utilisant une approche locale pour calculer les alignements. Dans cette même fiche, nous avons vu le principe de fonctionnement d'un algorithme itératif. Dans un premier temps, on compare toutes les paires de séquences. Dans le cas de DIALIGN, cette étape consiste à rechercher tous les fragments pour ne retenir que ceux qui sont compatibles. Un fragment consiste en une suite (la plus grande possible) de résidus (bases, AA) consécutifs, similaires entre deux séquences. Selon cette définition, on constate qu'un fragment ne peut pas contenir d'indels. Ensuite, on ne retient que ceux qui sont compatibles, c'est-à-dire des fragments qui ne se croisent pas.



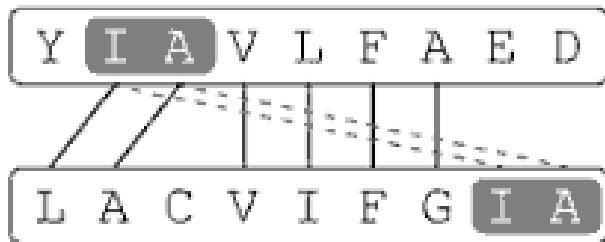


Figure 23.1. Exemple d'un fragment DIALIGN commun à deux séquences.

Sur la figure 23.1, par exemple, les deux séquences **IAVLFA** et **LACVIFG** sont similaires mais de longueurs différentes; ce ne sont donc pas des fragments au sens de la définition ci-avant. En revanche, **IA/IA** et **VLFA/VIFG** sont des fragments, mais **IA** n'est pas compatible puisqu'à gauche dans un cas et à droite dans l'autre. Notez que les fragments de DIALIGN s'appellent des mots dans d'autres programmes comme BLAST ou FASTA.

Après avoir repéré tous les fragments compatibles, les séquences sont triées en fonction du nombre total de fragments communs entre elles. La dernière étape de l'algorithme consiste à aligner itérativement les séquences, c'est-à-dire de la première à la dernière séquence de la liste. À chaque itération, des insertions sont ajoutées de manière à ce que les différents résidus soient correctement alignés. La figure 23.2 montre le cas d'un alignement de trois séquences protéiques.

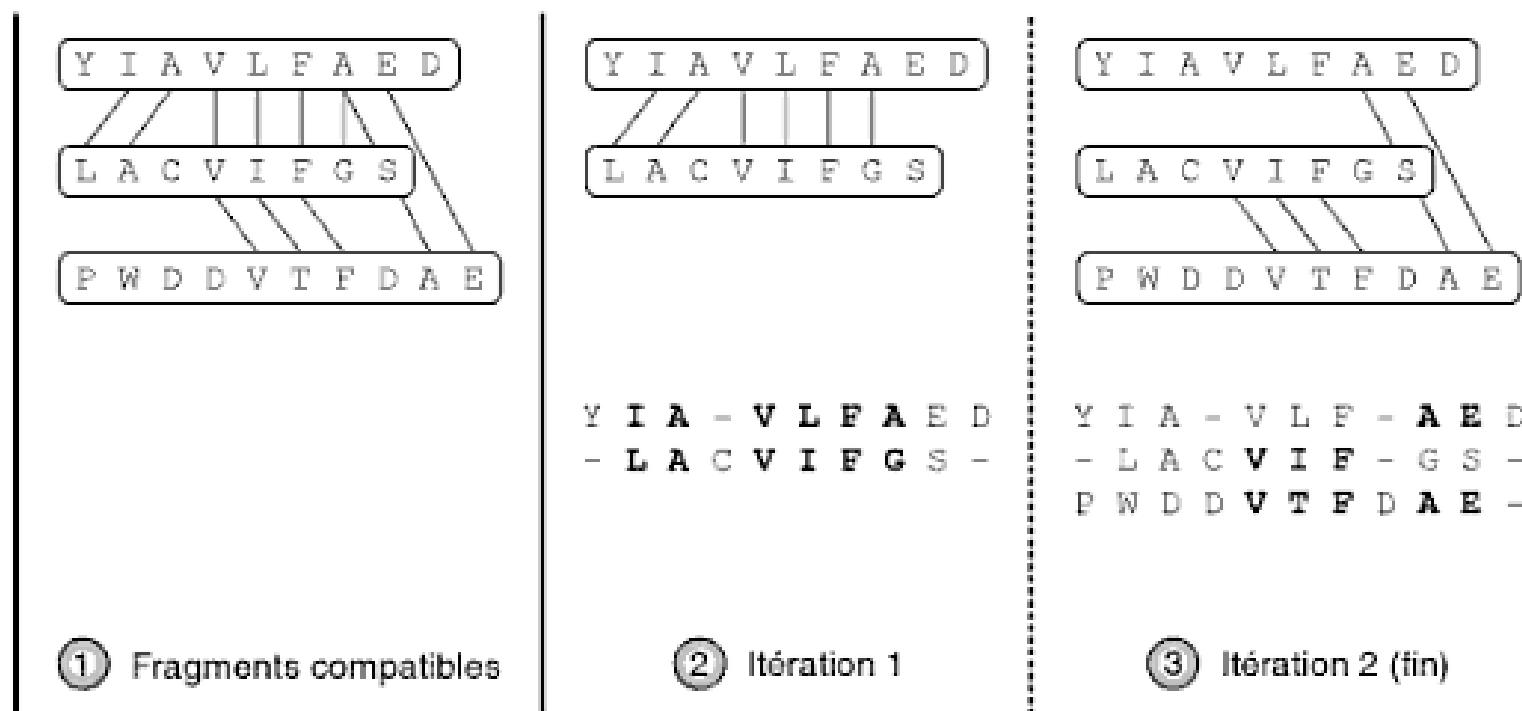


Figure 23.2. Les trois étapes de la méthode DIALIGN.

COMPARAISON DIALIGN CLUSTALW

SYI_ECOLI/1-938	51	filhdgPPYA NGSI HIGH SV NKILKDIIVK SKGLSGYDSP YVPGWDCHGL
SYL_ECOLI/1-860	42	-----PYP SGRL HMGH VR NYTIGDVIAR YQRMLGKNVL QPIGNDAFGL
SYM_ECOLI/1-677	15	-----PYA NGSI HLGH ML EHIQADVWVR YQRMRGHEVN FICADDAHGT
SYV_ECOLI/1-951	41	-----PPNV TGSL HMGH AF QQTIMDTMIR YQRMQGKNTL WQVGTDHAGI
		0000000555 7999 9999 78 8888888888 8888886665 5555554444
	
SYI_ECOLI/1-938	589	VLTHGFTVDG QGR KMSKS IG NTVSPQDVNN K----- -----
SYL_ECOLI/1-860	617	----- -MS KMSKS KN NGIDPQVMVE R----- -----
SYM_ECOLI/1-677	331	----- -GA KMSKS RG TFIKASTWLN H----- -----
SYV_ECOLI/1-951	541	VYMTGLIRDD EGQ KMSKS KG NVIDPLDMVD gislpellek rtgnmmqpql 1111111111 122 4444444 4443333333 2000000000 0000000000

Figure 23.3. Exemple de sortie d'alignement effectué sur le site Web de DIALIGN.

Toutes ces tRNA synthétases sont de type I et ont la particularité d'être globalement similaires, avec cependant la présence de longues insertions pour *SYL_ECOLI*. Par ailleurs, ces séquences sont connues pour posséder deux motifs caractéristiques propres aux tRNA synthétases de type I, **HIGH** et **KMSKS**.

COMPARAISON DIALIGN CLUSTALW

SYL_ECOLI/1-860 1 -----MQEQRPEEIESKVQLHWDERKTFEVTEDESKEKYCLSMPLYPGRLHMGHVRNYTIGDV 61
SYV_ECOLI/1-951 1 -----NEKTYNPQDIEQPLYEHWEKQGYFKPN--GDESGESFCIMIPPPNVTGSLHMGHAFQQTIMDT 61
SYI_ECOLI/1-938 1 MSDYKSTLNLPETGFPMRGDLAKREPGMLARWTDDLYGIIRAAKKGGKTFILHGPPYANGSI**KMSKS**HGVNKILEDI 77
SYM_ECOLI/1-677 1 -----MTQVAKKILVTCALPYANGSI**KMSKS**HLGHMLEHIQADV 34

SYL_ECOLI/1-860 335 VMAVEGHDQRDYEFA SKYGLNIKPVILAADGSEPDLSQQALTEKGVLFNSGEFNGLDHEAAFNAIADKLTAMGVGER 411
SYV_ECOLI/1-951 343 AVVAAVDAGLLEEIKPHDLTVFYGDROGVVIEPMILTDQWYVRAVLAKPAVEAVENGDIQFVPKQYENHYFSWMR- 418
SYI_ECOLI/1-938 382 IVVALLQEKGALLHVEMQHSYPCOCWRHKTPPIFRATPQWFVSMQKGLRAQSLSKEIKGVQWNPDMQGQARIESMVAN 458
SYM_ECOLI/1-677 286 STAELYHFIG-KDIVYFHSLFWPAMLEGSN---TRKPSNLFVNGYVTVNGA**KMSKS**RGTFIKASTMLNHFADSLR- 357

SYL_ECOLI/1-860 412 KVNYRLRDNCVERQRYNGAPIPMVTLEDGTVMPPTDQ---QLPVILPEDVVMDCITSPIKADPENAKTTVN---GMP 482
SYV_ECOLI/1-951 419 ----DIQDWCISRQLNWGERIPANYDEAGNVYVGRNEDEVRKENNIGADVVLRQDEDVLDTMPSALNTFS---TLG 488
SYI_ECOLI/1-938 459 ----RPDNCISRQRTWGVPMSELVHKDTEELH-----PRTLELNEEVAKRVEVDGIQANNDLDAKEILGDEADQ 523
SYM_ECOLI/1-677 358 -----YYTAKLSSRIDOID-----LNLEDVQRVNA DIVVNKVVNLAERNACFINKRF 405

SYL_ECOLI/1-860 483 ALRETDTFDTFMESSWYYARYTCPQYKEG-----NLDSEAANYNLPVD-IYIGGIEHAIMHLLYFRFFFKLMRD 550
SYV_ECOLI/1-951 489 NPENTDALRQFHPTSVMSVGFDIIFFWIARMIMMTMHFIKDENGKPVFPHTVYNTGLIIRDDEGQ**KMSKS**KGIVIDP 565
SYI_ECOLI/1-938 524 YVKVPTDLDVWPDSGSTHSSVVDVRPEFAG--HAADHYLEGSDQHRGMFSSLMISTAHKGXAPYRQLTHGFTVDG 598
SYM_ECOLI/1-677 406 DGVIASELADPQLYKTFDAAEVIG----- 430

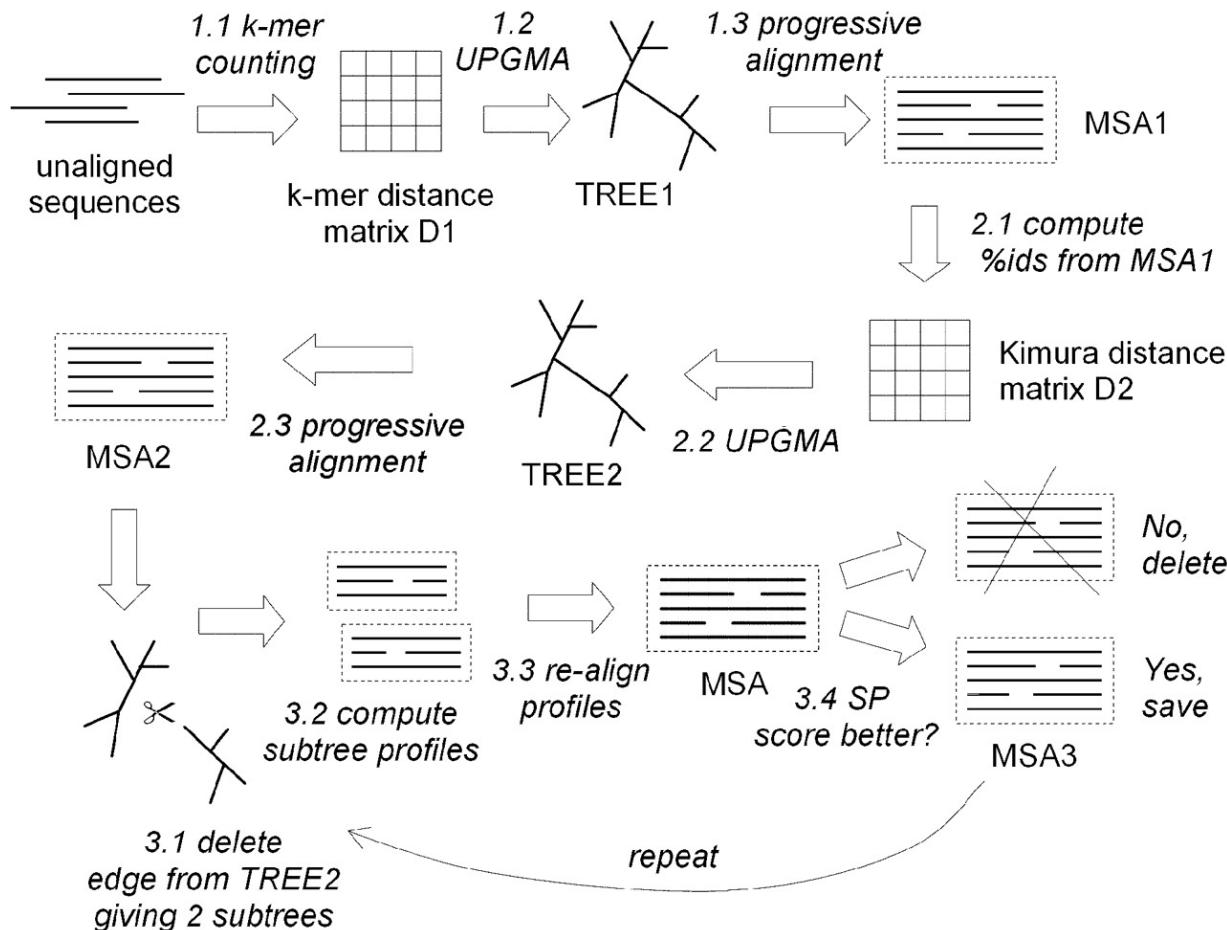
SYL_ECOLI/1-860 551 AGMVNSDEPAKQLLCQG--MVLADAFYYVGENGERNNV-----FVDAIVERDEKGRIVK 603
SYV_ECOLI/1-951 566 LDIVDGISLPELLERKRTGNMMQPQLADKIRKRTKEQFPNGIEPHGTDALRFTLAALASTGRDINNDMKRLEGYRNFC 642
SYI_ECOLI/1-938 599 QGR**KMSKS**ICNTVSPQD-----VNNKLGADILRLNVASTDYTG-----EMAVSDEILKRAADSY 652
SYM_ECOLI/1-677 431 ----- 445

SYL_ECOLI/1-860 604 AKDAAGHELVYTGMS**KMSKS**KNNGI-DPQVNVERYGADTVRLFNMMPASPADMTLEMQESG----VEGANRFLKRVWK 675
SYV_ECOLI/1-951 643 NKLWNASRFVLMNTEGQDCGFNGGE-MTLSLADRNLWIAEPNQTIKAYREALDSFRFDIAAGILYEFTNQFCDWYLE 718
SYI_ECOLI/1-938 653 RRIRNTARFLLANLNGFDPAKDNVKPEEMVVLDRWAVGCAKAAQEDILKAYEAYDFHEVVQRIMRFCSENGSFYLD 729
SYM_ECOLI/1-677 446 MALADLANRYVDEQAPWVVAKQEGRDAQIACSMGINLPRVLMTYLKPVLPKLTERAEAFINTELTWDGJQQPLLG 522

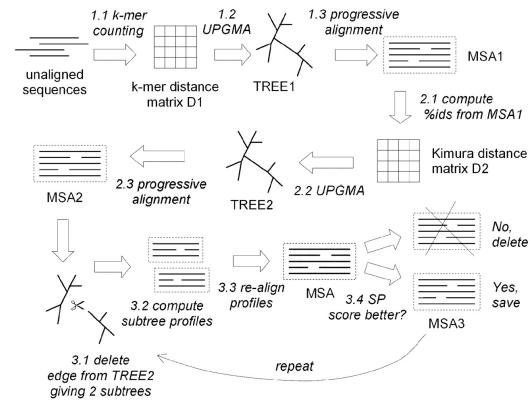
Figure 23.4. Alignement effectué par ClustalW montrant le non-alignement du motif **KMSKS**.

MUSCLE: MÉTHODE ITÉRATIVE

- MUSCLE: multiple sequence alignment by log-expectation



MUSCLE: MÉTHODE ITÉRATIVE



1. Un arbre guide est construit à partir de la matrice de distance qui a été calculé par UPGMA ou NJ (Neighbor Joining), et une racine est identifiée.

MUSCLE uses a much faster, but somewhat more approximate, method to compute distances: it counts the number of short sub-sequences (known as *k*-mers, *k*-tuples or words) that two sequences have in common, without constructing an alignment. This is typically around 3,000 times faster than CLUSTALW's method, but the trees will generally be less accurate. We call this step "*k*-mer clustering".

Un alignement progressif est construit en suivant l'ordre des branches de l'arbre guide, produisant un alignement multiple de l'ensemble des séquences.

2. La deuxième étape tente d'améliorer l'arbre guide et construit un nouvel alignement progressif selon cet arbre. Cette étape peut être répétée (méthode itérative).
3. Troisième étape : raffinement



MAFFT: MÉTHODE ITÉRATIVE (KATOH ET AL, 2002) D'ALIGNEMENT MULTIPLE

Principe

- Le logiciel MAFFT fait incontestablement partie des programmes de nouvelle génération qui tirent profit d'un certain nombre d'avancées réalisées dans ce domaine. MAFFT a été écrit dans le but explicite d'accélérer considérablement le processus d'alignement multiple, permettant ainsi d'aligner un grand nombre de séquences sans pour autant sacrifier à la qualité de l'alignement.
- On peut décomposer le processus en trois grandes étapes :
- i) chaque acide aminé est décrit par sa polarité et son volume, et les séquences sont réécrites dans ce système. Chaque suite de lettres (chaque séquence) est donc transformée en une suite de valeurs numériques. Les séquences nucléotidiques sont recodées en utilisant les fréquences locales des quatre bases. Puis les segments de similarité entre chaque paire de séquences sont repérés au moyen d'un algorithme de calcul appelé transformée de Fourier rapide, ou Fast Fourier Transform (FFT) en anglais. Les paires de séquences sont ensuite alignées sur la base de ces segments de similarité (cf. DIALIGN). Sauf pour des séquences très divergentes, ce procédé permet d'aligner toutes les paires de séquences environ dix fois plus vite que ClustalW;

MAFFT: ALIGNEMENT MULTIPLE

- ii) un arbre de guidage (cf. ClustalW, et MUSCLE) est ensuite calculé à partir des alignements précédents. Ici, le calcul des distances entre les séquences est simplifié et accéléré en recodant les séquences protéiques dans un alphabet réduit à six lettres: par exemple, les acides aminés hydrophobes I, L, M et V forment un seul groupe, de même que D, E, N et Q, etc. La distance entre deux séquences est estimée à partir du nombre de mots de six lettres que ces séquences partagent dans ce nouvel alphabet (cf. MUSCLE);
- iii) les séquences sont ensuite alignées progressivement en suivant l'ordre indiqué par l'arbre de guidage.

Plusieurs programmes sont proposés sur la page de garde du site Web de MAFFT. Ce que nous venons de décrire correspond à l'option FFT-NS-1. Ce nom un peu barbare signific *Fast Fourier Transform-New Scoring matrix-1 step*.



MAFFT: ALIGNEMENT MULTIPLE

Contrairement à ClustalW mais comme MUSCLE, MAFFT peut optionnellement procéder à un deuxième passage. Dans ce dernier, l'alignement réalisé précédemment sert à recalculer la distance entre chaque paire de séquence, un nouvel arbre de guidage et un nouvel alignement multiple. Cette option s'appelle FFT-NS-2.

De manière similaire à MUSCLE, MAFFT peut procéder à un raffinement de l'alignement. Dans ce cas, l'arbre de guidage est scindé en deux, puis les deux moitiés sont réalignées. On recommence ainsi tant que la procédure conduit à un gain dans le score d'alignement $\Delta S_i > \Delta S_{i-1}$. On procède alors à un nombre i d'itérations, i étant inconnu *a priori*. Cette option porte le nom de FFT-NS-i. On peut, sur la page de garde de MAFFT, limiter à deux le nombre d'itérations (« two cycles only »).



MAFFT: ALIGNEMENT MULTIPLE

Il faut par ailleurs noter que le site de MAFFT propose des programmes fondés non pas sur la transformée de Fourier rapide, mais sur l'algorithme de programmation dynamique . Ainsi le programme nommé G-INS-i aligne les paires de séquences suivant l'algorithme global de Needleman-Wunsch, comme ClustalW, calcule un arbre de guidage, aligne toutes les séquences suivant cet arbre et procède enfin à un raffinement de l'alignement comme décrit ci-avant. Les programmes L-INS-i et E-INS-i procèdent de la même façon, mais avec l'algorithme d'alignement local de Smith-Waterman. Bien entendu, ces programmes, nettement plus lents, ne conviennent pas pour un grand nombre de séquences. Enfin, le programme Q-INS-i est spécifiquement dédié à l'alignement de séquences d'ARN .



MAFFT : RÉSUMÉ

- Recodage des séquences en valeurs numériques (polarité et volume pour les acides aminés) (alphabet à 20 lettres => 6)
- Identification des segments locaux de similarité par Fast Fourier Transform
- Alignement des paires de séquences sur base des segments
- Construction d'un arbre de guide. Distance déduite du nombre de 6-mers partagés
- Alignement multiple progressif selon l'arbre, cf ClustalW
- 2^{ème} passage : nouvelle distance par paire => arbre => alignement2
- Raffinement itératif de l'alignement multiple (cf MUSCLE) : alignement coupé en 2, et moitiés ré-alignées...

MAFFT

NOMBREUSES OPTIONS DE MAFFT

1. Mode basique, rapide — juste progressif

- a) FFTNS1 (fftns --retree 1)
- b) FFTNS2 (fftns) (same as mafft --retree 2)

OK jusqu'à 1 000 séquences facilement alignables

2. Mode intermédiaire — progressif + itérations

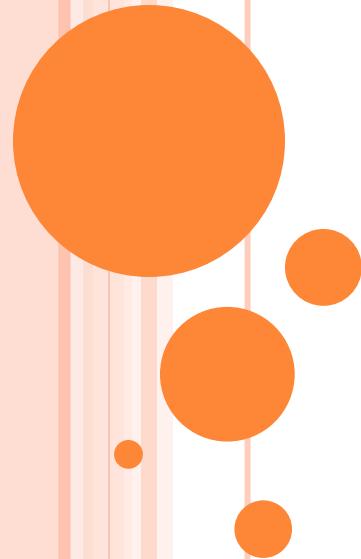
- a) FFTNSI (fftnsi) default two cycles, or e.g. fftnsi --maxiterate 1000
- b) NWNSI (nwnsi) same as FFTNSI, but no FFT, Needleman Wunsch only.

OK entre 100 et 500 séquences

3. Mode avancé — progressif + itérations + consistence (cf T-Coffee)

- a) EINSI (einsi) S-W (plusieurs rég. sim même ordre)
- b) LINSI (linsi) S-W stricte (1 rég. sim)
- c) GINSI (linsi) global N-W





AUTRES MÉTHODES

M-COFFEE: UN MÉTA-ALIGNEUR

T-Coffee

A collection of tools for Computing, Evaluating and Manipulating Multiple Alignments of DNA, RNA, Protein Sequences and Structures

Alignment

[T-Coffee](#) Aligns DNA, RNA or Proteins using the default T-Coffee >> Cite

[M-Coffee](#) Aligns DNA, RNA or Proteins by combining the output of popular aligners >> Cite

[R-Coffee](#) Aligns RNA sequences using predicted secondary structures >> Cite

[Expresso](#) Aligns protein sequences using structural information >> Cite

[PSI-Coffee](#) Aligns distantly related proteins using homology extension (slow and accurate) >> Cite

[TM-Coffee](#) Aligns transmembrane proteins using homology extension NEW >> Cite

[Pro-Coffee](#) Aligns homologous promoter regions NEW >> Cite

[Accurate](#) Automatically combine the most accurate modes for DNA, RNA and Proteins (experimental!)

[Combine](#) Combines two (or more) multiple sequence alignments into a single one >> Cite

Combine the output of your favorite alignment methods (Clustal, Mafft, Probcons, Muscle, etc.) into one unique alignment. **15** methods in all.

Evaluation

[Core](#) Evaluates your Alignment and outputs a Colored version indicating the local reliability. >> Cite

[iRMSD-APDB](#) Evaluates Multiple Sequence Alignment using structural information with APDB and iRMSD. >> Cite

[T-RMSD](#) Allows fine-grained structural clustering of a given group of related protein domains NEW >> Cite

Other

[Advanced](#) Run your alignment using full featured T-Coffee options. >> Cite



AQUA: UN AUTRE MÉTA-ALIGNEUR ALIGNMENT QUALITY ASSESSMENT

- Attention pas de serveur web !!
- http://bioinformatics.oxfordjournals.org/content/early/2009/11/19/bioinformatics.bt_p651.full.pdf
- Dans la méthode AQUA, seulement deux programmes d'alignement multiples (**Muscle** et **MAFFT**) sont implémentés mais une **méthode d'édition** (RASCAL) et une méthode d'**évaluation** de l'alignement (norMD) sont également implémentées.
- Les alignements multiples sont dans un premier temps générés par Muscle et MAFFT puis édités par RASCAL.
- La qualité de l'alignement est estimée grâce au score norMD (Thompson, et al., 2001).
- Ce score prend en compte la conservation des acides aminés dans une même colonne de l'alignement, la similarité des acides aminés mais également les caractéristiques propres des séquences brutes (nombre, longueur et similarité des séquences).
- L'alignement édité qui aura le meilleur score norMD sera celui qui sera choisi comme alignement final.



MULTIPLE SEQUENCE ALIGNERS

T-Coffee	http://www.tcoffee.org"
ClustalW	ftp://www.ebi.ac.uk/pub/clustalw"
MAFFT	http://www.biophys.kyoto-u.ac.jp/~katoh/programs/align/mafft/"
Dalign-tx	http://dalign-tx.gobics.de/"/>download
POA	http://www.bioinformatics.ucla.edu/poa/
ProbCons	http://probcons.stanford.edu/
Muscle	http://www.drive5.com/muscle/
PCMA	ftp://iole.swmed.edu/pub/PCMA/
Kalign	http://msa.cgb.ki.se
AMAP	http://bio.math.berkeley.edu/amap/
PRODA	http://bio.math.berkeley.edu/proda/
PRANK	http://www.ebi.ac.uk/goldman-srv/prank/



MULTIPLE RNA SEQUENCE ALIGNERS

Consan	http://selab.janelia.org/software/consan/
T-Lara	https://www.mi.fu-berlin.de/w/LiSA/Lara
STRAL	http://www.biophys.uni-duesseldorf.de/stral/
StemLoc	http://biowiki.org/StemLoc
RNAsampler	http://ural.wustl.edu/~xingxu/RNAsampler/index.html
Murlet	http://murlet.ncrna.org/
M-Locarna	http://www.bioinf.uni-freiburg.de/Software/LocARNA/
MARNA	http://biwww2.informatik.uni-freiburg.de/Software/MARNA/index.html
Foldalign	http://foldalign.kvl.dk/server/index.html



MULTIPLE STRUCTURAL ALIGNERS

FUGUE	http://www-cryst.bioc.cam.ac.uk/fugue/download.html
SAP	http://mathbio.nimr.mrc.ac.uk/wiki/Software
Mustang	http://www.cs.mu.oz.au/~arun/mustang/
TMalign	http://zhang.bioinformatics.ku.edu/TM-align/
DALI	http://ekhidna.biocenter.helsinki.fi/dali_lite/downloads/download.html



QUELLE MÉTHODE UTILISÉE?

	Avantages	Inconvénients
CLUSTALW	Peu d'espace mémoire Ok sur grandes séquences	Précision – Problème si beaucoup de séquences
DIALIGN	Alignements locaux => blocs communs (œur conservé)	Si similarité globale
MAFFT, MUSCLE	Équilibre rapidité/précision Options rapides (- iter) si beaucoup de séquences	
T-Coffee	Précision+	Temps de calcul et espace mémoire



LES PRINCIPAUX OUTILS DE MSA (MULTIPLE SEQUENCE ALIGNMENT)

ClustalW2

ClustalW2 is a general purpose multiple sequence alignment program for DNA or proteins. It produces biologically meaningful multiple sequence alignments of divergent sequences. It calculates the best match for the selected sequences, and lines them up so that the identities, similarities and differences can be seen. Evolutionary relationships can be seen via viewing Cladograms or Phylogenograms.

New users, please read the [FAQ](#).

>> [Download Software](#)

YOUR EMAIL:

Sequence: Interactive: Pair:

KTUP (WORD SIZE): def

WINDOW LENGTH: def percent def

SCORE TYPE: TOPDIAG PAIRGAP

MATRIX: NO END GAP

T-Coffee

T-Coffee is a multiple sequence alignment program. Multiple sequence alignment programs are meant to align a set of sequences previously gathered using other programs such as blast, fast, sw... The main characteristic of T-Coffee is that it will allow you to combine results obtained with several alignment methods. For instance if you have an alignment coming from ClustalW2, an other alignment coming from Dialign, and a structural alignment of some of your sequences, T-Coffee will combine all that information and produce a new multiple sequence having the best agreement with all these methods. By default, T-Coffee will compare all your sequences two by two, producing a global alignment and a series of local alignments (using lalign). The program will then combine all these alignments into a multiple alignment.

>> [Download Software](#)

EMAIL: RESULTS: RUN NAME: Sequence MATRIX: ORDER:

Enter or Paste a set of Sequences in any supported format: Help

Upload:

Progressif, alignement local

Upload a file: Parcourir... Run Reset

MUSCLE

MUSCLE stands for MULTiple Sequence Comparison by Log-Expectation. MUSCLE is claimed to achieve both better average accuracy and better speed than [ClustalW2](#) or [T-Coffee](#), depending on the chosen options.

[Download Software](#)

MUSCLE
multiple sequence alignment method
JOINTS: fast and time and space
SUBOPTIM: complexity

RESULTS: SEARCH TITLE: YOUR EMAIL:
OUTPUT FORMAT: OUTPUT TREE: OUTPUT ORDER:

Enter or Paste a set of Sequences in any supported format: Help

itératif, alignement global

MAFFT

MAFFT (Multiple Alignment using Fast Fourier Transform) is a high speed multiple sequence alignment program.

YOUR EMAIL: RESULTS: MATRIX (PROTEIN ONLY): GAP OPEN: 1.53 GAP EXTENSION: 0.123
TREE REBUILDING NUMBER: 1 MAXITERATE: 0 PERFORM FFTS:
REORDER: CLUSTAL OUTPUT:

Enter or Paste a set of DNA or protein sequences in fasta format: Help

Progressif, alignement global avec transformation de Fourier

Upload a file: Parcourir... Run Reset

LES PRINCIPAUX OUTILS DE MSA

Program	Function	Input	Output	Speed	Current Limitations
Blast 2.2.18	Sequence searching	Raw, FASTA	FASTA	Fast	None
MUSCLE 3.7	Multiple alignment	FASTA, EMBL/Uniprot, GenBank, PAUP*/Nexus	FASTA, Clustal, PHYLIP	Fast	<200 nucleic sequences, <6000 sites <200 protein sequences, <2000 sites
T-Coffee 6.85	Multiple alignment	FASTA, EMBL/Uniprot, GenBank, PAUP*/Nexus	FASTA, Clustal, PHYLIP	Very slow	<50 nucleic sequences, <2000 sites <50 protein sequences, <2000 sites
3D-Coffee 6.85	Multiple alignment using structural information	FASTA, EMBL/Uniprot, GenBank, PAUP*/Nexus	FASTA, Clustal, PHYLIP	Very slow	<50 nucleic sequences, <2000 sites <50 protein sequences, <2000 sites
ClustalW 2.0.3	Multiple alignment	FASTA, EMBL/Uniprot, GenBank, PAUP*/Nexus	FASTA, Clustal, PHYLIP	Fast	<200 nucleic sequences, <4000 sites <200 protein sequences, <2000 sites
ProbCons 1.12	Multiple alignment	FASTA, EMBL/Uniprot, GenBank, PAUP*/Nexus	FASTA, Clustal, PHYLIP	Slow	<200 nucleic sequences, <6000 sites <200 protein sequences, <2000 sites

QUELLE MÉTHODE UTILISÉE?

- Cela dépend du type de séquences à aligner . . .
- Plus les séquences sont divergentes, moins le résultat est fiable
- Attention à la « twilight zone »
La zone d'ombre est la zone à partir de laquelle il devient difficile de dire si deux protéines sont similaires ou si elles se sont alignées par chance. >10, 20, 30% identité
- Pas de méthode universelle
- Pas de confiance aveugle vis-à-vis du résultat obtenu



COMMENT AVOIR UN BON ALIGNEMENT ?

- séquences divergentes
 - le minimum d'indels
 - des *blocks* séparés par des indels
 - des résidus conservés dans les *blocks*
 - *Similarité, hydropathie*
 - jugement personnel et modifications manuelles
-
- Attention aux variations du nombre de domaines !



EXEMPLE: DOMAIN SH3

- *SH3 (Src homology 3) domains are often indicative of a protein involved in signal transduction related to cytoskeletal organization. The SH3 domain has a characteristic fold which consists of five or six beta-strands arranged as two tightly packed anti-parallel beta sheets. The linker regions may contain short helices.*
- Prosite PS50002

Séquences à aligner	longueur
-----	-----
1aboA	P00520
1yctsB	P04637
1pht	P27986
1ihvA	P00383
1vie	P12497

- séquences courtes
- similarité faible (< 25%) et diffuse



EXEMPLE: DOMAIN SH3

- Véritable alignement basé sur l'alignement des éléments de structure secondaire

1aboA	-NLFVALYDfvasgdntlsitkGEKLRVLgynhn-----
1ycsB	kGVIYALWDyepqnndelpmkeGDCMTIIhrede-----
1pht	gYQYRALYDykkereeedidlhlGDILTVNkgslvalgfsd
1ihvA	-NFRVYYRDsrd-----pvwkGPAKLLWkg-----
1vie	-drvrvkksga-----awqGQIVGWYctnlt-----
1aboA	-----gEWCEAQt--kngqGWVPSNYITPVN-----
1ycsB	-----deiEWWWARl--ndkeGYVPRNLLGLYP-----
1pht	ggearpeeiGWLNGYnettgerGDFPGTYVEYIGrkkisp
1ihvA	-----eGAVVIQd--nsdiKVVPRRKAKIIRd-----
1vie	-----peGYAVESeahpgsvQIYPVAALERIN-----

Alignment fourni par ClustalW

1aboA	n-LFVALYDFVASGDNTLSITK GEKLRLV I-----
1ycsB	kgVIYALWDYEPQNDDELPMKE GDCMTI Ihr----EDEDEI-----
1pht	gyQYRALYDYKKEREEDIDLHL GDILTVN KGSLVALGFSDgqearpee
1ihvA	-- NFRV --YYRD SRDPVWK GPAKLLW KGE GAVVIQ DNSDI-----
1vie	----- -----DRVRKKSGaa-W----- QQQI -----
1aboA	----GYNhng EWCEA QTNGQ GWV -----PSNYIt-----VN
1ycsB	----- EWWWA RLNDKE GYV -----PRNLLgLYP-----
1pht	gwlnGYN -----ETTGER GDF -----PGTYV-EYigRKKIsp--
1ihvA	----- Kv -----V-----PRr-----KAKIIRd-
1vie	-----VGWYCTNLTP EGYAv seahPGS VQ -IYPv-AALERIN

1aboA	-NLFV- ALYDF VASGDNTLSITK GEKL RV-----LGYNHNG
1ycsB	KGVIY -ALWDYEPQNDDELPMKE GDCMTI -----IHREDED
1pht	-GYQYRALYDYKKEREEDIDLHL GDILTVN KGSLVALGFSDGQ
1ihvA	----- NFRVYYRD SRD--PVWK GPAKLL -----WKGE G
1vie	-----DRVRKKSG--AAWQ GQIVGW -----YCTNL
1aboA	----- EWCEA --QTNGQ GWV PSNYItPVN-----
1ycsB	EI----- EWWA --RLNDKE GYV PRNLLGLYP-----
1pht	EARPEEI GWLNGY NETTGER GDFPGTYVEYIGRKKISP
1ihvA	----- AVVIQ --DNSDI KVVPRRKAKIIRD -----
1vie	TP---- EGYAVE SEAHPGS VQ IYPVAALERIN-----

Alignment fourni par Dialign2

LES PARAMÈTRES SONT MODIFIABLES !

ClustalW2

ClustalW2 is a general purpose multiple sequence alignment program for DNA or proteins. It produces biologically meaningful multiple sequence alignments of divergent sequences. It calculates the best match for the selected sequences, and lines them up so that the identities, similarities and differences can be seen. Evolutionary relationships can be seen via viewing Cladograms or Phylogenograms.

[New users, please read the FAQ.](#)

[Download Software](#)



YOUR EMAIL	ALIGNMENT TITLE	RESULTS	ALIGNMENT		
<input type="text"/>	Sequence	interactive	full		
KTUP (WORD SIZE)	WINDOW LENGTH	SCORE TYPE	TOPDIAG		
def	def	percent	def		
MATRIX	GAP OPEN	NO END GAPS	GAP EXTENSION		
pam	def	yes	def		
ITERATION		NUMITER			
none		1			
PHYLOGENETIC TREE					
OUTPUT FORMAT	OUTPUT ORDER	TREE TYPE	CORRECT DIST.	IGNORE GAPS	CLUSTERING
aln w/numbers	aligned	none	off	off	NJ

Enter or paste a set of sequences in any supported format: Help

Upload a file: Parcourir...

PSI-BLAST

BLAST stands for Basic Local Alignment Search Tool. The emphasis of this tool is to find regions of sequence similarity, which will yield functional and evolutionary clues about the structure and function of your novel sequence. Position specific iterative BLAST (PSI-BLAST) refers to a feature of BLAST 2.0 in which a profile is automatically constructed from the first set of BLAST alignments. PSI-BLAST is similar to [NCBI-BLAST2](#) except that it uses position-specific scoring matrices derived during the search, this tool is used to detect distant evolutionary relationships.

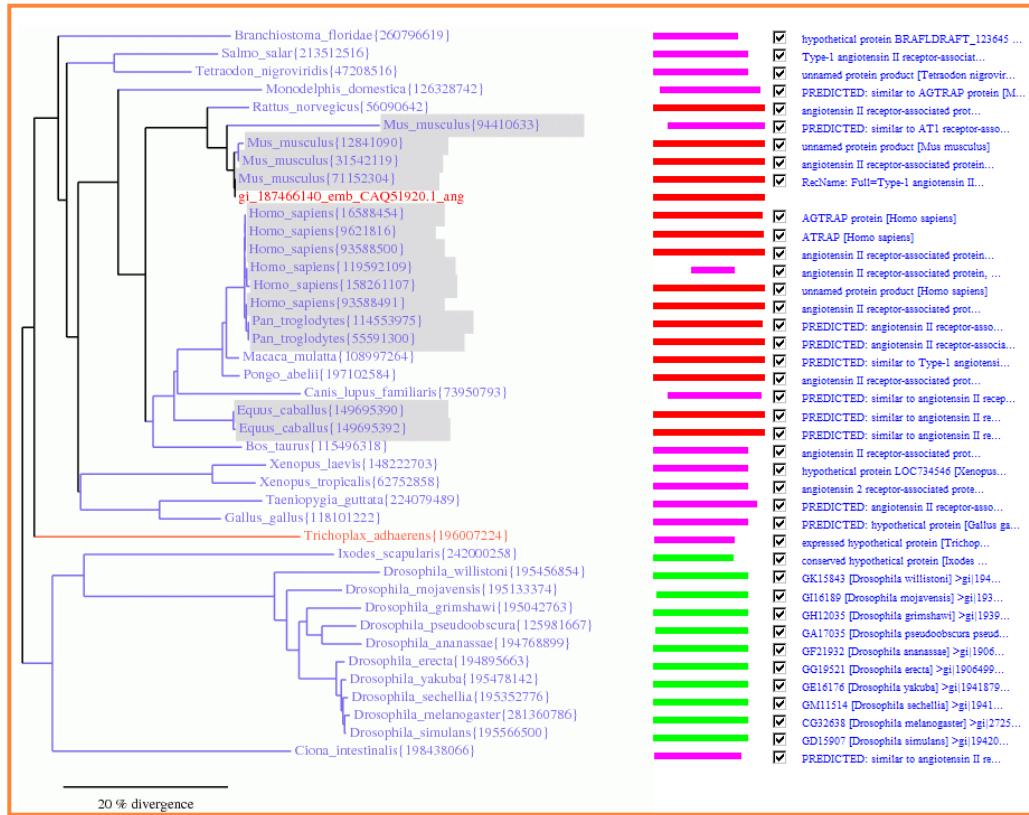
DATABASE	RESULTS	SEARCH TITLE	YOUR EMAIL
UniProt Knowledgebase	interactive	Sequence	<input type="text"/>
MATRIX	OPENGAP	EXTENDGAP	FILTER
BLOSUM62	default	default	false
BLOSUM45	DROPOFF	FINAL DROPOFF	
BLOSUM62	default	default	
BLOSUM80	ALIGNMENTS	SEQUENCE RANGE	ALIGN VIEWS
PAM30	500	START-END	pairwise
PAM70	FORMAT		default

UPLOAD A CHECKPOINT FILE (ASN.1 Binary Format) Parcourir...

Enter or Paste a PROTEIN Sequence in any format: Help

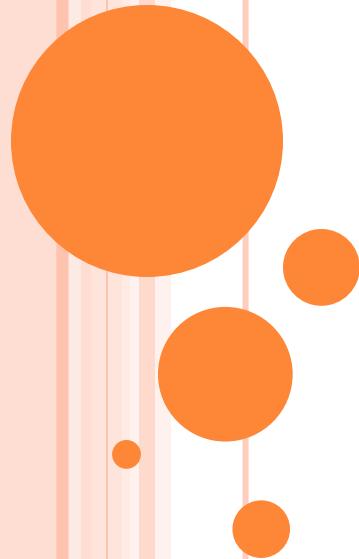
Upload a file: Parcourir...

ATTENTION AUX ARBRES EN SORTIE D'ALIGNEMENT!!



- hypothetical protein BRAFLDRAFT_123645 ...
- Type-I angiotensin II receptor-associat...
- unnamed protein product [Tetradon nigrovir...]
- PREDICTED: similar to AGTRAP protein [H...]
- angiotensin II receptor-associated prot...
- PREDICTED: similar to ATI receptor-asso...
- unnamed protein product [Mus musculus]
- angiotensin II receptor-associated protein...
- ReName Full=Type-I angiotensin II ...
- AGTRAP protein [Homo sapiens]
- ATRAP [Homo sapiens]
- angiotensin II receptor-associated protein...
- angiotensin II receptor-associated protein, ...
- unnamed protein product [Homo sapiens]
- angiotensin II receptor-associated prot...
- PREDICTED: angiotensin II receptor-asso...
- PREDICTED: angiotensin II receptor-asso...
- PREDICTED: similar to Type-I angiotensi...
- angiotensin II receptor-associated prot...
- PREDICTED: similar to angiotensin II recep...
- PREDICTED: similar to angiotensin II re...
- PREDICTED: similar to angiotensin II re...
- angiotensin II receptor-associated prot...
- hypothetical protein LOC734546 [Xenopus...]
- angiotensin 2 receptor-associated prot...
- PREDICTED: angiotensin II receptor-asso...
- PREDICTED: hypothetical protein [Gallus ga...
- expressed hypothetical protein [Trichop...
- conserved hypothetical protein [Ixodes ...]
- GK15843 [Drosophila willistoni] >gi|194...
- GH16189 [Drosophila mojavensis] >gi|193...
- GH12035 [Drosophila grimshawi] >gi|1939...
- GA17035 [Drosophila pseudoobscura pseud...
- GF21992 [Drosophila ananassae] >gi|1906...
- GG19521 [Drosophila erecta] >gi|1904499...
- GE16176 [Drosophila yakuba] >gi|1941879...
- GM11514 [Drosophila sechellia] >gi|1941...
- CG32658 [Drosophila melanogaster] >gi|2725...
- GD15907 [Drosophila simulans] >gi|19420...
- PREDICTED: similar to angiotensin II re...

ATTENTION !
Ce sont des arbres
de similarités et
non des arbres
phylogénétiques



APPLICATION DES ALIGNEMENTS MULTIPLES : L'IDENTIFICATION DE MOTIFS OU DE PATTERNS

UN PEU DE VOCABULAIRE

Familles, domaines, motifs, pattern, etc...

- **domaine protéique**: unité structurale (et fonctionnelle) indépendante, évolutivement conservée (doigt de zinc, boucle,...)
- **motifs protéiques**: plus courts
 - site de modification post-traductionnelle
 - site de liaison (ADN, métal,...)
 - site actif d'enzyme
- **famille protéique**: ensemble de protéines évolutivement reliées par un ou plusieurs domaines protéiques communs



DOMAINE CONSERVÉ

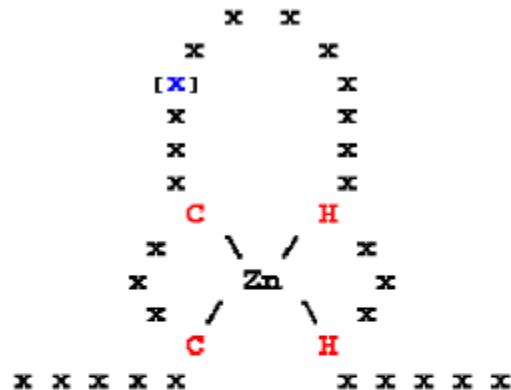
- Exemple doigt de zinc

TYY1_HUMAN/383-407	YVCPF-DG CN --KK F AQSTNLKSHILT--H
YKQ8_CAEEL/78-102	YK CT --V CR --KD I SSSESRLTHMF KQ -HH
BASO_HUMAN/719-742	F QCD --I CK --KT F KNACSVKI H HKN--MH
ZG29_XENLA/62-84	FV CT --V CG --KTY K HGLNT H LHS--H
P43_XENBO/106-130	L K CSV-PG CK --RSFRKKRALRIHVSE--H
IKAR_MOUSE/488-512	FE CN --M CG --Y H SQDRYEFSSHITRG-EH
Q92610/1043-1069	YT CG --Y C TEDSPSFPRPSLLES H ISL--MH
TRA1_CAEEL/306-331	YK CE F-AD CE --KAFSNASDRAK H QNR--T H
ZN10_HUMAN/383-405	YK CN --Q CG --IIFSQNSPFIV H QIA--H
GLI1_XENLA/283-310	FV CHW -Q D CSRELRP F KA Q YMLVV H MRR--H
XFIN_XENLA/276-298	FR CS --E C S--RS F THNSDLTAHMRK--H
TF3A_BUFAM/72-97	CK CET -EN CN --LAFTTASNMR L H F KR--AH
ZG58_XENLA/174-196	FV CT --E CN --LSFAGLANLRS H QHL--H
P43_XENBO/163-187	YR CSY -ED CQ --TVSPTWTALQT H LKK--H
TSH_DROME/354-378	FRC V --W CK --QS F PTLEALT H M KDS -K H
ZN76_HUMAN/165-189	FRC GY -KG CG --RLYTTAHHLK V HERA--H
TF3A_BUFAM/219-244	YRC P R-EN CD --RTYTT K FNLKSHILT--FH
SUHW_DROAN/349-373	YACK--I CG --KD F TRSYHLKR H Q KYS -SC
ZN76_HUMAN/285-309	YT CPE -PH CG --RG F TSATNYKNHVRI--H
SRYC_DROME/469-492	FK CN --Y CP --RD F TFNFPNL K H TRR --RH
EVI1_HUMAN/761-784	YR CK --Y CD --RS F SISSNLQRHVRN--IH
...	

Extrait de Pfam, entrée zf-C2H2

DOMAINE CONSERVÉ

- Exemple doigt de zinc



Motif Prosite:

C-x (2,4)-C-x (3)-[LIVMFYW]-x (8)-H-x (3,5)-H

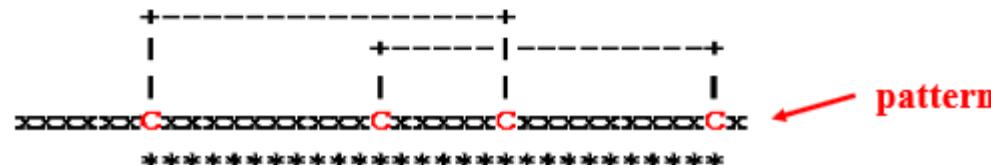


MOTIF (PATTERN)

GUN1_TRIRE/427-455	HWGQ CGGI --GYSGC--K-TCTSGTT CQYSNDYYSQCL
GUX1_TRIRE/481-509	HYGGQ CGGI --GYSGP--T-VCASGTT CQVLNPYYYSQCL
GUX1_PHACH/484-512	QWGGQ CGGI --GYTGS--T-TCASPYT CHVLNPYYYSQCY
GUX2_TRIRE/30-58	VWGGQ CGGI --NWSGP--T- CCASGSTC VYSNDYYSQCL
GUN5_TRIRE/209-237	LYGQ CGGA --GWTGP--T-TCQAPGT CKVQNQWYYSQCL
GUNF_FUSOX/21-49	IWGQ CGGN --GWTGA--T- TCASGLKCEKINDWYYQCV
GUX3_AGABI/24-52	VWGGQ CGGN --GWTGP--T- TCASGSTC VKQNDFYYSQCL
Q01763/473-500	--SQ CGGL --GYAGP--TgVCPSPYT CQALNIYYSQCI
GUX1_PENJA/505-533	DWAQ CGGN --GWTGP--T- TCVSPYTCTKQNDWYYSQCL
GUXC_FUSOX/482-510	QWGGQ CGGQ --NYSGP--T- TCKSPFTC KKINDFYYSQCR
GUX1_HUMGR/493-521	RWQQ CGGI --GFTGP--T-QCEEPYI CTKLNDWYYSQCL
GUX1_NEUCR/484-512	HWAQ CGGI --GFSGP--T- TCPEPYTCA KDHDIYYSQCV
Q9Y894/23-53	PWGQ CGGP --GWTGPttT- CCVTGCTCP VTND-YYSQCL
PSBP_PORPU/26-54	LYEQ CGGI --GFDGV--T- CCSEGLMC MKGPMYYYSQCR
GUNB_FUSOX/29-57	VWAQ CGGQ --NWSGT--P- CCTSGNKCV KLNDFYYSQCR
PSBP_PORPU/69-97	PYGGQ CGGM --NYSGK--T-MCSPGFK CVELNEFFSQCD
GUNK_FUSOX/339-370	AYYQ CGGSKSAYPNGN --L-ACATGSK CVKQNEYYSQCV
PSBP_PORPU/128-156	EYAA CGGE --MFMGA -K-CCKFGLVCYETSGK WSQCR

Extrait de Prosite, entrée PS00562

C-G-G-x(4,7)-G-x(3)-C-x(5)-C-x(3,5)-[NHG]-x-[FYWM]-x(2)-Q-C

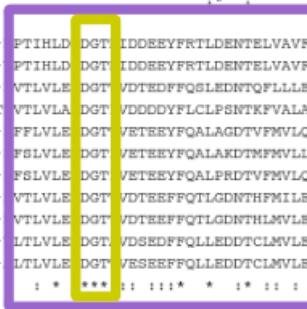


LE PROFIL DE CONSERVATION ISSU DE L'ALIGNEMENT MULTIPLE

CLUSTAL FORMAT LOI T-COFFEE VERSION_2.1.0 (http://www.t-coffee.org), CPU=1, 70 SEC, SCORE=-41, INSEG=11, LEN=304

```
tr|o61464|o61464_DROME      METAANSG-----D-----SKKPFVKVDVTRNIKKAVCA
tr|Q28ZV7|Q28ZV7_DROPS     MPNAMETT-----S-----SKKPFVKVDVTRNIKKAVCA
tr|Q66K97|Q66K97_XENTR     MGQGALDYNALSPKSLSRSVTVNGTSL/TRRVLFPPPLPE--PPQRPFRVNSDRSSKKGIVAA
unk|VIRIT1655|Blast_submission MEVTGDAG-----VPESEGEIN-----TLKPCCLLRNRYSEQNGVAA
Q96AQ7|CIDECK_HUMAN       MEYAMKSLSLLYPKSLSRHVSVRTSVTTQQLLSEPSPKAPRARPCRVSTADRSVRKGIMA
sp|P56198|CIDECK_MOUSE     MDYAMKSLSLLYPRSLSRHVAVSTAVVTQQLVSKPSRETPRARPCRVSTADRKVRKGIMA
tr|Q5X133|Q5X133_RAT       MDYAMKSLSLLYPRSLSRHVAVSTAVVTQQLVSEPSRETPRARPCRVSTADRKVRKGIMA
sp|o60543|CIDECK_HUMAN     MEAARDYAG----ALLRPLTFMGSQTKRLVFTP--LMHMPARPFRVNSNHDRSSRRGVMA
tr|A4FUX1|A4FUX1_BOVIN    METARDCAG----ALLRPLTFMGSQTKRLVFTP--FMHMPARPFRVNSNHDRSSRRGVMA
sp|o70303|CIDECK_MOUSE     ---MEYLSAFNPNGLLRSVSTVSSSELRRVWNNS---APPQRPFRVCIDHKRTVRKGGLTA
sp|Q3T191|CIDECK_BOVIN    ---MEYLSNLDPSSLLRSVSNMSADLGRKVWTS---APPRQRPFRVCIDNKRTTRKGGLTA
```

```
tr|o61464|o61464_DROME      SSLEEEIRSKVAEKFKECDH--PTIHLD DGT IDDEEYFRTLDENTELVAVFEGHNWID
tr|Q28ZV7|Q28ZV7_DROPS     ASLEEEIRDVKVAEKFKGKCDH--PTIHLD DGT IDDEEYFRTLDENTELVAVFEGHNWID
tr|Q66K97|Q66K97_XENTR     GTLKELIEKASETFLFIHSD--VTLVLE DGT VDTEDFFFQSLEDNTQFLLEQGQKNTQ
unk|VIRIT1655|Blast_submission SCLEDLRSKACDIADLKSLT VTLVLE DGT VDODDYFLCLPSNTKVALASNEKWAY
Q96AQ7|CIDECK_HUMAN       YSLEDLLLKVRTDLMLADK--FFLVLE DGT VETEEYFQALAGDTVMFVLQGQKWP
sp|P56198|CIDECK_MOUSE     HSLEDLLNKVQDILKLKDK--FSLVLE DGT VETEEYFQALAKDTMFVVLQGQKWP
tr|Q5X133|Q5X133_RAT       HSLEDLLKGVQDILKLKDK--FSLVLE DGT VETEEYFQALPRDTVMFVLQGQKWP
sp|o60543|CIDECK_HUMAN     SSLQELISKTLDALVIATG--VTLVLE DGT VDTEEFFQTLGDNNTHLMVLEQGQKNTP
tr|A4FUX1|A4FUX1_BOVIN    SSLQELLSKTLDALVVASQ--VTLVLE DGT VDTEEFFQTLGDNNTHLMVLEQGQKNTP
sp|o70303|CIDECK_MOUSE     ASLQELLDKVLETLLLRG--LTLVLE DGT VDSEDFFQILLEDDDTCLMVLEQGQNSP
sp|Q3T191|CIDECK_BOVIN    ATRQELLDKALEALVLSG--LTLVLE DGT VSEEEFFQLLEDDDTCLMVLEQGQNSP
```



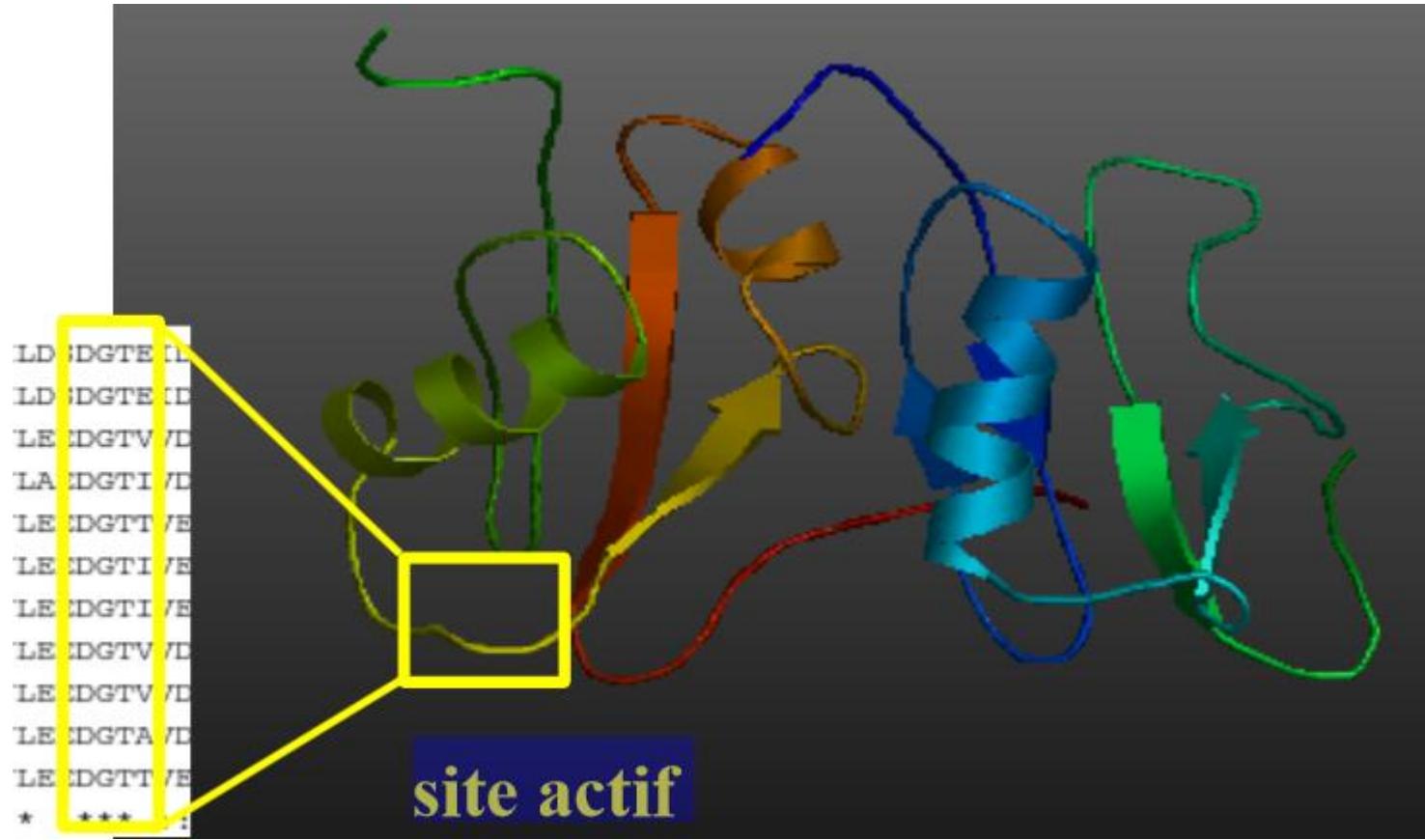
```
tr|o61464|o61464_DROME      PTHYVTTTPHGNNEAGTGNGELNGGGEG-----DTTDANNSES-ARIRQLVQQLQ
tr|Q28ZV7|Q28ZV7_DROPS     PTHYVTTTPHGSSETVTGNGDISSGGVGGGSCDGGTTDANHSESAARIRQLVQQLQ
tr|Q66K97|Q66K97_XENTR     ERNSKRAV-----Q
unk|VIRIT1655|Blast_submission NNSDGGTA-----WISQESF-----DVDETDSGAG-LWKKNVARQLK
Q96AQ7|CIDECK_HUMAN       PSEQGQTRH-----PLSL-----SH-----K
sp|P56198|CIDECK_MOUSE     PSEQRKRR-----AQLAL-----SQ-----K
tr|Q5X133|Q5X133_RAT       PSEQRKKK-----AQLSL-----SQ-----K
sp|o60543|CIDECK_HUMAN     GSQHVP-----TC-----S
tr|A4FUX1|A4FUX1_BOVIN    AGHQTP-----AR-----R
sp|o70303|CIDECK_MOUSE     KS-----M-----LSYGLG-----RE-----K
sp|Q3T191|CIDECK_BOVIN    RRSG-----V-----LSYGLG-----QE-----K
```

```
tr|o61464|o61464_DROME      NNLCNVSVVMNDADLDSLNSNMDPNSLVD-----ITGKEFMEQLKDAGRPLCAKRNAEDRL
tr|Q28ZV7|Q28ZV7_DROPS     NNLCNVSVVMNDADLDSLNSNMDPNSLVD-----ITGKEFMEQLKDAGRPLCAKRNAEDRL
tr|Q66K97|Q66K97_XENTR     HEKKTGIANLTFDLYKLNP-----
```

* = résidu parfaitement conservé
: = substitution conservative
. = substitution semi-conservative

Qu'est ce qu'il y a de si spécial ici ???

SITE ACTIF : CELL DEATH ACTIVATOR PROTEIN FAMILY



GHEGVGKVVKLGAGA
GHEKKGYFEDRGPSA
GHEGYGGRSRGGGYS
GHEFEGPKGCGALYI
GHELRGTTFMPALEC

GHE--G-----
GHE--G-----G---

GHE-x (2) -G-x (5) - [GA]

Consensus 100%
Consensus 60 %

pattern
signature

profil

précision
sensibilité

<A-x-[ST] (2)-x(0,1)-[APTL]-x(4,10)-C-{V}

<A en N terminal
x = n'importe quel AA
ST] (2) = Ser ou Thr 2 fois
x(0,1) 1 aa ou aucun
x(4,10) entre 4 et 10 aa quelconques
{V} tout sauf une Val



Code IUPAC pour les nucléotides

<u>Code</u>	<u>Description</u>
A	Adénine
C	Cytosine
G	Guanine
T	Thymine
U	Uracile
R	Purine (A ou G)
Y	Pyrimidine (C, T, ou U)
M	C ou A
K	T, U, ou G
W	T, U, ou A
S	C ou G
B	C, T, U, ou G (pas A)
D	A, T, U, ou G (pas C)
H	A, T, U, or C (pas G)
V	A, C, or G (pas T, pas U)
N	Toutes les bases (A, C, G, T, ou U)



NiceSite View of PROSITE: PS00191 - Mozilla

Eichier Edition Affichage Aller à Marque-pages Outils Fenêtre Aide

Précédent Suivant Actualiser Arrêter http://www.expasy.org/cgi-bin/nicesite.pl?PS00191 Rechercher Imprimer

Accueil Marque-pages Net2Phone Sorties

[ExPASy Home page](#) [Site Map](#) [Search ExPASy](#) [Contact us](#) [PROSITE](#)

Search PROSITE for P00174 Go Clear

NiceSite View of PROSITE: [PS00191](#)

General information about the entry

Entry name	CYTOCHROME_B5_1
Accession number	PS00191
Entry type	PATTERN
Date	APR-1990 (CREATED); DEC-2004 (DATA UPDATE); SEP-2005 (INFO UPDATE).
PROSITE documentation	PDOC00170

Pattern

Description Cytochrome b5 family, heme-binding domain signature.

Pattern **[FY]-[LIVMK]-{I}-{Q}-H-P-[GA]-G**

Numerical results

- UniProtKB/Swiss-Prot release number: **48.1**, total number of sequence entries in that release: **195058**.
- Total number of hits in UniProtKB/Swiss-Prot: **86 hits in 86 different sequences**
- Number of hits on proteins that are known to belong to the set under consideration: **80 hits in 80 different sequences**
- Number of hits on proteins that could potentially belong to the set under consideration: **0 hits in 0 different sequences**
- Number of false hits (on unrelated proteins): **6 hits in 6 different sequences**
- Number of known missed hits: **4**
- Number of partial sequences which belong to the set under consideration, but which are not hit by the pattern or profile because they are partial (fragment) sequences: **2**
- Precision (true hits / (true hits + false positives)): **93.02 %**
- Recall (true hits / (true hits + false negatives)): **95.24 %**

Comments

- Taxonomic range: **Eukaryotes, Prokaryotes (Bacteria), Eukaryotic viruses**
- Maximum known number of repetitions of the pattern in a single protein: **1**
- 'Interesting' site in the pattern: **5,heme_iron**
- VERSION: **1**

Cross-references

True positive hits:

ACO1_AJECA ([Q12618](#)), ACO1_YEAST ([P21147](#)), CYB2_HANAN ([P09437](#)),
CYB2_YEAST ([P00175](#)), CYB51_ARATH ([Q42342](#)), CYB52_ARATH ([Q48845](#)),
CYB52_SCHPO ([Q9USM6](#)), CYBSL_MIMIV ([Q5UR80](#)), CYBSL_NEUCR ([Q8X0J4](#)),
CYBSM_HUMAN ([O43169](#)), CYBSM_MOUSE ([Q9CQX2](#)), CYBSM_PONPY ([Q5RDJ5](#)),
CYBSM_RAT ([P04166](#)), CYBSR_DROME ([P10967](#)), CYBSR_DROS ([P50266](#))

Alignement

AATTGA

AGGTCC

AGGATG

AGGCGT

Matrice de position

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

Matrice de fréquences

	1	2	3	4	5	6
A	1	0.25	0	0.25	0	0.25
C	0	0	0	0.25	0.25	0.25
G	0	0.75	0.75	0	0.50	0.25
T	0	0	0.25	0.50	0.25	0.25



Matrice de fréquences

$$\log \left[\frac{f_b}{p_b} \right]$$

Matrice de poids de position
(Position Weight Matrix ou Position Specific Scoring Matrix PSSM)

	1	2	3	4	5	6
A	1.2	0	-1.6	0	-1.6	0
C	-1.6	-1.6	-1.6	0	0	0
G	-1.6	0.96	0.96	-1.6	0.59	0
T	-1.6	-1.6	0	0.59	0	0



SOUMISSION D'UNE SÉQUENCE D'INTÉRÊT DANS LES BASES DE DONNÉES DU TYPE PFAM

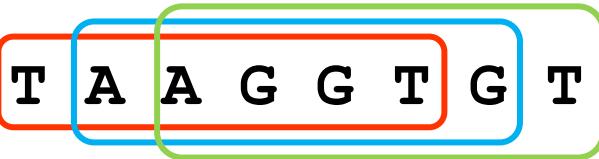
C G T A T G T A A G G T G T A C G T A G

Pour trouver si la séquence contient un motif, les bases de données appliquent sur les séquences soumises les matrices qu'elles ont correspondant toutes à des motifs décrits, et notamment la matrice ci-dessous.

	1	2	3	4	5	6
A	1.2	0	-1.6	0	-1.6	0
C	-1.6	-1.6	-1.6	0	0	0
G	-1.6	0.96	0.96	-1.6	0.59	0
T	-1.6	-1.6	0	0.59	0	0

Calcul de score par fenêtre glissante

C G T A T G **T** A **A** G G T G T A C G T A G



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

	1	2	3	4	5	6
A	1.2	0	-1.6	0	-1.6	0
C	-1.6	-1.6	-1.6	0	0	0
G	-1.6	0.96	0.96	-1.6	0.59	0
T	-1.6	-1.6	0	0.59	0	0

Score = -4.21

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

	1	2	3	4	5	6
A	1.2	0	-1.6	0	-1.6	0
C	-1.6	-1.6	-1.6	0	0	0
G	-1.6	0.96	0.96	-1.6	0.59	0
T	-1.6	-1.6	0	0.59	0	0

Score = 0.56



C G T A T G T A **A G G T G T** A C G T A G

	1	2	3	4	5	6
A	1.2	0	-1.6	0	-1.6	0
C	-1.6	-1.6	-1.6	0	0	0
G	-1.6	0.96	0.96	-1.6	0.59	0
T	-1.6	-1.6	0	0.59	0	0

Score = 4.3

Si dans la base de données interrogée, le seuil minimal de détection de motif est par exemple $S = 4$ alors seule la séquence AGGTGT sera considérée comme pattern

EXEMPLE : RECHERCHE DE MOTIF SUR UNE SÉQUENCE VEM-1 DE *CAENORHABDITIS ELEGANS*

Vema (Mammalian ventral midline antigen) related protein 1, isoform a

UniProtKB/Swiss-Prot: Q9TY05

[GenPept](#) [Graphics](#)

```
>gi|75024827|sp|Q9TY05|Q9TY05_CAEEL Vema (Mammalian ventral midline antigen)
related protein 1, isoform a
MYTVSVTFLHKSSFTMDLSSWFEFTMYDAVFLVVVLGFFFYWLTRSEQPLPAPPKE
LAPLPMPSDMTVEEL
RKYDGVKNEHILFGLNGTIYDVTRGKGFYGPBKAYGTLAGHDATRALGTMDQNAVSSE
WDDHTGISADEQ
ETANEWETQFKFKYLTVGRLVKNSSKADYGNRKSFVRGAESLDIINGGDEGTKKD
```



[HOME](#) | [SEARCH](#) | [BROWSE](#) | [FTP](#) | [HELP](#) | [ABOUT](#)

Pfam 27.0 (March 2013, 14831 families)

The Pfam database is a large collection of protein families, each represented by **multiple sequence alignments** and **hidden Markov models (HMMs)**. [More...](#)



QUICK LINKS

[SEQUENCE SEARCH](#)

[VIEW A PFAM FAMILY](#)

[VIEW A CLAN](#)

[VIEW A SEQUENCE](#)

[VIEW A STRUCTURE](#)

[KEYWORD SEARCH](#)

[JUMP TO](#)

ANALYZE YOUR PROTEIN SEQUENCE FOR PFAM MATCHES

Paste your protein sequence here to find matching Pfam families.

Go

Example

This search will use an E-value of 1.0. You can set your own search parameters and perform a range of other searches [here](#).



EXEMPLE : RECHERCHE DE MOTIF SUR UNE SÉQUENCE VEM-1 DE *CAENORHABDITIS ELEGANS*

EXEMPLE: RECHERCHE DE MOTIF SUR UNE SÉQUENCE VEM-1 DE *CAENORHABDITIS ELEGANS*



[HOME](#) | [SEARCH](#) | [BROWSE](#) | [FTP](#) | [HELP](#) | [ABOUT](#)



[keyword search](#) [Go](#)

Sequence search results

[Show](#) the detailed description of this results page.

We found 2 Pfam-A matches to your search sequence (1 significant and 1 insignificant). You did not choose to search for Pfam-B matches.

Cyt-b5

[Show](#) the search options and sequence that you submitted.

[Return](#) to the search form to look for Pfam domains on a new sequence.

Significant Pfam-A Matches

[Show](#) or [hide](#) all alignments.

Family	Description	Entry type	Clan	Envelope		Alignment		HMM		HMM length	Bit score	E-value	Predicted active sites	Show/hide alignment
				Start	End	Start	End	From	To					
Cyt-b5	Cytochrome b5-like Heme/Steroid binding	Domain	n/a	64	161	65	160	2	75	76	43.6	1.8e-11	n/a	Show

Insignificant Pfam-A Matches

[Show](#) or [hide](#) all alignments.

Family	Description	Entry type	Clan	Envelope		Alignment		HMM		HMM length	Bit score	E-value	Predicted active sites	Show/hide alignment
				Start	End	Start	End	From	To					
Exo_endo_phos_2	Endonuclease-reverse transcriptase	Domain	CL0530	106	189	113	162	30	72	119	11.2	0.21	n/a	Show

Comments or questions on the site? Send a mail to pfram-help@sanger.ac.uk. Our [cookie policy](#).

The Wellcome Trust

Family: Cyt-b5 (PF00173)



Summary

Domain organisation

Clan

Alignments

HMM logo

Trees

Curation & news

Species

Interactions

Structures

Jump to... ↴

enter ID/acc



Go

Summary: Cytochrome b5-like Heme/Steroid binding domain

Pfam includes annotations and additional family information from a range of different sources. These sources can be accessed via the tabs below.

[Wikipedia: Cytochrome b5](#) [Pfam](#) [InterPro](#)

This is the Wikipedia entry entitled "[Cytochrome b5](#)". [More...](#)

Cytochrome b5 [Edit Wikipedia article](#)

Cytochromes b₅ are ubiquitous electron transport hemoproteins found in [animals](#), [plants](#), [fungi](#) and [purple phototrophic bacteria](#). The [microsomal](#) and [mitochondrial](#) variants are membrane-bound, while bacterial and those from [erythrocytes](#) and other [animal tissues](#) are water-soluble. The family of cytochrome b₅-like proteins includes (besides cytochrome b₅ itself) hemoprotein domains covalently associated with other redox domains in flavocytochrome cytochrome b₂ ([L-lactate dehydrogenase; EC 1.1.2.3](#)), sulfite oxidase ([EC 1.8.3.1](#)), plant and fungal nitrate reductases ([EC 1.7.1.1](#), [EC 1.7.1.2](#), [EC 1.7.1.3](#)), and plant and fungal cytochrome b₅/acyl lipid desaturase fusion proteins.

Contents [hide]

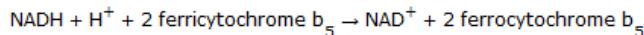
- 1 Structure
- 2 Cytochrome b₅ in some biochemical reactions
- 3 See also
- 4 References
- 5 External links

Structure

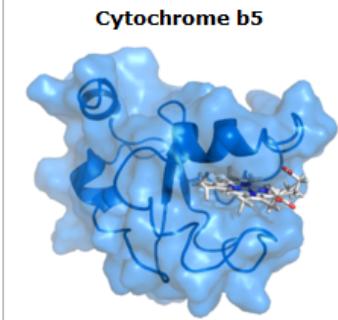
3-D structures of a number of cytochrome b₅ and yeast flavocytochrome b₂ are known. The fold belongs to the α-β class, with two hydrophobic cores on each side of a β-sheet. The larger hydrophobic core constitutes the heme-binding pocket, closed off on each side by a pair of helices connected by a turn. The smaller hydrophobic core may have only a structural role and is formed by spatially close N-terminal and C-terminal segments. The two [histidine](#) residues provide the fifth and sixth heme ligands, and the propionate edge of the heme group lies at the opening of the heme crevice. Two isomers of cytochrome b₅, referred to as the A (major) and B (minor) forms, differ by a 180° rotation of the heme about an axis defined by the α- and γ-meso carbons.

Cytochrome b₅ in some biochemical reactions

[EC 1.6.2.2](#) cytochrome-b₅ reductase



[EC 1.10.2.1](#) L-ascorbate—cytochrome-b₅ reductase



Rat cytochrome b5 bound to heme

Identifiers

Symbol	CYB5A
Alt. symbols	CYB5
Entrez	1528
HUGO	2570
OMIM	250790
PDB	1JEX
RefSeq	NM_001914
UniProt	P00167

Other data

Locus	Chr. 18 q23
-------	-------------

Cytochrome b5

Identifiers

Symbol	Cyt_B5
Pfam	PF00173
InterPro	IPR001199
PROSITE	PDOC00170

Family: Cyt-b5 (PF00173)

 Loading page components (1 remaining)...

 60 architectures
  1547 sequences
  2 interactions
  316 species
  63 structures

Summary

Domain organisation

Alignments

HMM logo

Trees

Curation & models

Species

Interactions

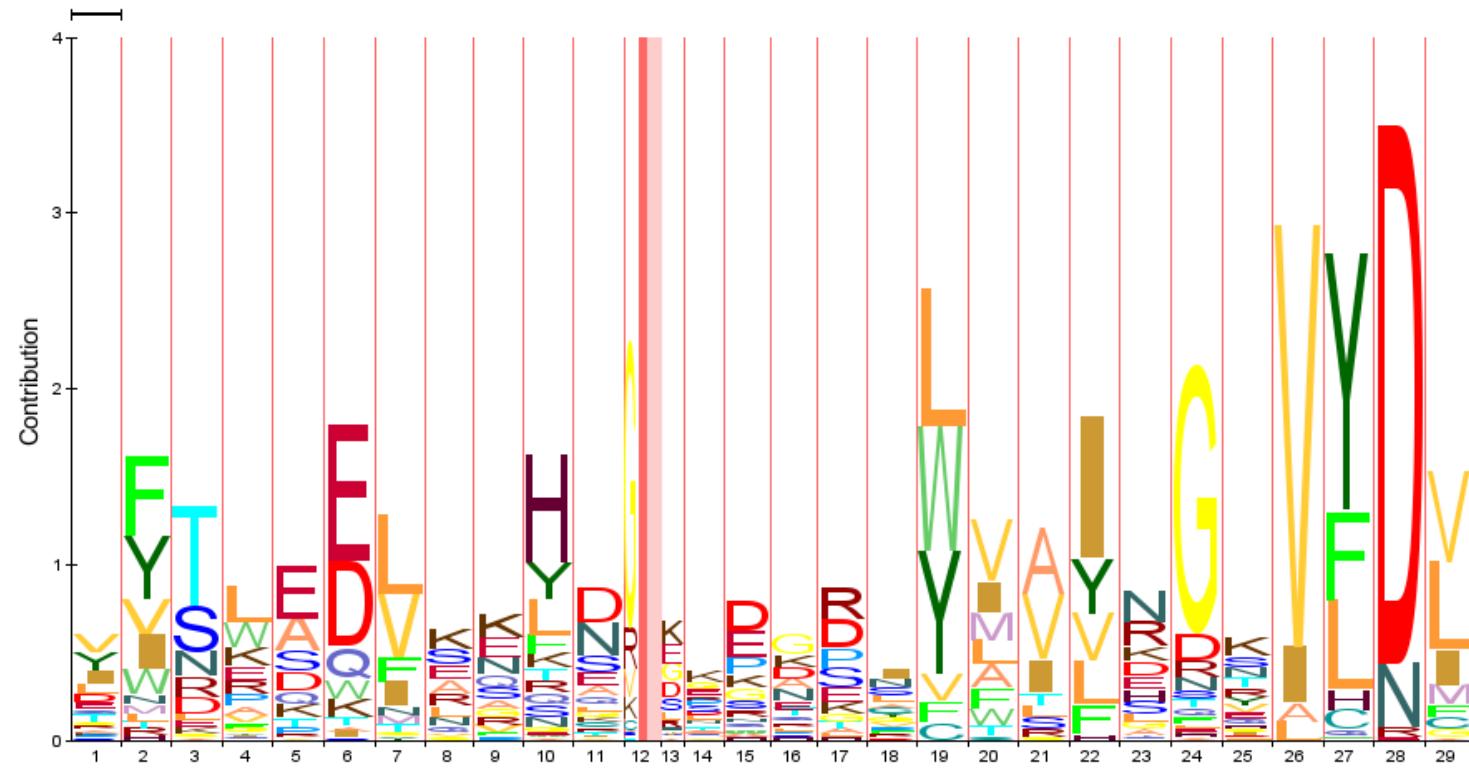
Structures

Jump to...

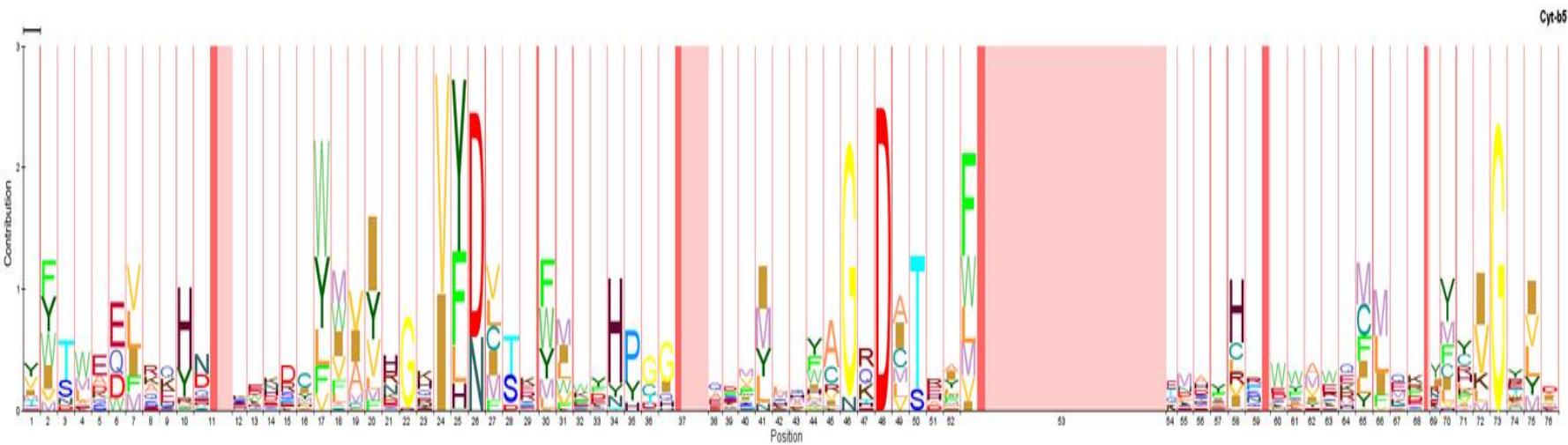
enter ID/acc

HMM logo

HMM logos are one way of visualising profile HMMs. They provide a quick overview of the properties of an HMM in a graphical form. You can see a more detailed description of HMM logos and find out how you can interpret them [here](#). [More...](#)



WEB LOGO ENTIER DE CYT-B5



DESCRIPTION DE CYT-B5 DANS LA CONSERVED DOMAINS DATABASE DU NCBI

NCBI

Conserved Domains

Entrez **CDD** **Structure** **Protein** **Help**

pfam00173: Cyt-b5

Cytochrome b5-like Heme/Steroid binding domain
 This family includes heme binding domains from a diverse range of proteins. This family also includes proteins that bind to steroids. The family includes progesterone receptors. Many members of this subfamily are membrane anchored by an N-terminal transmembrane alpha helix. This family also includes a domain in some chitin synthases. There is no known ligand for this domain in the chitin synthases.

Links **BioAssay Targets and Results** **Statistics** **Structure**

Structure View
 Program: Cn3D Drawing: All Atoms Aligned Rows: up to 10

Statistics

PSSM-Id: 201057
 View PSSM: pfam00173
 Aligned: 211 rows
 ThresholdBitScore: 38.7241
 ThresholdSettingGi: 148887356
 Created: 21-Dec-2011
 Updated: 16-Jan-2013

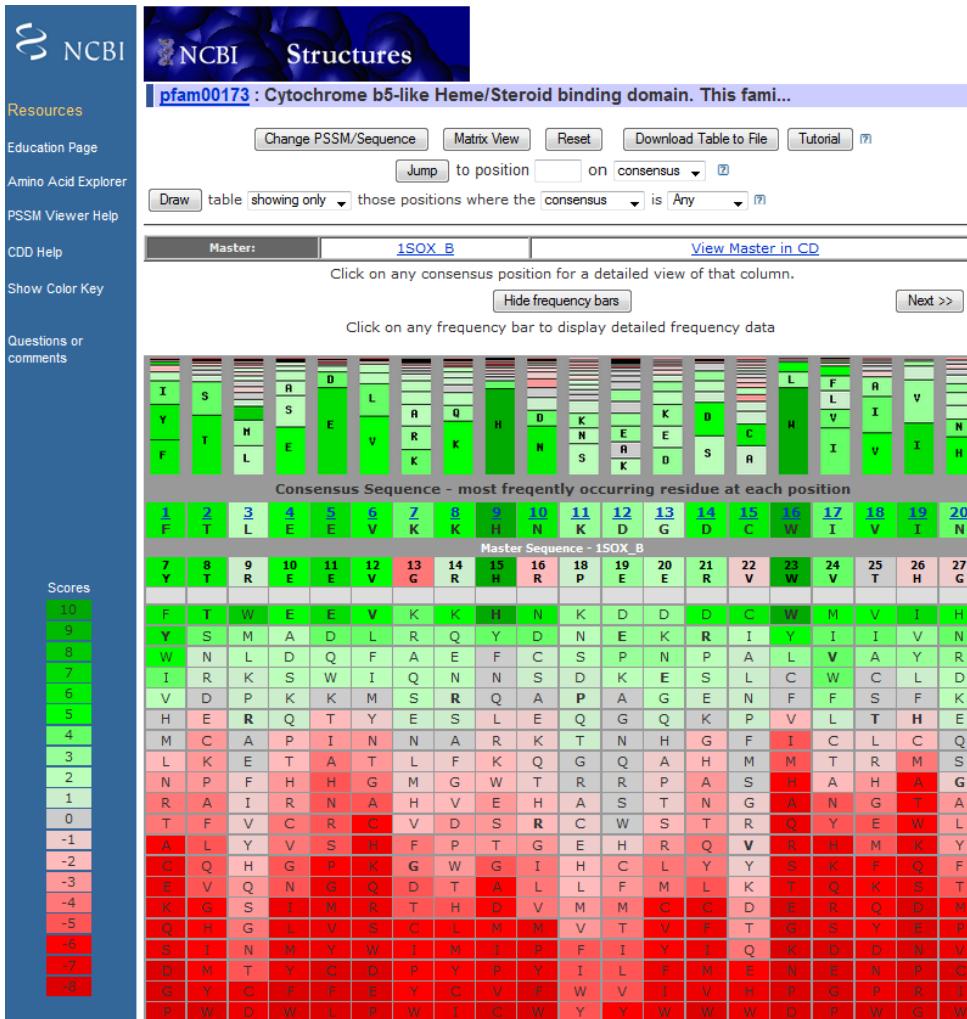
from porcine liver
 Sm. 1998 Jul;
 binding domain. Genome

Sequence Alignment

Format: Hypertext Row Display: up to 10 Color Bits: 2.0 bit Type Selection: the most diverse members

	10	20	30	40	50	60	70	80
1SOX_B*******
gi_5921753	7 YTREEVGRHRs	-PEERVWVTHGTDVFVTD-FVELHPGG						-PDKILLAAQGALEPFWa
gi_166203133	247 YSWDQVAKLE	---NYMVIDGYVLNNSP-YLSFNPTAvegdev--dsiirhvlvlsNQTGSGKDATRLF						60
gi_148887356	408 FTWADIRNNs	---RNLVVYSGHVLDL-LHWFNDTqvtypafkeirdknTAGNQAIRGRDITHAF						306
gi_122065155	403 VSLQWNNTD	---PARNLAVYRGSVLDLNR-LNNLTGGLsypeI---						475
gi_74582634	304 YNWTDI--HE	--PGTSLMVFNGNVLDLsR-IRYlTPNplpiq---						470
gi_6648047	47 RTLSKFNGHG	--DEKIFIAIRGKVYDCTRgRQFYGPSPG						462
gi_75024827	65 MTVEELRKYD	gvKNEHILFLGLNGTIVDVR-GKGFYGP6						361
gi_91206848	1291 VRRADMENL	--LDGSRICILLAGYVCDLSG-YNCESETL						107
2KEO_A	26 VRIADLENHN-	-NDGGFWIVIDGKVYDID-FQTQSLTE						128
								1342
								77
	90	100	110					
1SOX_B*******
gi_5921753	61 -----LYAVHGePHVLELLQQ-YKVGELS							
gi_166203133	307 -----FNRQVP-QDAVGCMKaryYAGGRID							329
	476 -----OSSKD-KOIAECFEEiiKVGsVD							497

PSSM



PSI-BLAST (Position-Specific Iterative)

- alignements multiples de hits ayant les meilleurs scores dans un blast classique
- génération d'un profil en calculant un score pour chacune des positions de l'alignement (PSSM)
- utilisation façon itérative de ce profil pour faire de nouvelles recherches et affinage à chaque itération

PHI-BLAST (Pattern Hit Initiated BLAST)

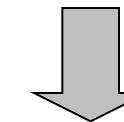
Pattern donné par l'utilisateur puis PSI-BLAST

Intérêt : recherche de familles de protéines
détecter des membres que BLAST ne trouve pas

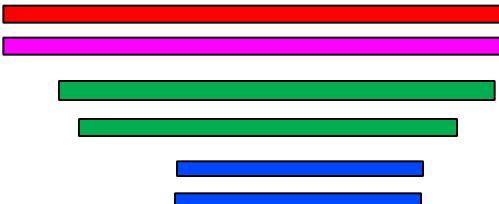
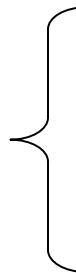
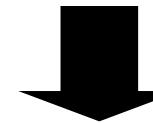


PSI-BLAST (Position-Specific Iterative)

Query



Banque



P	C	Master	A	G	I	L	V	M	F	W	P	C	S	T	Y	N	Q	H	K	R	D	E
			-3	-3	2	0	3	1	-2	-4	-2	-4	-1	1	4	-4	1	-3	-2	-2	0	0
1	V	5 - K	-3	-3	2	0	3	1	-2	-4	-2	-4	-1	1	4	-4	1	-3	-2	-2	0	0
2	Y	6 - I	-4	-1	3	-1	1	0	4	4	-6	-4	-3	3	-3	0	0	-5	-3	4	-5	5
3	X	7 - S	-2	-3	-4	-3	-3	-4	-3	-5	-2	-2	3	3	-4	1	-3	-4	-2	1	0	-1
4	L	8 - P	0	-3	-2	2	-1	2	0	4	1	-5	-2	-3	-1	-4	-2	-1	0	-4	0	
5	T	9 - A	2	-3	-5	-5	-3	-4	-6	-5	-2	-3	2	-1	-6	3	1	-2	1	-2	2	7
6	X	10 - E	-3	-5	-3	-5	-5	-4	-6	-2	-4	-5	-3	-2	-5	-2	2	-3	0	-3	4	6
7	V	11 - Y	-3	-3	2	3	5	0	2	-4	-5	-4	-4	-2	0	-3	-5	-5	-5	-6	4	-6
8	K	12 - A	2	-3	-1	-2	-1	-3	-5	-4	1	-3	-4	0	2	-1	3	2	-3	1	-1	
9	K	13 - K	0	-1	-5	-4	-2	-4	-1	-2	-5	0	-2	-4	1	3	-3	4	1	-2	2	
10	F	14 - H	-4	-5	-1	-5	-3	0	-1	-5	-2	-2	3	0	0	9	-1	-1	-4	-2	-2	
P	C	Master	A	G	I	L	V	M	F	W	P	C	S	T	Y	N	Q	H	K	R	D	E
11	W	15 - N	0	-2	-3	-2	-4	-4	-5	-6	-4	-1	1	-1	-5	0	-1	-1	-2	4	0	
12	K	16 - K	-1	0	-4	-2	-2	-2	-4	-4	1	-1	2	1	-4	2	1	-1	3	0	1	
13	D	17 - P	0	0	-4	-4	-3	-3	-3	0	2	-3	0	-2	-4	0	0	-1	2	0	3	2
14	G	18 - O	0	2	-5	-3	-4	-3	-5	-1	-5	-1	-1	-4	2	1	0	3	-1	3	2	
15	P	19 - D	-2	-2	-5	-4	-5	-4	-5	-6	-4	2	-5	2	-2	-4	-1	-2	1	3	5	1
16	C	20 - C	1	0	2	1	-1	0	-1	-5	1	6	0	-2	-1	1	-2	-4	-1	-1	-2	
17	W	21 - W	-5	-6	-2	-1	-3	1	11	-6	0	-5	-4	-4	-6	-5	-4	-5	-6	-7	-6	
18	I	22 - V	-1	-6	4	1	3	4	3	4	-5	-1	-4	-1	-3	-6	-5	-5	-6	-6	-6	
19	V	23 - Y	3	-3	4	-1	-4	-2	-4	-5	-4	0	0	0	-4	-5	-4	-2	-1	-5	-4	
20	I	24 - J	-1	-6	6	1	3	-2	1	-4	-5	-2	-5	-3	-3	-5	-1	-5	-6	4	6	
P	C	Master	A	G	I	L	V	M	F	W	P	C	S	T	Y	N	Q	H	K	R	D	E
21	N	25 - N	-2	-1	-5	-3	-5	-4	-3	-5	-4	-5	0	-3	-2	4	0	3	2	2	2	1
22	G	26 - G	-3	6	-4	-5	-6	-5	-3	-5	-5	-5	-2	-3	-5	-2	-1	-3	0	1	-3	
23	K	27 - Y	-3	-2	-1	-3	0	-4	-2	-5	-4	-5	-1	0	1	1	3	4	2	-2	1	
24	Y	28 - W	-2	-2	-3	-2	-3	-5	-5	-3	-4	-3	-4	-5	-5	-6	-5	-6	-7	-6	-5	
25	Y	29 - Y	-4	-3	-3	0	-3	-3	0	-2	-1	-4	-4	9	-5	-3	-3	-5	-4	-5	-5	
26	O	30 - D	-4	-4	-6	-5	-6	-5	-7	-6	-7	-4	-5	-3	-3	-4	-3	-3	-2	8	-1	
27	V	31 - L	-1	-1	3	2	3	1	0	-5	-5	3	-4	-3	-4	-5	-5	-6	-5	-6	-6	
28	T	32 - T	-3	-2	-4	-4	-3	-4	-5	-4	-5	-4	6	-4	0	-2	-4	-2	-2	0	-2	
29	R	33 - R	-1	0	-5	-3	-5	-2	-5	-5	-2	-5	1	-2	-4	0	0	-1	2	3	2	
30	F	34 - F	-6	-2	-3	0	-4	-1	7	6	-6	-5	-3	-4	-4	-5	-4	-3	-2	-1	4	-5
P	C	Master	A	G	I	L	V	M	F	W	P	C	S	T	Y	N	Q	H	K	R	D	E

PSSM (*position specific score matrix*)
(matrice de poids de position)

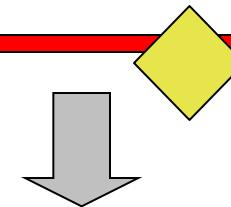


PHI-BLAST (Pattern Hit Initiated)

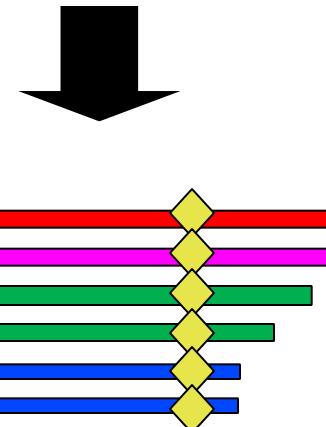
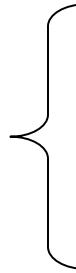
Query

C - C - {P} - {P} - x - C - [STDNEKPI] - x(3) - [LIVMFS] - x(3) - C

Pattern Prosite



P	C	Master	A	G	I	L	V	M	F	W	P	C	S	T	Y	N	Q	H	K	R	D	E
1	V	5 - K	-3	-3	2	0	3	1	-2	-4	-2	-4	-1	1	4	-4	1	-3	-2	-2	0	0
2	Y	6 - I	-4	3	-1	1	0	4	4	-6	-4	-3	-3	-3	-5	0	-5	-3	-4	-5	0	
3	X	7 - S	-2	-3	-4	-3	-4	-3	-5	-2	-2	3	3	-4	-1	-3	-4	-2	1	0	-1	
4	L	8 - P	0	-3	-2	2	-1	2	0	1	-5	-2	-3	-1	-4	-2	-2	1	0	-4	0	
5	T	9 - A	2	-3	-5	-5	-3	-4	-6	-5	-2	-3	2	-1	-5	3	-1	-2	1	-2	2	
6	X	10 - E	-3	-5	-3	-5	-5	-4	-6	2	-4	-5	-3	-2	-5	-2	2	-3	0	-3	4	6
7	V	11 - Y	-3	-3	2	3	5	0	2	-4	-5	-4	-4	-2	0	-3	-5	-5	-5	-4	0	-5
8	K	12 - A	2	-3	-4	-1	-2	-1	-3	-5	-4	1	-3	-4	0	2	-1	3	2	-3	1	
9	K	13 - K	0	-1	-5	-4	-2	-4	-1	-2	-5	0	-2	-4	1	3	-3	4	1	-2	2	
10	F	14 - H	-4	-5	-1	-5	-3	0	-1	-5	-2	-2	3	0	0	9	-1	-1	-4	-2	-2	
P	C	Master	A	G	I	L	V	M	F	W	P	C	S	T	Y	N	Q	H	K	R	D	E
11	W	15 - N	0	-2	-3	-2	-4	-4	-5	-6	-4	-1	1	-1	-5	0	-1	-1	-2	-4	0	
12	K	16 - K	-1	0	-4	-2	-2	-3	-4	-1	-1	2	1	-4	2	1	-1	3	0	1	0	
13	D	17 - P	0	0	-4	-4	-4	-3	-3	0	2	-3	0	-2	-4	0	0	-1	2	0	3	2
14	G	18 - O	0	2	-5	-3	-3	-5	-5	-1	-1	-1	-4	2	1	0	3	-3	1	3	2	
15	P	19 - D	-2	-2	-5	-4	-5	-4	-5	-6	2	-5	2	-2	-4	-1	-2	1	3	5	1	
16	C	20 - L	1	0	2	1	-1	-1	-5	1	6	0	-2	-1	1	-2	-4	-1	-1	-2	-4	
17	W	21 - W	-5	-6	-2	-1	-3	1	11	-6	0	-5	-4	-4	-6	-5	-4	-5	-5	-6	-5	
18	I	22 - V	-1	-6	4	1	3	4	3	-4	-5	-1	-4	-1	-3	-8	-5	-5	-5	-6	-6	
19	V	23 - Y	3	-3	4	-1	-4	-2	-4	-5	-4	0	0	0	-4	-5	-4	-2	-1	-5	-4	
20	I	24 - J	-2	-6	6	1	3	-2	1	-4	-5	-2	-5	-3	-3	-5	-1	-5	-6	-4	-6	
P	C	Master	A	G	I	L	V	M	F	W	P	C	S	T	Y	N	Q	H	K	R	D	E
21	N	25 - N	-2	-1	-5	-3	-5	-4	-3	-5	-4	0	-3	-2	4	0	3	2	2	2	1	
22	G	26 - G	-3	6	-4	-5	-6	-5	-3	-5	-5	-2	-3	-5	-2	-1	-3	0	1	-3		
23	K	27 - Y	-3	-2	-1	0	-4	-2	-5	-4	-5	-1	0	1	1	3	4	2	-2	1		
24	W	28 - W	-2	-2	-7	-2	-2	-5	-5	-3	-4	-3	-4	-5	-5	-6	-5	-6	-5	-5		
25	C	29 - Y	-4	-3	-3	0	-3	-3	0	-6	-1	-4	-4	9	-5	-3	-5	-4	-5	-5		
26	P	30 - D	-4	-4	-6	-5	-6	-5	-7	-4	-5	-3	-3	-4	3	-3	-3	-2	8	-1		
27	V	31 - L	-1	-1	3	2	3	1	0	-5	-5	3	-4	-3	-4	-5	-5	-6	-5	-6		
28	I	32 - T	-3	-2	-4	-4	-3	-4	-5	-4	-4	6	-4	0	-2	-4	-2	-2	0	-2		
29	R	33 - R	-1	0	-5	-3	-5	-2	-5	-5	-2	1	-2	-4	0	0	-1	2	3	2	1	
30	F	34 - F	-6	-2	-3	0	-4	-1	7	6	-6	-5	-3	-4	-4	-5	-4	-3	-2	-1	-6	-5
P	C	Master	A	G	I	L	V	M	F	W	P	C	S	T	Y	N	Q	H	K	R	D	E



PSSM (*position specific score matrix*)
(matrice de poids de position)

