

# Introduction à la bioinformatique (UE SSV3U15)

## TP4. Alignement par paire et alignement multiple

Yvan Perez et Andreas Zanzoni

## Objectifs

- L'objectif de ce TP est de découvrir les sites web permettant de générer des alignements de séquences protéiques par paire ou multiples, d'essayer d'y localiser les domaines et de modéliser la famille pour en détecter tous les membres.

## Notions mises en pratique

- **Recherche par similarité** : alignement par paire d'une séquence d'intérêt (subject) avec toutes les séquences
  - Alignement local
  - Eléments de l'alignement: matches, mismatches, indels
- **Alignements multiples**
  - constater les blocs de conservation, et les régions plus variables, les domaines fonctionnels
  - résolution des insertions versus délétions, qui n'était pas résoluble en alignement par paire
  - constater les substitutions fréquentes
- **Matrices de similarité** – Liens avec la biochimie.

N'oubliez pas que vous pouvez à tout moment consulter le [glossaire du cours](#) pour obtenir une définition sommaire des principaux termes utilisés.

## Etapes

- Exercice 1. **Alignement par paire et alignement multiple au NCBI**
  - Observation et compréhension des résultats de BLAST
  - Téléchargement d'un ensemble de séquences homologues
  - Alignement multiple et MSA viewer au NCBI
- Exercice 2. **Alignement multiple à l'EBI**
  - Différents formats de sortie (Pearson/FASTA, ClustalW)
  - Outil de visualisation et d'édition d'un alignement multiple

## Complétion

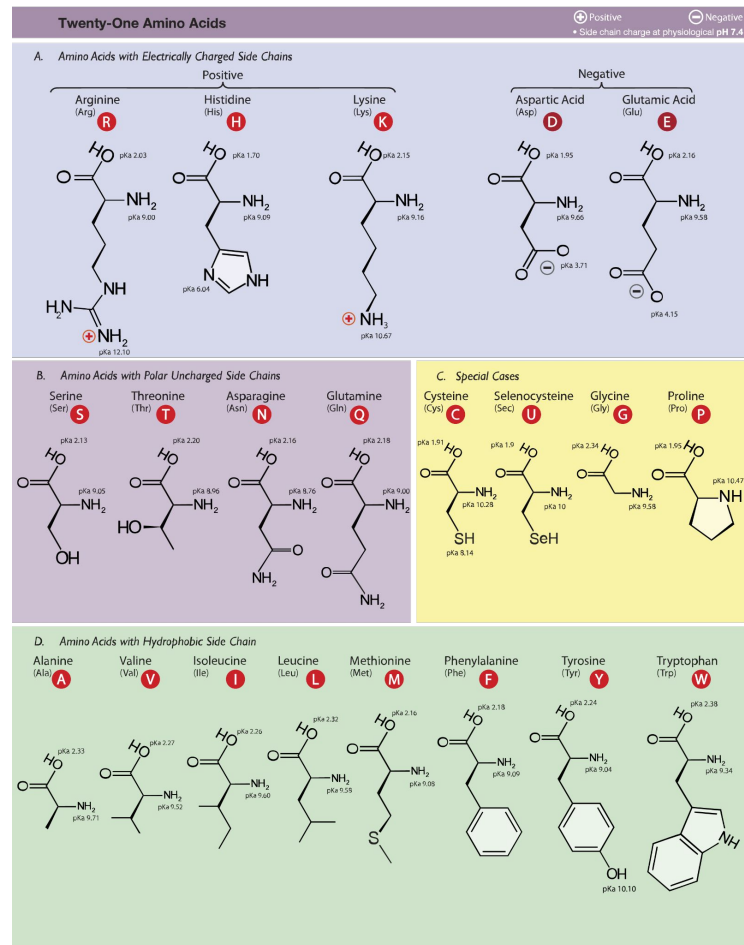
- Tous les exercices doivent être réalisés par chaque étudiant.
- En principe, les deux exercices devraient être faits en séance (avec explications par les enseignants).
- Si nécessaire, ils peuvent être terminés ultérieurement.

# Rappels des définitions



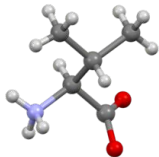
# Rappel – Nomenclature et composition des acides aminés

Amino Acid	Abbrev	1-lettre	Codon(s)
Alanine	Ala	A	GCA, GCC, GCG, GCT
Arginine	Arg	R	CGA, CGC, CGG, CGT, AGA, AGG
Aspartic acid	Asp	D	GAC, GAT
Asparagine	Asn	N	AAC, AAT
Cysteine	Cys	C	TGC, TGT
Glutamic acid	Glu	E	GAA, GAG
Glutamine	Gln	Q	CAA, CAG
Glycine	Gly	G	GGA, GGC, GGG, GGT
Histidine	His	H	CAC, CAT
Isoleucine	Ile	I	ATA, ATC, ATT
Leucine	Leu	L	CTA, CTC, CTG, CTT, TTA, TTG
Lysine	Lys	K	AAA, AAG
Methionine	Met	M	ATG
Phenylalanine	Phe	F	TTC, TTT
Proline	Pro	P	CCA, CCC, CCG, CCT
Serine	Ser	S	TCA, TCC, TCG, TCT, AGC, AGT
Threonine	Thr	T	ACT, ACC, ACG, ACT
Tryptophan	Trp	W	TGG
Tyrosine	Tyr	Y	TAC, TAT
Valine	Val	V	GTA, GTC, GTG, GTT
STOP	-	-	TAG, TAA, TGA

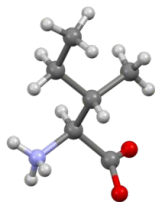


## Similarités chimiques entre acides aminés

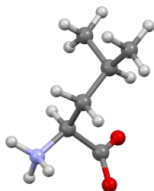
### Valine (Val)



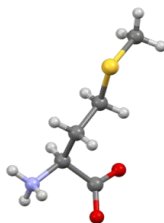
### Isoleucine (Ile)



## Leucine (Leu)

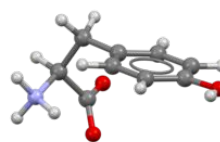


### Méthionine (Met)

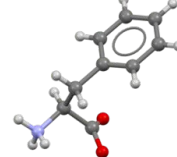


## Hydrophobes

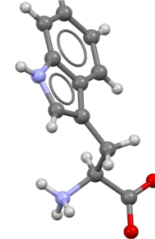
### Tyrosine (Tyr)



### Phénylalanine (Phe)

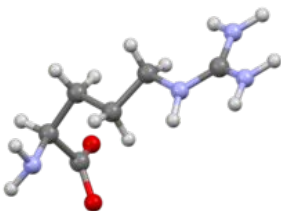


### Tryptophane (Trp)

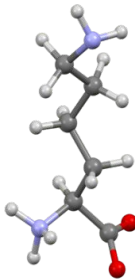


## Aromatiques

### Lysine (Lys)



### Arginine (Arg)

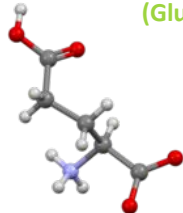


## Chargés +

Acide aspartique (Asp)



**Acide glutamique (Glu)**



### Chargés -

## Matrice de substitutions entre acides aminés

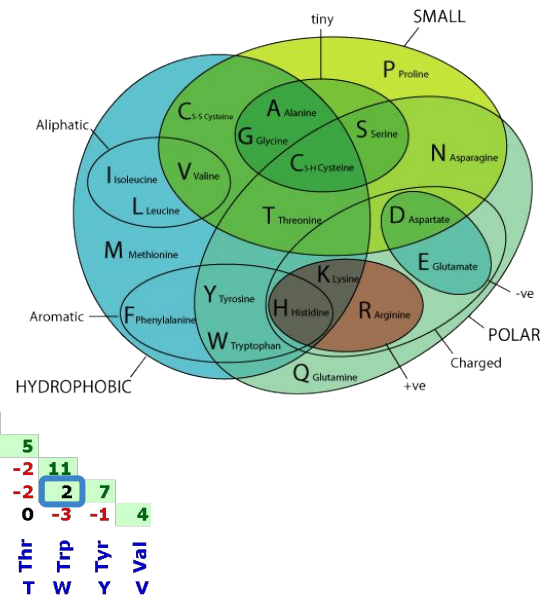
Hydrophobicity scale diagram showing the relative hydrophobicity of 20 amino acids. The scale ranges from -4 (Ala) to 4 (Val). A pie chart on the right categorizes amino acids as Aliphatic, Aromatic, or HYDROPHOBIC.

Amino Acid	Hydrophobicity
Ala	-4
Arg	-1
Asn	0
Asp	-2
Cys	-3
Gln	-1
Glut	-1
Gly	0
His	-2
Ile	-1
Leu	-1
Lys	-1
Met	-1
Phe	-2
Pro	-1
Ser	1
Thr	0
Trp	-3
Tyr	-2
Val	0

Aliphatic: Ala, Arg, Asn, Asp, Cys, Glu, Gly, His, Ile, Leu, Lys, Met, Phe, Pro, Ser, Thr, Val

Aromatic: Phe, Tyr

HYDROPHOBIC: Trp, Val



# Recherche de similarité dans les bases de données

- La ligne entre les séquences "Query" et "Sbjct" indique les correspondances entre acides aminés.
- Identités
- Substitutions "conservatives": paires de résidus distincts mais dont la substitution est généralement moins délétère que pour d'autres paires de résidus.
- Substitutions non conservatives
- Positives: identités + substitutions conservatives.
- Gaps: lacunes insérées dans une séquence afin d'optimiser l'alignement des fragments avoisinants.

```
>gi|16127996|ref|NP_414543.1| bifunctional: aspartokinase I  
(N-terminal); homoserine dehydrogenase I (C-terminal)  
[Escherichia coli K12]  
Length = 820
```

```
Score = 344 bits (882), Expect = 2e-95
```

```
Identities = 247/821 (30%) Positives = 410/821 (49%) Gaps = 44/821 (5%)
```

```
Query: 16 KFGGSSLADV KCYLRVAGIMAEYSQPDMM-VVSAAGSTTNQLINWLKLSQTDRLSAHQV 74  
          KFGG+S+A+ + +LRVA I+ ++ + V+SA TN L+ ++ + + + + +  
Sbjct: 5 KFGGTSVANAERFLRVADILESMAROGQVATVLSAPAKITNHLVAMIEKTISGQDALPNI 64  
  
Query: 75 QQTLRRYQCDLISGLLPAREADSL--ISAFVSDLERLAALLDSGIN-----DAVYAEVV 126  
          R + +L++GL A+ L + FV + GI+ D++ A ++  
Sbjct: 65 SDAERIF-AELLTGLAAACPGFPLAQLKTFVDQEFQAQIKHVLHGISLLGQCPSINAALI 123  
  
Query: 127 GHGEVWSARLMSAVLNQOGLPAAWLDAREFLRAER---AAQPQVDEGLSYPLLQQLLVQH 183  
          GE S +M+ VL +G +D E L A + + E ++ H  
Sbjct: 124 CRGEKMSIAIMAGVLEARGHNVTVIDPVEKLLAVGHYLESTVDIAESTRRIAASRIPADH 183  
  
Query: 184 PGKRLVVTGFI SRNAGETVLLGRNGSDYSATQIGALAGVSRVTIWSDVAGVYSADPRKV 243  
          +++ GF + N GE V+LGRNGSDYSA + A IW+DV GVY+ DPR+V  
Sbjct: 184 ---MVL MAGFTAGNEKGELVVLGRNGSDYSAAVLAACLRADCCEIWTDVDGVYTCDPRQV 240  
  
Query: 244 KDA CL LPLRLDEASELARLAAPVLHARTLQPVSGSEIDLQLRCSYTPDQ-----GSTRI 298  
          DA LL + EA EL+ A VLH RT+ P++ +I ++ + P G++R  
Sbjct: 241 PDARLLKMSYQEAMELSYFGAKVLHPRTITPIAQFQIPCLIKNTGNPQAPGTLIGASRD 300  
  
Query: 299 ERVLASGTGARIVTSHDDVCLIEFQVPASQDFKLAHKEIDQILKRAQVRPLAVGVHNDRQ 358  
          E L + +++ +++ + P + + + RA++ + + +  
Sbjct: 301 EDELP---VKGISNLNNMAMFSVSGPGMKGMVGMMAARVFAAMSRARISVVLITQSSEY 356  
  
Query: 359 LLQFCYTSEVADSALKILDEA-----GLPGELRLRQGLALVAMVGAGVTRNPLHCHRF 411
```



## Résultat de BLAST – Requête peptidique vs DB de peptides

Exemple de résultat de recherche par similarité de séquences.

- Requête (**query**): metaA
- Protéine identifiée dans la base de données: (**subject**): thrA.

Le premier critère d'évaluation d'un résultat de BLAST:

- La **e-valeur (expect)** indique le nombre de faux-positifs attendus au hasard, si l'on plaçait le seuil au niveau du score observé (**344 bits** dans ce cas-ci).
- **Plus la e-valeur est faible, plus le résultat est statistiquement significatif.** Dans le cas présent, il est très significatif (**Expect = 2e-95**)
- **Si la e-valeur est  $\geq 1$ , le résultat n'est pas significatif** (on s'attendrait à trouver un alignement « aussi bon » avec des séquences aléatoires.

```
>gi|16127996|ref|NP_414543.1| bifunctional: aspartokinase I
      (N-terminal); homoserine dehydrogenase I (C-terminal)
      [Escherichia coli K12]
      Length = 820
```

```
Score = 344 bits (882), Expect = 2e-95
Identities = 247/821 (30%), Positives = 410/821 (49%), Gaps = 44/821 (5%)
```

```
Query: 16  KFGGSSLADVCKYLRVAGIMAEYSQPDDMM-VVSAAGSTTNQLINWLKLSQTDRLSAHQV 74
           KFGG+S+A+ + +LRVA I+ ++ + V+SA TN L+ ++ + + + + +
Sbjct: 5   KFGGTSVANAERFLRVADILESNAHQGVATVLSAPAKITNHLVAMIEKTISGQDALPNI 64

Query: 75  QQTLRRYQCDLISGLLPAAEEADSL--ISAFVSDLERLAALLDSGIN-----DAVYAEVV 126
           R + +L++GL A+ L + FV + GI+ D++ A ++
Sbjct: 65  SDAERIF-AELLTGLAAAQPGFPLAQLKTFVDQEFQAQIKHVLHGISLLGQCPDSINAALI 123

Query: 127 GHGEVWSARLMSAVLNQQGLPAAWLDAREFLRAER---AAQPQVDEGLSYPLLQQLLVQH 183
           GE S +M+ VL +G +D E L A + + E ++ H
Sbjct: 124 CRGEKMSIAIMAGVLEARGHNVTVIDPVEKLLAVGHYLESTVDIAESTRRIAASRIPADH 183

Query: 184 PGKRLVVTGFI SRNNAGETVLLGRNGSDYSATQIGALAGVSRVTIWSDVAGVYSADPRKV 243
           +++ GF + N GE V+LGRNGSDYSA + A IW+DV GVY+ DPR+V
Sbjct: 184 ---MVL MAGFTAGNEKGELVVLGRNGSDYSAAVLAACL RADCC EIWTDVDGVYTCDPRQV 240

Query: 244 KDA CLLPLRLDEASELARLAAPVLHARTLQPVSGSEIDLQLRCSYTPDQ-----GSTRI 298
           DA LL + EA EL+ A VLH RT+ P++ +I ++ + P G++R
Sbjct: 241 PDARLLKSMYSQEAMELSYFGAKVLHPRITITPIAQFQIPCLIKNTGNPQAPGTLIGASRD 300

Query: 299 ERVLASGTGARIVTSHDDVCLIEFQVPASQDFKLAHKEIDQILKRAQVRPLAVGVHNDRQ 358
           E L + +++ +++ + P + + + RA++ + + +
Sbjct: 301 EDELP---VKGISNLNNMAMFSVSGPGMKGMVGM AARVFAAMSRRARISVVLITQSSEY 356

Query: 359 LLQFCYTSEVADSALKILDEA-----GLPGELRLRQGLALVAMVGAGVTRNPLHCHRF 411
```

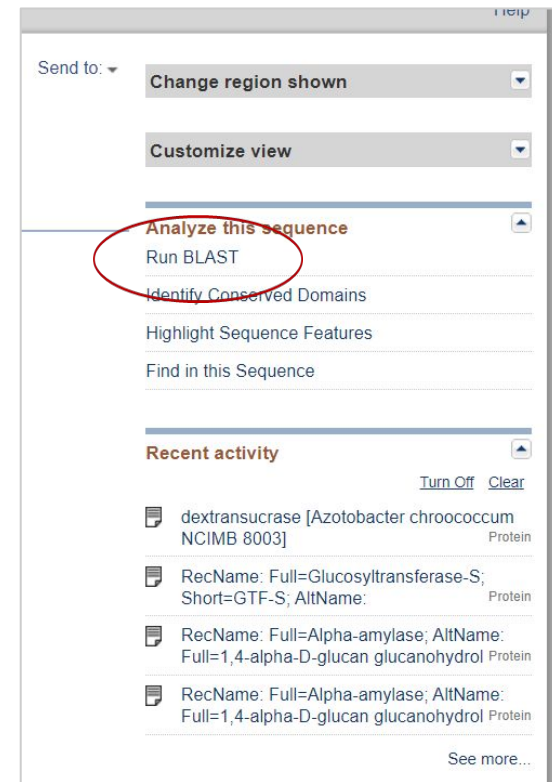
# Tutoriel et exercices

## Exercice 1. Alignements par paires et multiples au NCBI

Nous travaillerons à partir de la protéine **dextransucrase** ([AJE22990.1](#)) de la bactérie *Azotobacter chroococcum*. Cet enzyme catalyse notamment la biosynthèse du dextrane à partir du sucrose. Pour collecter des séquences protéiques homologues de la dextransucrase, nous allons lancer des recherches BLASTP en utilisant l'interface web du NCBI. L'objectif sera de sélectionner un sous-ensemble d'homologues dans les résultats et de les télécharger afin de générer ensuite un alignement multiple (Exercice 2).

- Dans la base de données [protein du NCBI](#), ouvrez la fiche de la protéine [AJE22990.1](#).
- Dans la section “**Analyze this sequence**” (colonne de droite) cliquez sur “**Run BLAST**”. Une fenêtre BLASTP s'ouvre.
- Dans un premier temps, choisissez comme database **UniProtKB/Swiss-Prot**, qui ne contient que des séquences vérifiées par des humains et est plus petite et donc plus rapide.
- Lancez la recherche en cliquant sur le bouton **BLAST** en bas de page.

Au bout de quelques secondes ou minutes, le résultat devrait s'afficher.



# BLAST – Observer, comprendre le tableau de résultats

Explorez le tableau (onglet **Descriptions**) :

- Combien de séquences retourne BLAST?
- Avec quelle E-value ?
- Quels sont les % d'identité et de couverture (Cover) ?

Ces statistiques sont importantes car elles sont nécessaires à la bonne appréciation de la qualité de l'alignement, notamment la significativité de la similarité.

Une significativité élevée permet de conclure à l'homologie des séquences (émettre l'hypothèse que ces séquences sont issues d'un ancêtre commun).

**NIH National Library of Medicine**  
National Center for Biotechnology Information

BLAST® » blastp suite » results for RID-GMBW1SYS013

Home Recent Results Saved Strategies Help

[Edit Search](#) [Save Search](#) [Search Summary](#) [How to read this report?](#) [BLAST Help Videos](#) [Back to Traditional Results Page](#)

Job Title: gblAJE22990.1  
RID: GMBW1SYS013 Search expires on 10-13 16:12 pm [Download All](#) [Citation](#)  
Program: BLASTP  
Database: swissprot [See details](#)  
Query ID: AJE22990.1  
Description: dextranucrase [Azotobacter chroococcum NCIMB 8003]  
Molecule type: amino acid  
Query Length: 780  
Other reports: [Distance tree of results](#) [Multiple alignment](#) [MSA viewer](#)

**Filter Results**

Organism only top 20 will appear ☐ exclude  
Type common name, binomial, taxid or group name  
[Add organism](#)

Percent Identity:  to  E value:  to  Query Coverage:  to   
[Filter](#) [Reset](#)

**Descriptions** Graphic Summary Alignments Taxonomy

**Sequences producing significant alignments** Download Select columns Show 100

☒ select all 25 sequences selected [GenPept](#) [Graphics](#) [Distance tree of results](#) [Multiple alignment](#) [MSA Viewer](#)

	Description	Scientific Name	Max Score	Total Score	Query Cover	E value	Per. Ident	Acc. Len	Accession
<input checked="" type="checkbox"/>	RecName: Full=Glucosyltransferase-I; Short=GTF-I; AltName: Full=Dextranucrase; AltName: Full=Sucrose 6-glucos...	Streptococcus m...	149	149	43%	2e-35	29.06%	1476	P08987.3
<input checked="" type="checkbox"/>	RecName: Full=Glucosyltransferase-I; Short=GTF-I; AltName: Full=Dextranucrase; AltName: Full=Sucrose 6-glucos...	Streptococcus do...	144	144	58%	3e-34	26.92%	1597	P11001.1
<input checked="" type="checkbox"/>	RecName: Full=Glucosyltransferase-SI; Short=GTF-SI; AltName: Full=Dextranucrase; AltName: Full=Sucrose 6-gluc...	Streptococcus m...	140	140	43%	9e-33	28.82%	1455	P13470.2
<input checked="" type="checkbox"/>	RecName: Full=Glucosyltransferase-S; Short=GTF-S; AltName: Full=Dextranucrase; AltName: Full=Sucrose 6-gluc...	Streptococcus do...	139	215	72%	2e-32	27.64%	1365	P29336.1
<input checked="" type="checkbox"/>	RecName: Full=Glucosyltransferase-I; Short=GTF-I; AltName: Full=Dextranucrase; AltName: Full=Sucrose 6-glucos...	Streptococcus do...	138	138	58%	3e-32	26.57%	1592	P27470.1
<input checked="" type="checkbox"/>	RecName: Full=Glucosyltransferase-S; Short=GTF-S; AltName: Full=Dextranucrase; AltName: Full=Sucrose 6-gluc...	Streptococcus m...	136	217	71%	1e-31	26.51%	1462	P49331.3
<input checked="" type="checkbox"/>	RecName: Full=Alpha-amylase; AltName: Full=1.4-alpha-D-glucan glucanohydrolase; AltName: Full=BLA; Flags: Pre...	Bacillus lichenifor...	82.8	145	62%	3e-15	34.78%	512	P06278.1
<input checked="" type="checkbox"/>	RecName: Full=Alpha-amylase; AltName: Full=1.4-alpha-D-glucan glucanohydrolase; Flags: Precursor [Bacillus amyl...	Bacillus amyloliqu...	82.4	145	61%	4e-15	35.58%	514	P00692.1
<input checked="" type="checkbox"/>	RecName: Full=Dextranucrase 1; AltName: Full=Glucansucrase 1; AltName: Full=Sucrose 6-glucosyltransferase 1 [...]	Leuconostoc mes...	79.3	79.3	18%	6e-15	30.81%	284	B2MUJ06
<input checked="" type="checkbox"/>	RecName: Full=Glucan 1.4-alpha-maltohexaosidase; AltName: Full=Exo-maltohexaosidase; AltName: Full=G6-am...	Bacillus sp. 707	79.0	137	63%	5e-14	33.13%	518	P19571.1
<input checked="" type="checkbox"/>	RecName: Full=Alpha-amylase; AltName: Full=1.4-alpha-D-glucan glucanohydrolase; Flags: Precursor [Vigna mungo]	Vigna mungo	63.9	63.9	12%	2e-09	37.37%	421	P17859.1

# BLAST – Interprétation d'un alignement

Sur **Ametice**, répondez aux questions du **Questionnaire 1** "**Alignement local par paire avec BLAST**".

Pour plus de détails, vous pouvez cliquer sur l'onglet "**Alignements**", qui affiche un à un chacun des alignements entre votre protéine de requête et les protéines similaires trouvées dans UniprotKB/Swiss-Prot.

- Observez les positions de début/fin de la séquence requête et de la protéine similaire ("Subject") dans les alignements.
- Évaluez les nombres et pourcentages d'identités, de positifs et de gaps.
- Retrouvez les correspondances entre ces chiffres donnés et les caractères de l'alignement.
- Évaluez également la significativité de l'alignement, sur base de la E-valeur ("Expect").

The screenshot displays the BLAST interface on the Ametice platform, showing two alignment results. The first result is for **Streptococcus mutans UA159** (Sequence ID: **P08987.3**, Length: 1476, Number of Matches: 1). The second result is for **Streptococcus downei** (Sequence ID: **P1001.1**, Length: 1597, Number of Matches: 1). Red circles highlight specific alignment details: the 'Query' and 'Sbjct' labels, the 'Expect' value, the 'Identities' percentage, and the 'Positives' percentage.

**Alignment 1: 404 to 795**

Score	Expect	Method	Identities	Positives	Gaps
149 bits (37)	2e-35	Compositional matrix adjust.	118/406 (29%)	183/406 (45%)	79/406 (19%)

Query 434 FEF LVGNDLTIREDVQEQELNWKYLLDFG-----FDGFRIDAASHTNWKD 483  
+EFL+ ND+D VQ EQLNW ++L+FG FD RDA ++D+L+  
Sbjct 404 YEFL LANDVNSNPVQAEQLNWLHFLNFGNIYANDPDANFDSIRVDADVNDADLLQI 463

Query 484 E-----VTQRLNHFAGEDVNEHLSYIESYVTQQVDFLSNNYGMQADAGPFGMLFSGFR 539  
+ H + N+HLS E++ +L + + MD LFS +  
Sbjct 464 AGDY LKAAKGIKNDKAANDHLSILEAWSNDTPYLHDDGDMINMDNKLRLSLFLSLAK 523

Query 540 ----DWAPLRYAFEASLIDRVNGGP----ALPNWSFVNNDQEHNLVTVPTEEEAGGYEP 593  
+ SL++R + A++SF+ HD E L+ + E P  
Sbjct 524 PLNQRSGMNPLITNSLVNRTDDNAETAAPVSYFIRAHDSVQDLIRDIKAE----INP 579

Query 594 NSQPYEL-----RQLEKYDADRNSVEKQWAPHNVPAMYAILLTQDTYPTVFYGMFVS 647  
+ E Y+ D + EK++ +N YALLT K +Y+ V+YGMF  
Sbjct 580 NVGYSTMEIEIKKAFIYNKOLLATEKKYHTNALSALLTNKSSVPRVYGMFTD 701

Query 648 SKPYMSTPTPYRDDIVNLKQKQFAKEQVIRYENSTGNGEDLVSNIRLN----- 739  
YH+ T + I +LK R++ G Q+R N G++ ++++R G  
Sbjct 640 DGQYMAHKTINYEATETLLKARIKYVSGGQAMR--NQQVGN--EITTSVRYGKALKAT 695

Query 702 -----DRKTGVAVVAGNNPAL-----DTTITVDMGAQRNQWFDAMGYQPERLKTDK-- 749  
R +GAV+ GNNP+L + V+MGA H+NQ Y+P L TD  
Sbjct 696 DTGDRTRTSGVAVIEGNNPSRLRLKASDRVNVNMGAAHKNQ-----VNPRLLTNDNGI 749

Query 750 -----DGR-----TVQVKGTQNVGVKYLAAW P 774  
G L +KG N V GYL W/P  
Sbjct 750 KAYHSDQEAAGLVRYTNDRGELIFTAIDIKYANPQVSGYLGW P 795

**Alignment 2: 236 to 798**

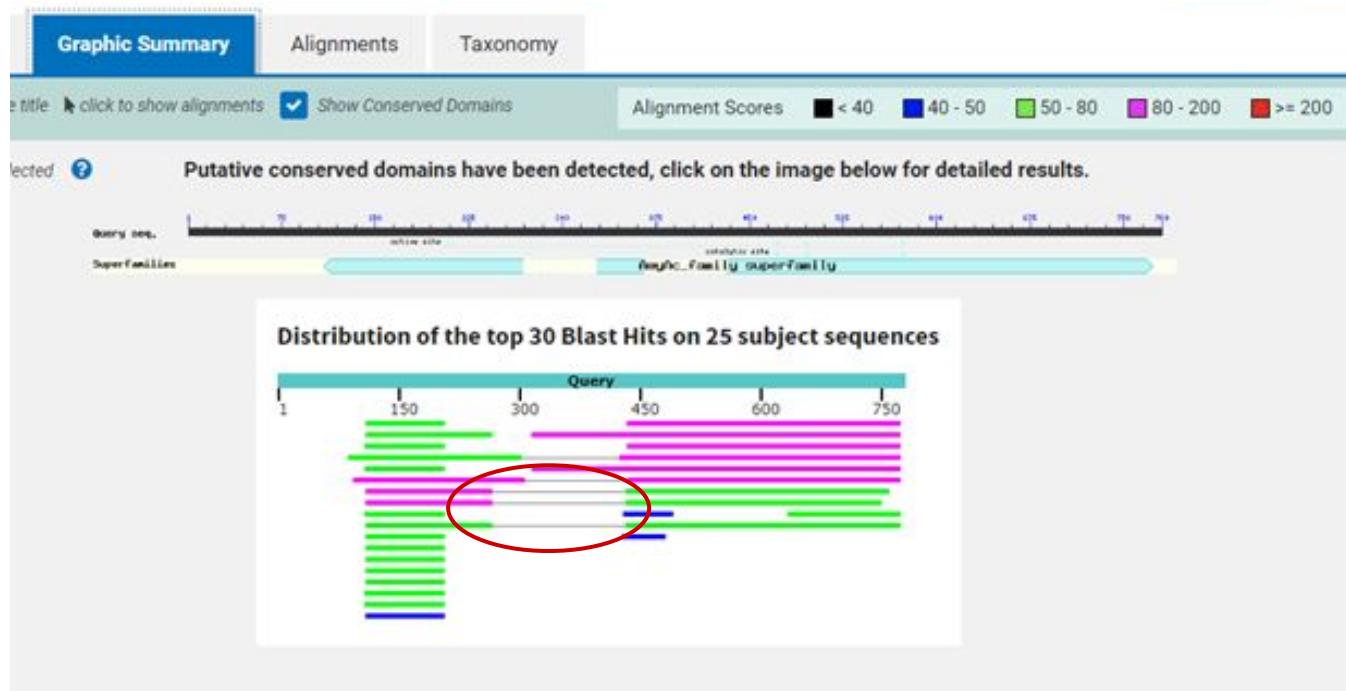
Score	Expect	Method	Identities	Positives	Gaps
144 bits (364)	3e-34	Compositional matrix adjust.	154/572 (27%)	243/572 (42%)	122/572 (21%)

Query 316 IDGYLLADTWFAVEN-----AESNAVYAPLFLYY-----EPRNGV-----VEQ 355  
ID YL AD+W+ ++ ES + PL + + E RN V +++  
Sbjct 236 IDNYLTADSWYRPKSLKDGKGTWTESSKDDFRLLMAWPDTEKRNRYNVMNKVGVGIDK 295

Query 356 TFMDFARENGYTGSDIEDIRATMLAELRMTPNP--IGPLMDEYLAAQPGYKSKSE----DD 409  
+ + + T + F + + + + + N + + + + + NP + + + CF N

## BLAST – Graphic summary

Cliquez maintenant sur l'onglet “**Graphic Summary**” pour retrouver ces informations avec une représentation plus visuelle. Vous remarquerez que certaines régions alignées sont reliées par un fin trait gris; il s'agit de cas où une même protéine subject (Number of Matches: 2) comporte à plusieurs régions disjointes similaires à la protéine requête des alignements. La barre grise marque l'intervalle qui sépare les deux régions alignées sur la protéine requête.





## Visualisation des alignements multiples – MSA Viewer

L'interface web du NCBI-BLAST propose de voir un alignement multiple (**Multiple Sequence Alignment** ou **MSA**) en cliquant sur **MSA Viewer**. Cependant attention : il ne s'agit pas d'un vrai alignement MSA mais de l'alignement de chaque résultat sur la séquence "requête".

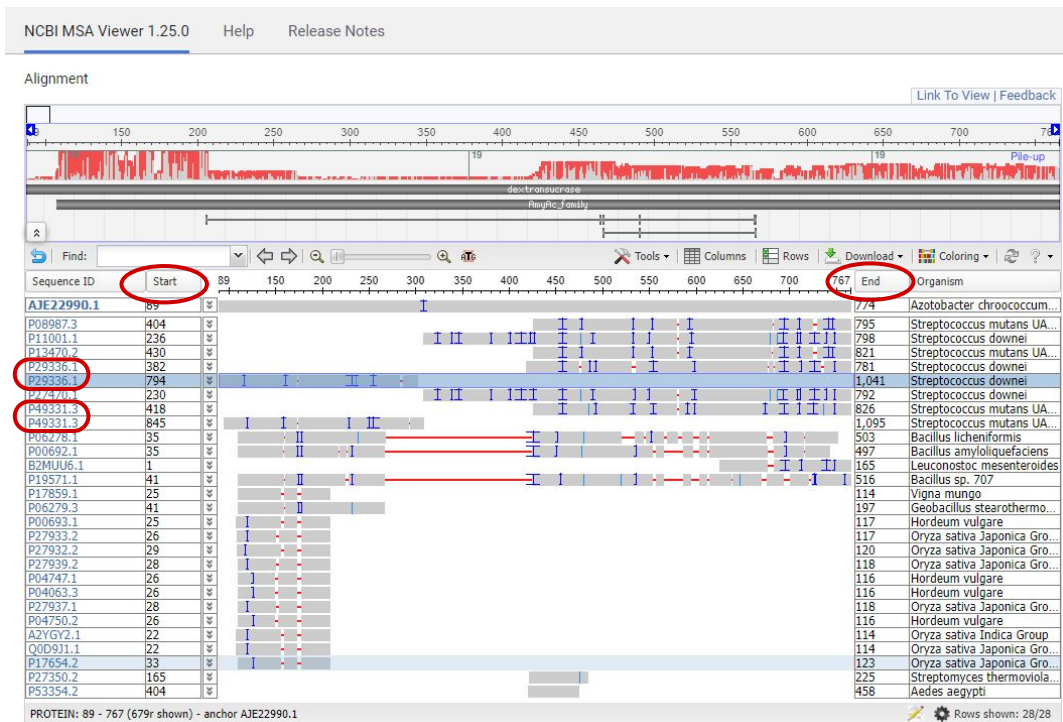
MSA viewer propose donc une compilation d'alignements par paires.

**Note:** un vrai MSA peut être généré grâce au logiciel Cobalt en cliquant sur **Multiple Alignment**.

- Retournez dans l'onglet **Descriptions** et cliquez sur le lien **MSA viewer**.
- Comparez cette présentation avec le **Graphic summary**.
- Regardez bien attentivement les positions de début et de fin des fragments alignés dans le cas des protéines avec deux hits retournés par BLAST, au besoin faites un petit schéma.

## Que remarquez-vous ?

(observez, réfléchissez, puis consultez l'explication à la diapo suivante)



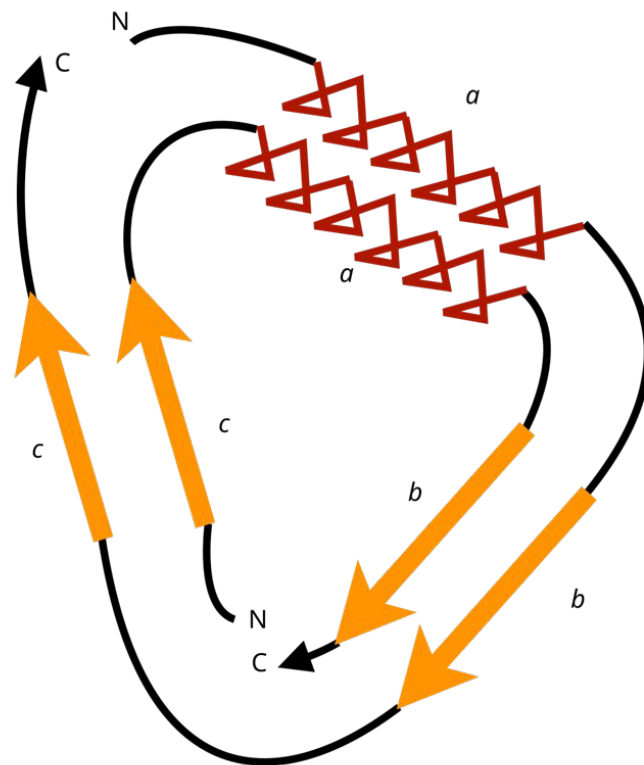
On appelle cet évènement une [permutation circulaire](#).

La **permutation circulaire** au sein d'un groupe de protéines fait référence à un réarrangement de l'ordre des domaines ou des motifs dans la structure de ces protéines. Le résultat est une organisation protéique différente, mais une forme tridimensionnelle (3D) globalement similaire.

En génomique, la permutation circulaire est souvent étudiée à l'aide de techniques bioinformatiques pour analyser la structure et l'évolution des protéines en comparant les arrangements entre différentes espèces.

**Répondez aux questions du Questionnaire 2.**

**Alignements multiples et MSA viewer.**



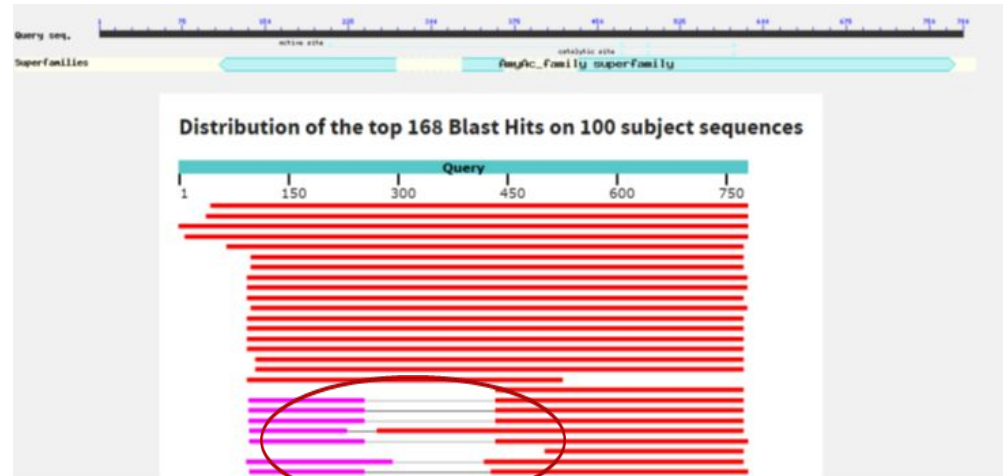


# Construction d'un jeu de séquences homologues pour réaliser un alignement multiple

Dans un deuxième temps, nous allons chercher un ensemble plus large d'homologues à partir d'une plus grande base de données. A partir de la fiche NCBI de la protéine dextranase ([AJE22990.1](https://www.ncbi.nlm.nih.gov/protein/AJE22990.1)), relancez une recherche BLAST en choisissant la base de données “**refseq\_select**” dans le menu déroulant “**Databases**”.

- Dans les représentations “**Graphic Summary**” et “**Alignments**”, repérez dans ce nouveau BLAST l'endroit de la “transition” entre protéines homologues “normales” et “permutées”.
- En appliquant la même méthode que précédemment, identifiez la **première séquence “permutée”** à l'aide du “Graphic Summary”.

Considérez cette séquence comme un seuil qui vous permettra de télécharger uniquement les séquences « normales » (alignées sans permutation).



# Construction d'un jeu de séquences homologues pour réaliser un alignement multiple

Basez-vous sur ce seuil pour télécharger uniquement le groupe de protéines homologues “normales”. Pour cela, retournez dans l'onglet “**Descriptions**”, puis sélectionnez les séquences “normales” dans le début du tableau

**Astuce :** pour effectuer facilement votre sélection, **désélectionnez toutes les séquences** en cliquant l'option “**Select all**”, puis sélectionnez la séquence au-dessus de la séquence seuil. Ensuite, remontez en début de liste et sélectionnez la première séquence de la liste en maintenant la touche Shift enfoncée. Cette action vous permet de sélectionner en un clic toutes les séquences dans l'intervalle.

- Pour télécharger cet ensemble de séquences au format FASTA, cliquez sur le menu déroulant “**Download**” et sélectionnez **FASTA (complete sequence)**.
- Sauvegardez ce fichier afin de pouvoir l'utiliser dans l'exercice 2.

**Sur Ametice, répondez au questionnaire 3.**  
**Sélection des séquences homologues non**  
**permutées**

Molecule type: amino acid  
Query Length: 780  
Other reports: [Distance tree of results](#) [Multiple alignment](#) [MSA viewer](#) ?

Descriptions | Graphic Summary | Alignments | Taxonomy

Sequences producing significant alignments

☒ select all 100 sequences selected

Description	Scientific Name
<input checked="" type="checkbox"/> glycoside hydrolase family 70 protein [Azotobacter chroococcum]	<a href="#">Azotobacter chroococcum</a>
<input checked="" type="checkbox"/> glycoside hydrolase family 70 protein [Azotobacter salinestris]	<a href="#">Azotobacter salinestris</a>
<input checked="" type="checkbox"/> glycoside hydrolase family 70 protein [Pseudomonas cavemicola]	<a href="#">Pseudomonas cavemicola</a>
<input checked="" type="checkbox"/> glycoside hydrolase family 70 protein [Frateruia defendens]	<a href="#">Frateruia defendens</a>
<input checked="" type="checkbox"/> glycoside hydrolase family 70 protein [Paenibacillus vietnamensis]	<a href="#">Paenibacillus vietnamensis</a>
<input checked="" type="checkbox"/> glycoside hydrolase family 70 protein [Paenibacillus caui]	<a href="#">Paenibacillus caui</a>
<input checked="" type="checkbox"/> glycoside hydrolase family 70 protein [Paenibacillus baumannii]	<a href="#">Paenibacillus baumannii</a>

Download ▾ Select column

- FASTA (complete sequence)
- FASTA (aligned sequences)
- GenBank (complete sequence)
- Hit Table (text)
- Hit Table (CSV)
- Text
- Descriptions Table (CSV)
- XML
- ASN.1

## Exercice 2. Alignement multiple à l'EBI

L'European Bioinformatics Institute (EBI) met à disposition un ensemble d'outils d'alignement multiple sur une page dédiée aux [MSA](#). Observez les descriptifs des différentes méthodes. Ces outils ont été développés par différents groupes de recherches, pour affiner les résultats dans des cas un peu particuliers et propres à leurs questions de recherche/objectifs (ex : un grand nombre de séquences, une meilleure précision, un alignement structural, etc) mais dans l'ensemble les résultats seront très similaires. Lors de ce TP, nous allons utiliser **ClustalO** pour aligner les séquences.

- Allez sur la page [MSA](#) de l'EBI.
- Cliquez sur **“Launch Clustal Omega”**.
- Assurez-vous que **“protein”** est sélectionné dans le champ **“Sequence type”**.
- Cliquez sur **“Choose File”** et téléversez le fichier fasta obtenu dans l'exercice 1.
- Nommez votre requête **“ClustalW”** dans le champ 'Title'.
- Dans **“Parameters”**, cliquez sur **“More options”** et dans le menu **Order** sélectionnez **“Input”**. Cette option permet de conserver l'ordre des séquences du fichier FASTA dans l'alignement.
- Cliquez sur **submit**.
- Observez le résultat.
- Cliquez sur **Resubmission**.

The screenshot shows the Clustal Omega web interface. The 'Input sequence' section has 'Protein' selected under 'Sequence Type'. A text box contains an example sequence. Below it, the 'Choisir un fichier' button is circled in red. The 'Parameters' section shows 'OUTPUT FORMAT' set to 'ClustalW with character counts' (circled in red). Other parameters like 'DEALIGN INPUT', 'MBED-LIKE CLUSTERING GUIDE-TREE', 'MBED-LIKE CLUSTERING ITERATION', 'COMBINED ITERATIONS', 'MAX GUIDE TREE', 'MAX HMM ITERATIONS', 'ORDER' (set to 'Input', circled in red), 'DISTANCE MATRIX', and 'OUTPUT GUIDE TREE' are also visible. At the bottom, there is a 'Less options' link.

## Exercice 2. Alignement multiple à l'EBI

- Relancez un calcul en sélectionnant cette fois-ci l'option **"Pearson fasta"** dans le menu **OUTPUT FORMAT** et en renommant votre requête **"Pearson"** dans le champ **"Title"**. Cliquez à nouveau **Submit**.
- Une fois cette opération effectuée, cliquez sur le bouton **"Your Jobs"**: un tableau de l'historique de vos alignements s'affiche. **Ouvrir les deux alignements** en format Pearson et clustalW dans deux fenêtres séparées (click droit > ouvrir dans une nouvelle fenêtre).

The screenshot shows the EBI Clustal Omega web interface. A green notification box states "YOUR JOB IS FINISHED" with a "Close" button. Below it, a message says "Please note that results can only be retrieved for jobs submitted within the last seven days." and provides a "Job ID: clustalo-I20241011-125516-0751-81092089-p1m". At the bottom, the "Your Jobs" button is circled in red, with a red arrow pointing to a separate window showing a table of job history.

Job Title (ID) [ALL]	Status	Last update	Delete
Clustal (clustalo-I20241011-125433-0155-49015409-p1m)	✓	2 minutes ago	🗑️
Pearson (clustalo-I20241011-125516-0751-81092089-p1m)	✓	1 minute ago	🗑️

Job status: ✓ Success | ✖ Failed, Error | ⓘ Not found

## Exercice 2. Alignement multiple à l'EBI

- Pour l'alignement clustalW, affichez la page '**Alignments**' : votre alignement multiple sera visible.
- Familiarisez-vous avec l'interface d'exploration de l'alignement multiple: **zoom**, glissement de la **fenêtre d'observation**, **schéma de coloration**.
- Testez différents schémas de coloration, notamment la possibilité de ne montrer que certaines catégories tels que les AA chargés, aromatiques etc.

**Sur Ametice, répondez au questionnaire 4. Alignement multiple**

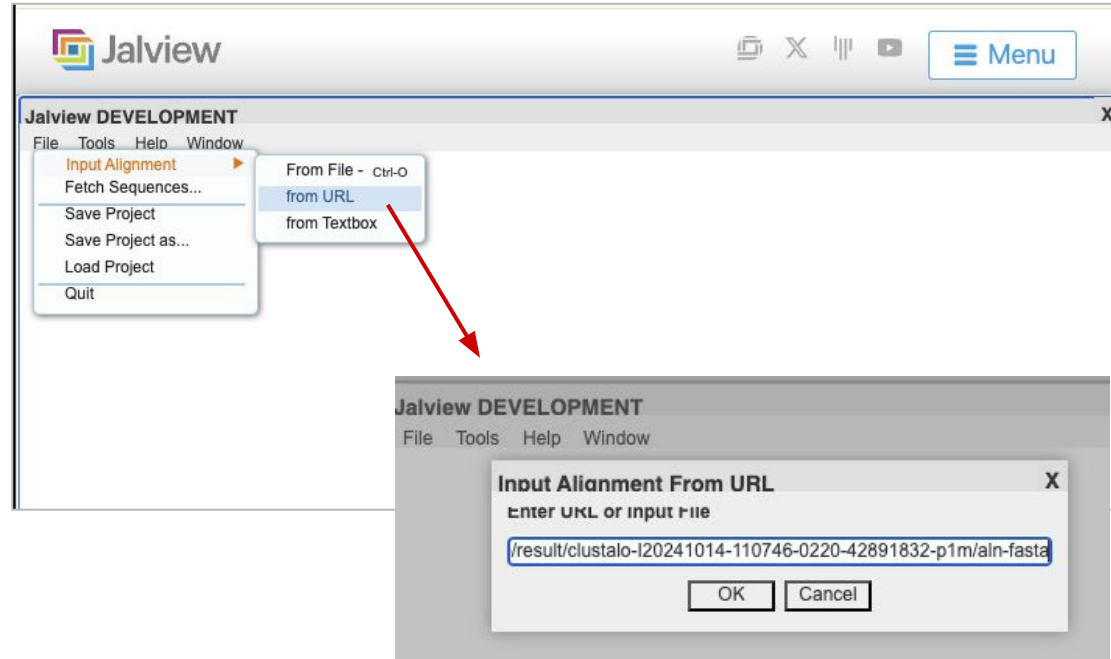
[illegible]

# Jalview – Outil de visualisation et d'édition d'un alignement multiple

La page web MSA ne nous permet pas de modifier ou de réordonner les séquences dans l'alignement.

Afin d'éditer cet alignement, allez, dans l'onglet “**Result viewers**”, copiez le lien de la sortie. Ce lien va nous permettre d'ouvrir l'alignement dans le programme d'alignement multiple [JALVIEW](#) pour visualiser et éditer des alignements.

- A l'ouverture d'une page [JALVIEW](#), une fenêtre “**Jalview development**” dédiée à l'application apparaît devant celle du navigateur. Vous pouvez positionner cette fenêtre en haut à gauche puis la redimensionner vers le bas à droite pour qu'elle occupe bien votre écran.
- Dans cette fenêtre, cliquez sur le menu “**File**”, sélectionnez “**Input alignment**” puis l'option “**From URL**”.
- Collez le lien de votre alignement ClustalW généré dans l'exercice 2.
- Cliquez sur **OK**.





Une nouvelle fenêtre montrant l'alignement multiple.

- Déplacez le curseur blanc au bas de la fenêtre pour avancer vers le coeur de l'alignement (dépassés les gaps initiaux, généralement peu informatifs)
- Modifiez l'affichage de l'alignement en changeant la coloration (menu **Colours**) en **BLOSUM62** ou **Pourcentage d'identité**. Vous devriez observer que les colonnes colorées sont celles qui affichent les valeurs les plus élevées de conservation dans le **profil de conservation** sous l'alignement.
- Essayez ensuite la coloration **ClustalX**, qui combine conservation et groupes de propriétés physico-chimiques des acides aminés.



# Jalview – Ré-ordonnancement des séquences en fonction de leur proximité sur un arbre

Les gaps peuvent provenir d'un événement évolutif réel d'insertion ou de délétion, pour s'en assurer il faut évaluer la cohérence entre les gaps de plusieurs protéines. Ainsi il est impossible de déterminer l'origine des gaps sur base d'un alignement par paire. Par contre, dans un alignement multiple, on peut dans certains cas évaluer si des gaps sont consistants avec un évènement d'insertion ou de délétion.

Pour mieux visualiser cette information, nous pouvons ré-ordonner les séquences dans JALVIEW afin de rapprocher dans l'alignement multiple celles qui sont les plus similaires.

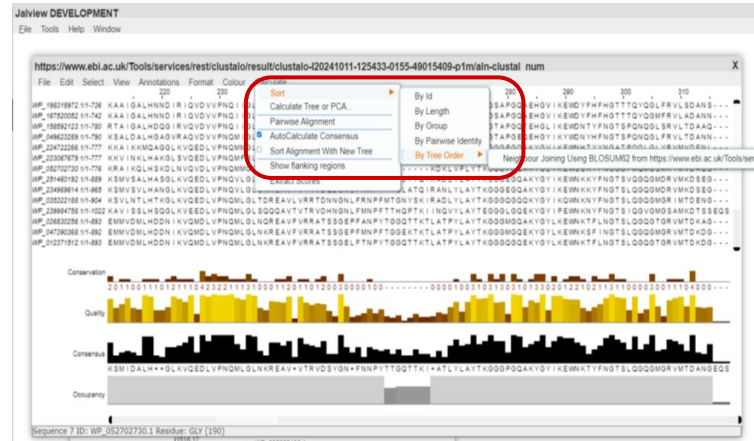
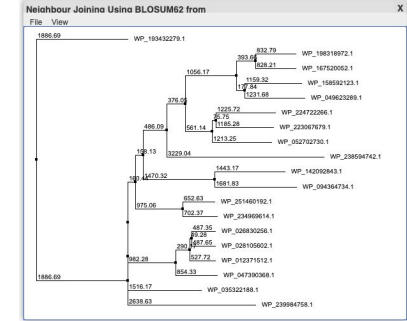
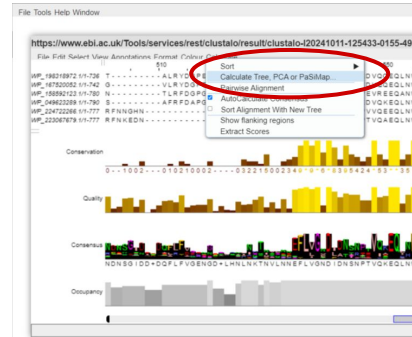
Pour cela nous procédons en deux temps

- Construction d'un arbre à partir de l'alignement multiple: menu **Calculate > Calculate Tree or PCA**, et laisser les paramètres par défaut.
- Ré-ordonnancement des séquences de l'alignement multiple en fonction de leur proximité dans l'arbre : **Calculate > Sort > By tree order**. Vous pouvez ensuite fermer la fenêtre contenant l'arbre.

Rappel : dans un arbre phylogénétique, la distance entre deux séquences se calcule en faisant la somme des longueurs des branches (horizontales sur la représentation de Jalview).

**Dans Ametice, répondez au Questionnaire 5.**

**Edition d'un alignement multiple dans JALVIEW.**





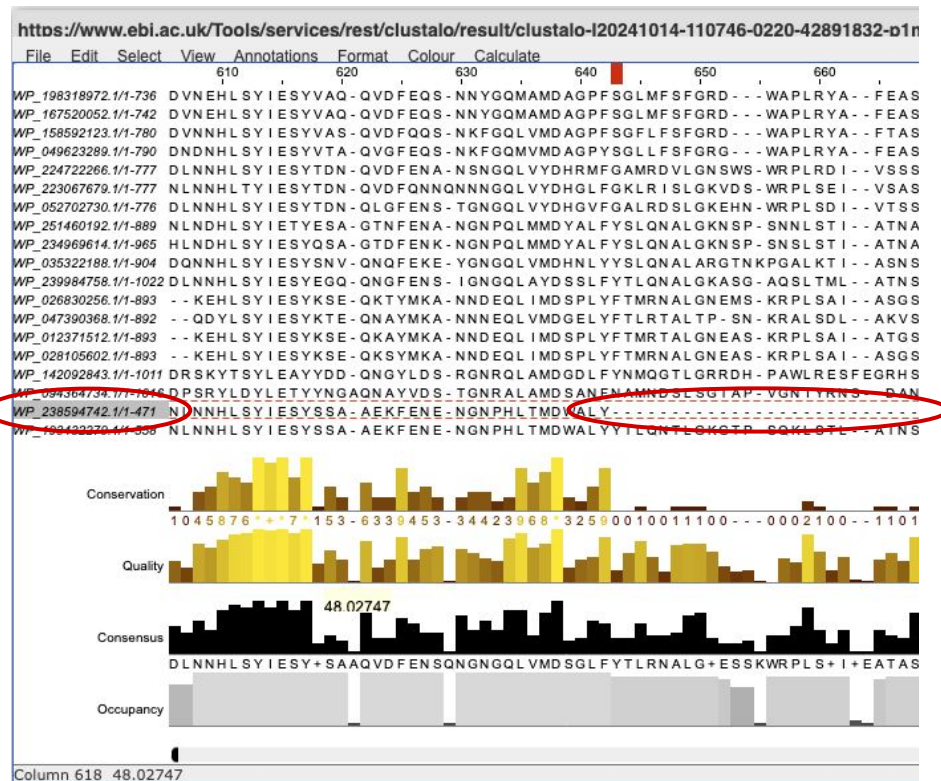
# JalView - Edition du fichier d'alignement multiple

Une des premières étapes lorsque l'on édite un alignement multiple est de rechercher d'éventuelles séquences fragmentaires et de les effacer/enlever. Dans notre cas, la séquence WP\_238594742.1 est beaucoup plus courte que les autres. Il convient donc de la retirer.

- Fermez la fenêtre de visualisation de l'alignement multiple dans JALVIEW, et ouvrez à nouveau votre fichier d'alignement multiple (**File > Input alignment > From URL**).
- Collez le lien de l'alignement ClustalW généré dans l'exercice 2. Cliquez sur **OK**.
- Sélectionnez la séquence WP\_238594742.1 en cliquant sur son identifiant.
- Utilisez le raccourci **Ctrl+X** pour la supprimer.

**Astuce:** au cas où vous devriez supprimer plusieurs séquences d'un alignement, JalView permet également de sélectionner :

- plusieurs séquences en cliquant successivement leurs identifiants tout en maintenant la touche Ctrl enfoncée
- un bloc de séquences en cliquant sur la première, puis en maintenant la touche Shift avant de cliquer sur la dernière.



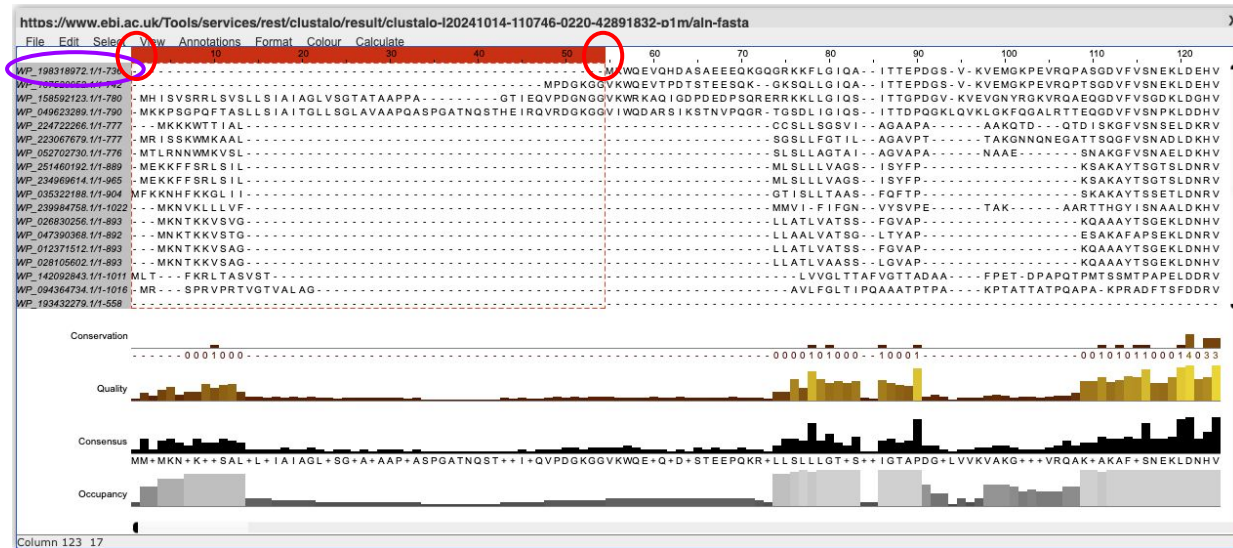
# Suppression des séquences terminales excédentaires

Une fois les protéines fragmentaires éliminées, on observe souvent dans les alignements que les portions N- et C-terminale sont moins conservées (plus variables). Cela peut s'expliquer par la présence d'un domaine fonctionnel plus conservé que ce qui le précède ou le suit. Nous souhaitons donc borner notre alignement en fonction de la longueur de **notre séquence référence** **WP\_198318972.1**. Pour cela, on doit supprimer les portions N- et C-terminales en amont et en aval de cette séquence référence.

- Repérez la séquence de référence.

En principe c'est la première du fichier que vous venez de recharger.

- Sélectionnez les colonnes à supprimer en cliquant au-dessus de la première position de l'alignement puis en étirant la sélection jusqu'à la position souhaitée (c'est à dire jusqu'à la position qui précède le premier acide aminé de la séquence référence).
- Utilisez le raccourci **Ctrl+X** pour supprimer la région sélectionnée.

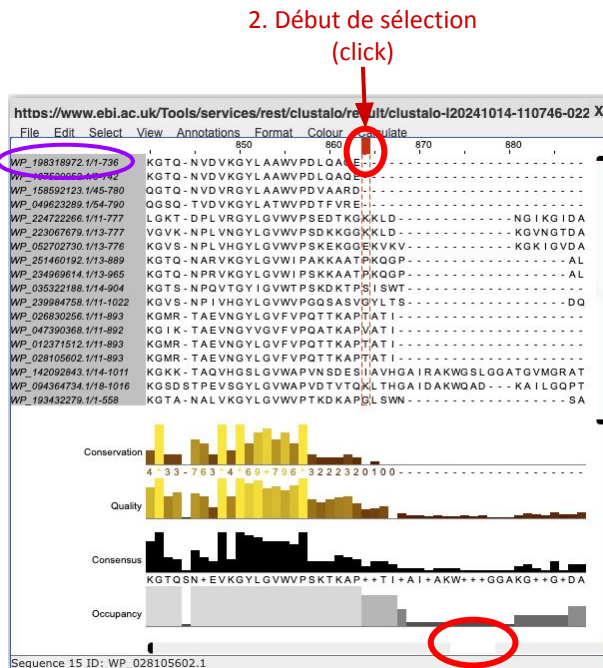


# Suppression des séquences terminales excédentaires

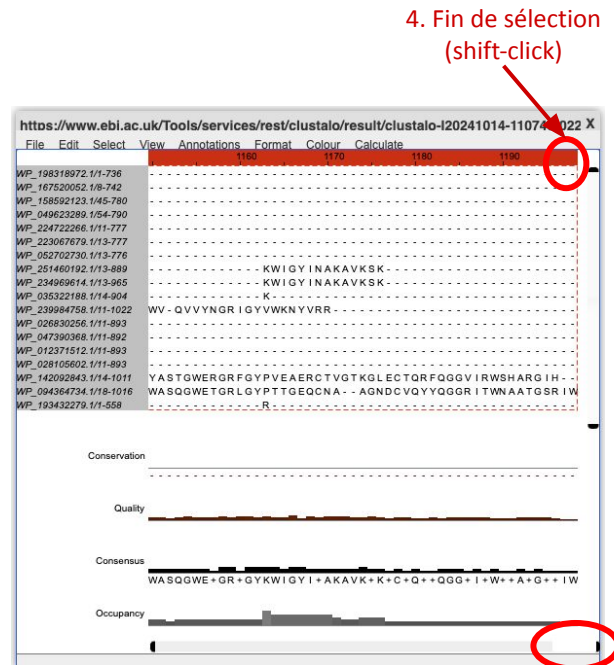
- Faite de même pour **supprimer la région C-terminale** qui dépasse de la séquence de référence.

**Astuce:** vous pouvez sélectionner un nombre de colonnes qui dépasse la taille de votre fenêtre en cliquant sur les en-têtes de colonnes tout en maintenant la touche **Shift-Click enfoncée**

1. **Positionnez le curseur** (rectangle blanc en bas de la fenêtre) sur la première colonne qui suit la protéine référence.
2. **Cliquez sur l'en-tête de la première colonne à sélectionner** (celle qui suit immédiatement la séquence référence)
3. **Déplacez le curseur** jusqu'à la dernière colonne à supprimer (dans votre cas, la dernière de l'alignement)
4. **Shift-click sur l'en-tête de cette colonne.**
5. **Ctrl-X** pour supprimer toutes les colonnes sélectionnées



1. Positionner



Recherchez dans l'alignement, les acides aminés catalytiques D427, E469 et D528 de la séquence référence [WP 198318972.1](#). Pour cela, vous pouvez vous aider des acides aminés qui précèdent et qui suivent chacun de ces sites catalytique:

- D427 est compris entre I et A
- E469 est compris entre I et S
- D528 est compris entre H et Q

**Astuce:** pour trouver ces trois fragments dans l'alignement, sélectionnez la séquence référence, placez vous en début d'alignement et lancez une recherche de caractères à l'aide de l'outil Ctrl-F.

**Sur Ametice, répondez au questionnaire 6. Recherche de sites catalytiques**

C'est fini !