

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 (requête, "query") avec toutes les séquences d'une base de données ("subject")
 - 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 substitution** – 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

Matrice de substitutions

- Une **matrice de substitution** associe un score à chaque paire de résidus qu'on peut trouver dans un alignement.
 - Chaque ligne et chaque colonne représente l'un des résidus (4 nucléotides, 20 acide aminés).
 - La **diagonale** correspond aux **identités**.
 - Le triangle inférieur correspond à des substitutions.
 - Le triangle supérieur est symétrique au triangle inférieur, il n'est pas nécessaire d'indiquer les nombres.
 - Les **scores négatifs** sont considérés comme des pénalités associées à certaines substitutions qu'on n'observe que rarement dans les alignements. Les algorithmes d'alignements tenteront donc d'éviter ces substitutions.
 - Les **scores positifs** correspondent à des substitutions qu'on observe plus souvent que prévu, dans les alignements d'un grand nombre de séquences. Ceci suggère que ces substitutions particulières sont moins dommageable que d'autres, et on les qualifie donc de « **substitutions conservatives** » ou encore de « **mutations ponctuelles acceptées** » (**PAM**).

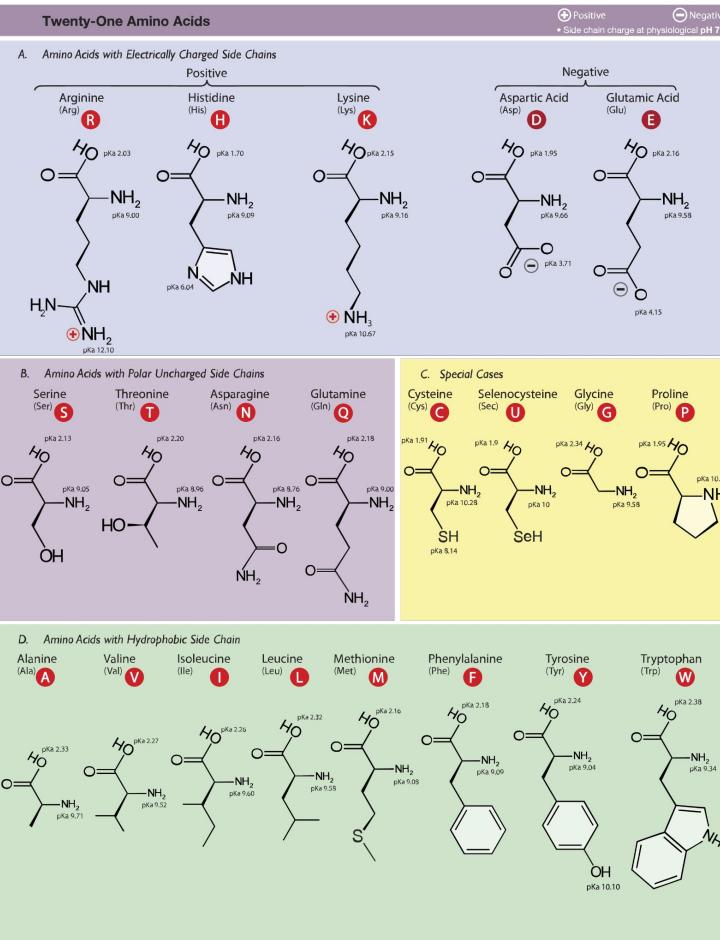
Matrice de substitutions entre nucléotides

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

Matrice de substitutions entre acides aminés

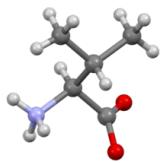
Rappel – Nomenclature et composition des acides aminés

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

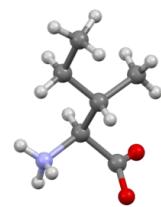


Similarités chimiques entre acides aminés

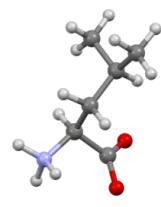
Valine (Val)



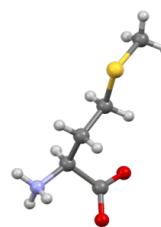
Isoleucine (Ile)



Leucine (Leu)

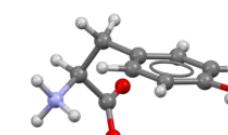


Méthionine (Met)

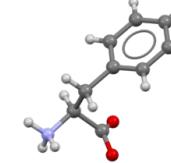


Hydrophobes

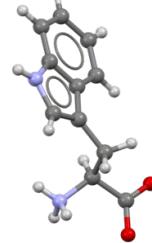
Tyrosine (Tyr)



Phénylalanine (Phe)

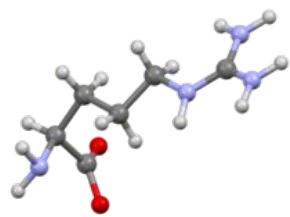


Tryptophane (Trp)

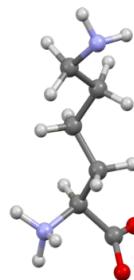


Aromatiques

Lysine (Lys)

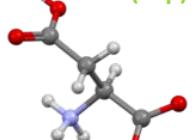


Arginine (Arg)

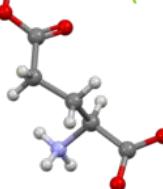


Chargés +

Acide aspartique (Asp)



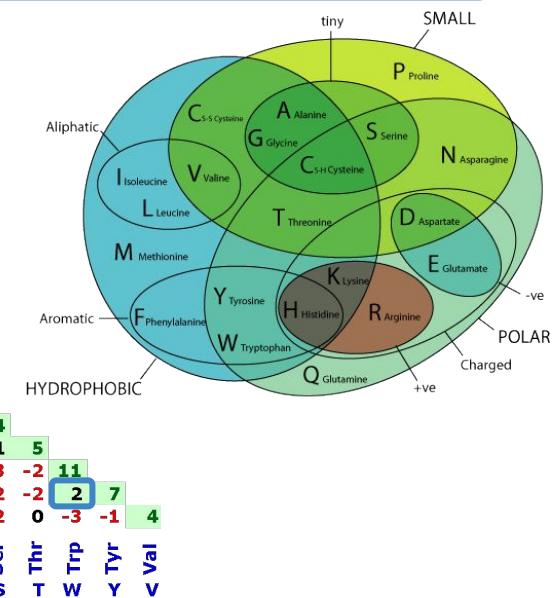
Acide glutamique (Glu)



Chargés -

Matrice de substitutions entre acides aminés

	Ala	A	R	N	D	C	C	Q	Gln	E	Glu	Gly	His	H	Ile	I	Leu	L	Lys	M	Phe	F	Pro	P	Ser	S	Thr	T	W	Trp	Y	Tyr	V	Val	
Ala	4																																		
Arg	-1	5																																	
Asn	-2	0	6																																
Asp	-2	-2	1	6																															
Cys	0	-3	-3	-3	9																														
Gln	-1	1	0	0	-3	5																													
Glu	-1	0	0	-4	2	5																													
Gly	0	-2	0	-1	-3	-2	-2	6																											
His	-2	0	1	-1	-3	0	0	-2	8																										
Ile	-1	-3	-3	-3	-1	-3	-3	-4	-3	4																									
Leu	-1	-2	-3	-4	-1	-2	-3	-4	-3	2	4																								
Lys	-1	2	0	-1	-3	1	1	-2	-1	-3	-2	5																							
Met	-1	-1	-2	-3	-1	0	-2	-3	-2	1	2	-1	5																						
Phe	-2	-3	-3	-3	-2	-3	-3	-3	-1	0	0	-3	6																						
Pro	-1	-2	-2	-1	-3	-1	-1	-2	-2	-3	-3	-1	-2																						
Ser	1	-1	1	0	-1	0	0	0	0	-1	-2	-2	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1				
Thr	0	-1	0	-1	-1	-1	-1	-2	-2	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1				
Trp	-3	-3	-4	-4	-2	-2	-3	-2	-2	-3	-2	-3	-2	-3	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1				
Tyr	-2	-2	-2	-3	-2	-1	-2	-3	2	-1	-1	-2	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1				
Val	0	-3	-3	-3	-1	-2	-2	-3	-3	3	1	-2	1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1				



Exemple d'alignement par paire

- 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.

Note: le mode de représentation des identités, substitutions conservatives et gaps varie d'un outil à l'autre.

>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 KFGGSSLADVKCYL^{LRVAG}IMAEY^{SQ}PDDMM-VVSAAGSTTNQLINWLKLSQTDRLSAHQV 74
KFGG+S+A+ + LRVA I+ ++ + V+SA TN L+ ++ + + + + +
Sbjct: 5 KFGGT^SVANAER^LRVADILES^MAROGQVATVLSAPAKITNHLVAMIEKTISGQDALPNI 64

Query: 75 QQTLRRYQC^DLISG^LLPABEADSL--ISAFVSDLERLAALLDSGIN-----DAVYAEVV 126
R + + L++GL A+ L + FV + GI+ D++ A ++
Sbjct: 65 SDAERIF-AELLTGLAAACPGFPLAQLKTFVDQEFAQIKHVLHG^ILLGQCPD^SINAALI 123

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

Query: 184 PGKRLVVTGFISRNNAGETVLLGRNGSDYSATQIGALAGVSRVTIWSDVAGVYSADPRKV 243
+++ GF + N GE V+LGRNGSDYSA + A IW+DV GVY+ DPR+V
Sbjct: 184 ---MVL^MAGFTAGNEKGELVVLGRNGSDYSAAVLAACLRADCCEIWTDVDGVYTC^DPRQV 240

Query: 244 KDACLLPLLRLDEASELARLAAPVILHARTLQPVSGSEIDLQLRC^SYTPDQ----GSTRI 298
DA LL + EA EL+ A VLH RT+ P++ +I ++ + P G++R
Sbjct: 241 PDARLLKSM^SYQEAMEL^SYFGAKVLHPRTITPIAQFQIPCLIKNTGNPQAPGTLIGASRD 300

Query: 299 ERVLASGTGARI^TSHDDVCLIEFQVPASQDFKLAHKEIDQILKRAQVRPLAVGVHNDRQ 358
E L + +++ +++ + P + + RA++ + + +
Sbjct: 301 EDEL^P---VKGISNLNNMAMFSVSGPGMKGMVGMAARVFAAMSRARISVVLITQSSSEY 356

Query: 359 LLQFCYTSEVADSALKIDEA-----GLPGELRLRQGLALVAMVGAGVTRNPLHCHRF 411

Résultat de BLAST – Requête peptidique vs DB de peptides

Exemple de résultat de recherche par similarité de séquences.

- Requête (**query**): metA
- 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 ≥ 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 KFGGSSLADVKCYLRVAGIMAEYSQPDDMM-VVSAAGSTTNQLINWLKLSQTDRLSAHQV 74
Sbjct: 5 KFGGTsvanaerflrvadilesnarqgqvatvlsapakitnHLVAMIEKTISGQDALPNI 64

Query: 75 QQTLRRYQCQLISGLLPAEAEADSL--ISAFVSDLERLAALLDSGIN-----DAVYAEVV 126
Sbjct: 65 SDAERIF-AELLTGLAAAQPGFPLAQLKTFVDQEFAQIKHVLHGISLLGQCPDSINAALI 123

Query: 127 GHGEVWSARLMSAVLNQQGLPAAWLDAREFLRAER---AAQPQVDEGLSYPLLQQLLVQH 183
Sbjct: 124 CRGEKMSIAIMAGVLEARGNVTVIDPVEKLLAVGHYLESTVDIAESTRRIAASRIPADH 183

Query: 184 PGKRLVVTGFISRNNAGETVLLGRNGSDYSATQIGALAGVSRTIWSDVAGVYSADPRKV 243
Sbjct: 184 ---MVL MAGFTAGNEK GELVVLGRNGSDYSAAVLAACLRADCCEIWTDVDGVYTCDPRQV 240

Query: 244 KDACLLPLLRLDEASELARLAAPV L HARTLQPVSGSEIDLQLRC SYTPDQ----GSTRI 298
Sbjct: 241 PDARLLKSMSYQEAME LSYFGAKVLHPRTITPIAQFQI PCLIKNTGNPQAPGTLIGASRD 300

Query: 299 ERVLASGTGARIVTS HDDVCLIEFQVPASQDFKLAHKEIDQILKRAQRPLAVGVHNDRQ 358
Sbjct: 301 EDEL P----VKGI S NLNNMAMFSVSGPGMKGMVGMAARVFAAMS RARISVVLITQSSSEY 356

Query: 359 LLQFCYTSEVADSALKILDEA-----GLPGELRLRQGLALVAMVGAGVTRNPLHCHR F 411

Tutoriel et exercices

Exercice 1. Alignements par paires et multiples au NCBI

Nous travaillerons à partir de la protéine **dextranucrase** ([AJE22990.1](#)) de la bactérie *Azotobacter chroococcum*. Cet enzyme catalyse notamment la biosynthèse du dextrane à partir du sucre. Pour collecter des séquences protéiques homologues de la dextranucrase, 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 [AJE 22990.1](#).
- 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, le résultat devrait s'afficher.

BLAST – Observer, comprendre le tableau de résultats

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

- Combien de séquences retourne BLAST?
 - Observez l'étendue des e-valeurs (Expect). Comment interprétez-vous les premiers et les derniers alignements en terme de significativité statistique ?
 - 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).

National Library of Medicine
National Center for Biotechnology Information

Log in

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 gb|AJE22990.1|

RID GMBW1SYS013 Search expires on 10-13 16:12 pm Download All ▾

Program BLASTP ? Citation ▾

Database swissprot See details ▾

Query ID AJE22990.1

Description dextrantraserase [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 E value Query Coverage

to to to to

Filter Reset

Descriptions Graphic Summary Alignments Taxonomy

Sequences producing significant alignments Download Select columns Show 100 ▾ ?

select all 25 sequences selected

	Description	Scientific Name	Max Score	Total Score	Query Cover	E value	Per. Ident.	Acc. Len.	Accessory
✓ RecName: Full=Glycosyltransferase-I; Short=GTF-I; AltName: Full=Dextrantraserase; AltName: Full=Sucrose 6-glucos... Streptococcus m...	Streptococcus m...	149	149	43%	2e-35	29.06%	1476	P08987.3	
✓ RecName: Full=Glycosyltransferase-I; Short=GTF-I; AltName: Full=Dextrantraserase; AltName: Full=Sucrose 6-glucos... Streptococcus do...	Streptococcus do...	144	144	58%	3e-34	26.92%	1597	P11001.1	
✓ RecName: Full=Glycosyltransferase-SI; Short=GTF-SI; AltName: Full=Dextrantraserase; AltName: Full=Sucrose 6-glucos... Streptococcus m...	Streptococcus m...	140	140	43%	9e-33	28.82%	1455	P13470.2	
✓ RecName: Full=Glycosyltransferase-S; Short=GTF-S; AltName: Full=Dextrantraserase; AltName: Full=Sucrose 6-glucos... Streptococcus do...	Streptococcus do...	139	215	72%	2e-32	27.64%	1365	P29336.1	
✓ RecName: Full=Glycosyltransferase-S; Short=GTF-S; AltName: Full=Dextrantraserase; AltName: Full=Sucrose 6-glucos... Streptococcus do...	Streptococcus do...	138	138	58%	3e-32	26.57%	1592	P27470.1	
✓ RecName: Full=Glycosyltransferase-S; Short=GTF-S; AltName: Full=Dextrantraserase; AltName: Full=Sucrose 6-glucos... Streptococcus m...	Streptococcus m...	136	217	71%	1e-31	26.51%	1462	P49331.3	
✓ RecName: Full=Alpha-amylase; AltName: Full=1,4-alpha-D-glucan glucanohydrolase; AltName: Full=BLA; Flags: Pre... Bacillus licheniformis	Bacillus licheniformis	82.8	145	62%	3e-15	34.78%	512	P06278.1	
✓ RecName: Full=Alpha-amylase; AltName: Full=1,4-alpha-D-glucan glucanohydrolase; Flags: Precursor [Bacillus amyl... Bacillus amyloliquif...	Bacillus amyloliquif...	82.4	145	61%	4e-15	35.58%	514	P00692.1	
✓ RecName: Full=Dextrantraserase 1; AltName: Full=Glucansucrase 1; AltName: Full=Sucrose 6-glucosyltransferase 1 [... Leuconostoc mes...	Leuconostoc mes...	79.3	79.3	18%	6e-15	30.81%	284	B2MUU6.	
✓ RecName: Full=Glucan 1,4-alpha-maltohexaosidase; AltName: Full=Exo-maltohexaohydrolase; AltName: Full=G6-am... Bacillus sp. 707	Bacillus sp. 707	79.0	137	63%	5e-14	33.13%	518	P19571.1	
✓ RecName: Full=Alpha-amylase; AltName: Full=1,4-alpha-D-glucan glucanohydrolase; Flags: Precursor [Vigna mungo] 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 “**Alignments**”, 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”).

Descriptions Graphic Summary Alignments Taxonomy

Alignment view Pairwise ? Restore defaults Download

25 sequences selected ?

Download GenPept Graphics ▾ Next ▾ Previous ◀ Descriptions

RecName: Full=Glucosyltransferase-I; Short=GTF-I; AltName: Full=Dextranucrase; AltName: Full=Sucrose 6-glucosyltransferase; Flags: Precursor [Streptococcus mutans UA159]

Sequence ID: P08987.3 Length: 1476 Number of Matches: 1

Range 1: 404 to 795 GenPept Graphics ▾ Next Match ▾ Previous Match

Score	Expect	Method	Identities	Positives	Gaps
149 bits(37%)	2e-35	3e-34 Compositional matrix adjust.	118/406(29%)	183/406(45%)	79/406(19%)

Query 434 FELVGNNDLTIRDVQEQEOLNWKQYLLDFG-----FDGFRIDAASHNTDWWLNU 483
+EFL+ ND+D VQ EQLNW +L++FG FD R+DA +++ D+L+
Sbjct 404 YEFLLANDVNDSNPVVQAQLNLWLFHMNGNIYANDPDDANFDSIRVADVNDAADLLI 463

Query 484 -----VTQRLNNHFAGEDVNNEHLSYIESVTTQVDFLQSNNYGQAMADAGPFSGLMFSGR 539
+ H + N-HLS +E++ +L + + MD L+FS +
Sbjct 464 AGDYLAALKGIIHKNDKAANDHSLLEAWSNDTTPYLDHDGDNMINNDKRLRSLLFSLAK 523

Query 540 ---DWAPLRYAFEAFLIDVRNGP----ALPNWSFVNHHQDEHNILTVPLTEEEAGGYEP 593
+ + SL+R + A+P+SF+ HD E L+ + E P
Sbjct 524 PLNRSGMPLITNSLVNRTDDNAETAAVPSYSFIRAHSEVQDLIRDOIKAEC----INP 579

Query 594 NSQPYEL-----ROLEKYYADDRNSVEKQWAPHNVPMAYAIIILTKDTVTPTVFYGDMS 647
N Y + E Y+ D + EK++ +N YA+LL K +V+ V-YGDMF
Sbjct 580 NVVGYSFTMEEIKKAEFYINQKDOLLAETKYYHTNLLTALSYNNKSPRYYGDMFTD 639

Query 648 SKPYMSTPTPYRDDIVILKLKRQFAKGEQVIRYENSNTGSNGEDLVSNIRLQN----- 701
VM+ T + I + LK R ++ G Q +R N G++ ++++++ R G
Sbjct 640 DGQYMAHKINTYEAETLKLARIKYVSGGQAMR---NQVGNS--EITTSVRYKGALKAT 695

Query 702 -----DRKTGVAVVAGNPAL-----DTTIVDPMGAHQRNQWFVDAMGYQPERLKTD-- 749
R +GVAV+ GNP+L + V+MG A+H+NQ Y+P L TD
Sbjct 696 DTGDRTRTSGVAVIEGNPLRNLKASDRVNVNMGAAHNQ---AVRNLLTTDNGI 749

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

Download GenPept Graphics ▾ Next ▾ Previous ◀ Descriptions

RecName: Full=Glucosyltransferase-I; Short=GTF-I; AltName: Full=Dextranucrase; AltName: Full=Sucrose 6-glucosyltransferase; Flags: Precursor [Streptococcus downei]

Sequence ID: P11001.1 Length: 1597 Number of Matches: 1

Range 1: 236 to 798 GenPept Graphics ▾ Next Match ▾ Previous Match

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

Query 316 IDGYLLADTWFAVEN-----AESENAYAPLFLYY----EERPNGV-----VQ 355
ID YL AD+W+ ++ E S + PL + + E RN V +++
Sbjct 236 IDNYLTADSWYRPKSILKDGKTWTTESSKDDFRPLLMAWPDTETKRNVYNYMNKVKVGDK 295

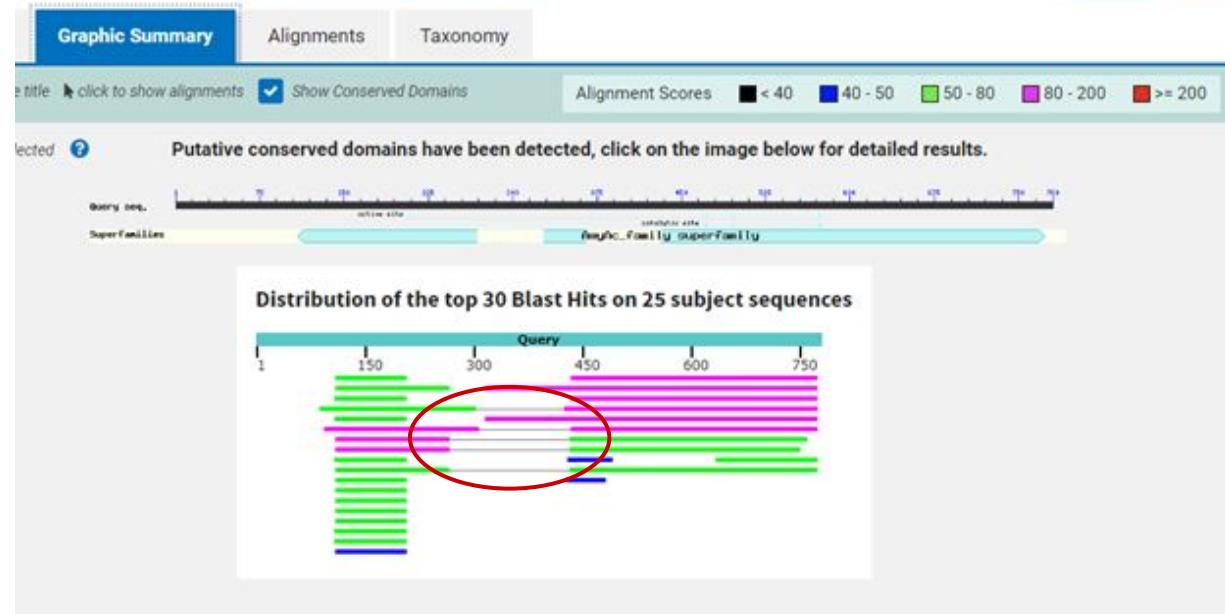
Query 356 TFMDFARENGTYGTGSDEDIATMLAELRMTNP-----IGPLMDEYLAQPGYSKKSE----DD 409
T+ + T + F + A + ++ N + + ++ NP ++ LSF - n

Related Information
AlphaFold Structure - 3D structure displays

BLAST – Graphic summary

Cliquez maintenant sur l'onglet “Graphic Summary” pour retrouver ces informations avec une représentation plus visuelle.

- Chaque rectangle coloré correspond à un alignement local (un “hit”).
- La couleur indique le degré de significativité, selon le score en bits (légende au dessus).
- Attention, plusieurs hits peuvent apparaître à la même hauteur sur le graphique mais correspondre à des protéines subject différentes.
- Les régions alignées reliées par un fin trait gris indiquent des cas où une même protéine subject 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.

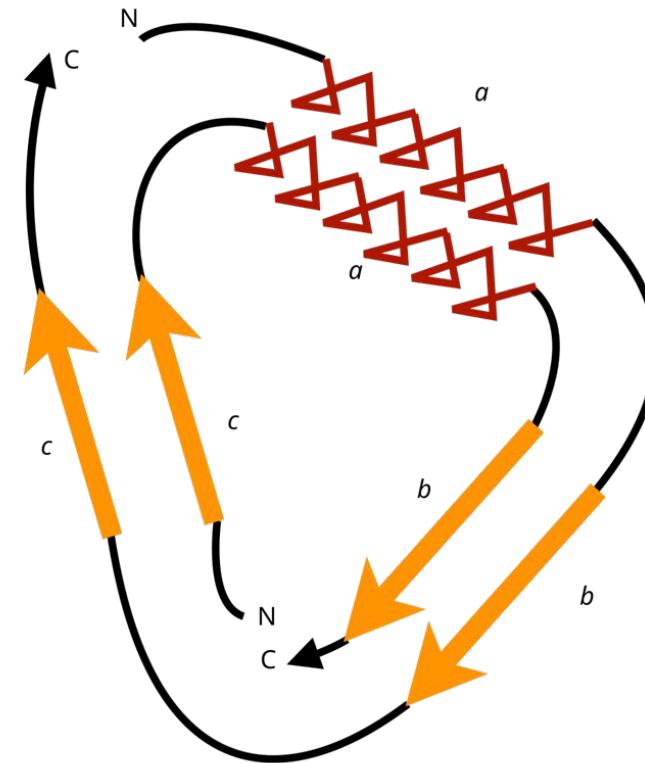


Permutation circulaire

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 topologique (succession des domaines sur la séquence) différente, mais une structure 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.**



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.

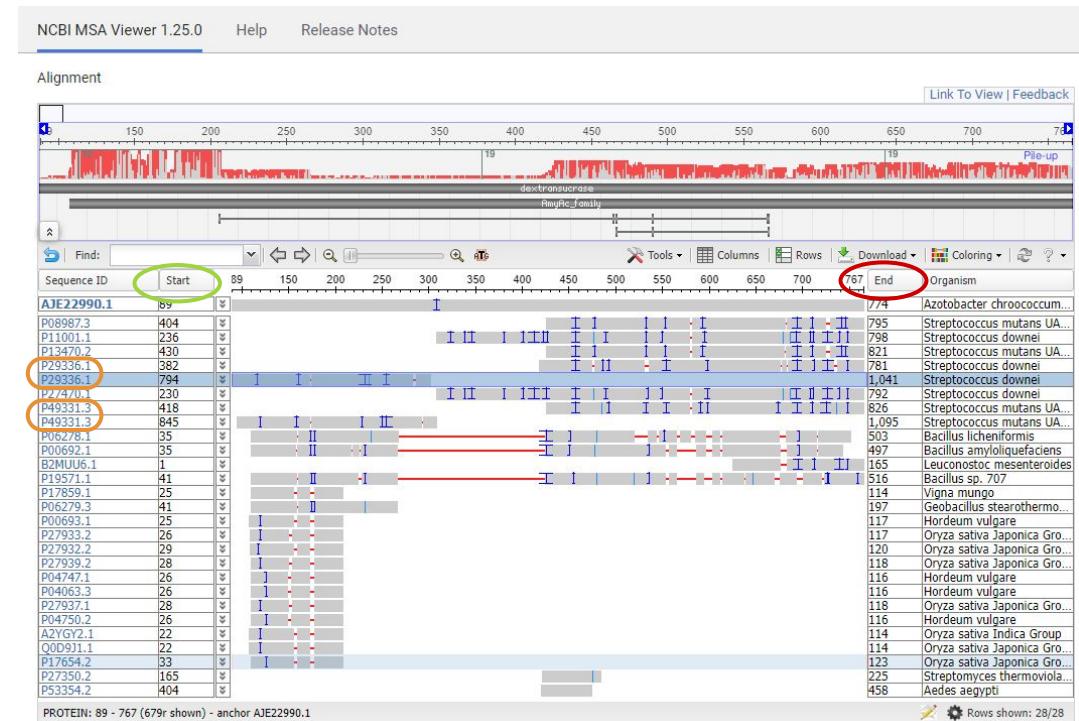
Facultatif: un vrai MSA peut être généré grâce au logiciel Cobalt en cliquant sur **Multiple Alignment**.

- Au-dessus du Graphic Summary, 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)

Rappel : AJE22990.1 est le numéro d'accession de votre protéine requête (query), les autres numéros d'accessions sont les protéines sujettes (subjects).

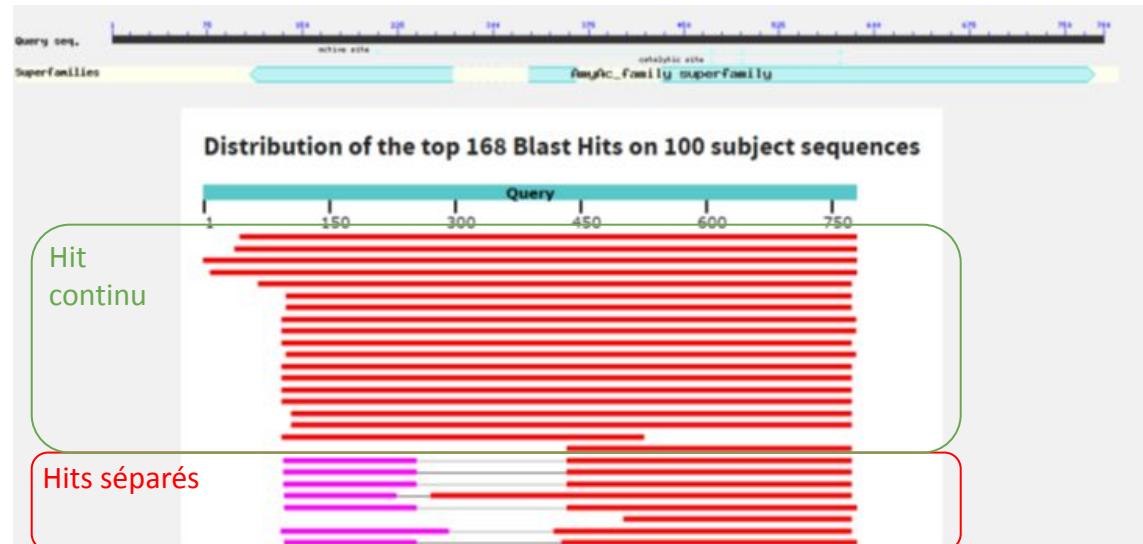


Construction d'un jeu de séquences homologues pour réaliser un alignement multiple

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 dextranucrasse ([AJE22990.1](#)), relancez une recherche BLAST en choisissant la base de données “**refseq_select**” dans le menu déroulant “**Databases**”.

Astuce : les séquences sélectionnées dans l'onglet “Descriptions” vont être utilisées pour les onglets “Graphic Summary”, “Alignements” et “Taxonomy” avec le même ordre (c'est-à-dire de l'alignement le plus significatif au moins significatif), ce qui permet de visualiser les différences entre les alignements successifs selon notre sélection (ex : nombre de hits, recouvrement, etc).

- Dans les représentations “**Graphic Summary**” et “**Alignments**”, repérez dans ce nouveau BLAST l'endroit de la “transition” entre protéines homologues avec un seul ou avec 2 hits séparés.
- En appliquant la même méthode que précédemment, identifiez la **première séquence “hits séparés”** à l'aide du “**Graphic Summary**”.
- Identifiez cette séquence comme un jalon, qui vous permettra ensuite de sélectionner uniquement les séquences avec un hit continu.



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

Basez-vous sur le hit “jalon” défini précédemment pour télécharger uniquement le groupe de protéines homologues avec un hit continu. Pour cela, retournez dans l’onglet “**Descriptions**”, puis sélectionnez les séquences “hit unique” dans le début du tableau. Vous pouvez vérifier votre sélection dans l’onglet “Graphic Summary” en vous assurant de n’avoir que des alignements avec un seul hit.

Astuce : pour effectuer facilement votre sélection:

- **désélectionnez toutes les séquences en cliquant l’option “Select all”, puis**
- **sélectionnez la séquence juste au-dessus de la séquence seuil.**
- **Ensuite, remontez en début de liste, enfoncez la touche Shift de votre clavier, et cliquez sur la première séquence de la liste en maintenant la touche Shift enfoncée. Cette action vous permet de sélectionner en un click 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.

Molecule type: amino acid
Query Length: 780
Other reports: Distance tree of results, Multiple alignment, MSA viewer, ?

to [] to []

Descriptions Graphic Summary Alignments Taxonomy

Sequences producing significant alignments

GenPept	Description	Scientific Name	Download	Select column
<input checked="" type="checkbox"/>	glycoside hydrolase family 70 protein [Azotobacter chroococcum]	Azotobacter chroococcum	FASTA (complete sequence)	FASTA (aligned sequences)
<input checked="" type="checkbox"/>	glycoside hydrolase family 70 protein [Azotobacter salinestris]	Azotobacter salinestris	GenBank (complete sequence)	Hit Table (text)
<input checked="" type="checkbox"/>	glycoside hydrolase family 70 protein [Pseudomonas cavernicola]	Pseudomonas cavernicola	Hit Table (CSV)	Text
<input checked="" type="checkbox"/>	glycoside hydrolase family 70 protein [Fraterula defendens]	Fraterula defendens	Descriptions Table (CSV)	E value
<input checked="" type="checkbox"/>	glycoside hydrolase family 70 protein [Paenibacillus vietnamensis]	Paenibacillus vietnamensis	XML	0.0
<input checked="" type="checkbox"/>	glycoside hydrolase family 70 protein [Paenibacillus caui]	Paenibacillus caui	ASN.1	0.0
<input checked="" type="checkbox"/>	glycoside hydrolase family 70 protein [Paenibacillus haemolyticus]	Paenibacillus haemolyticus		0.0
			631 631 86%	0.0
			625 625 98%	0.0

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

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. In the 'Input sequence' field, a protein sequence from Azotobacter chroococcum is pasted. The 'Sequence Type' is set to 'Protein'. Below the sequence, there is a text area for 'Paste your sequence here - or use the example sequence' with a 'Use the example' button. Under 'Parameters', the 'OUTPUT FORMAT' is set to 'ClustalW with character counts'. The 'ORDER' dropdown is circled in red and has 'input' selected. Other parameters like 'DEALIGN INPUT', 'MBED-LIKE CLUSTERING GUIDE-TREE', and 'MAX HMM ITERATIONS' are also visible.

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 onglets séparés (**click droit > ouvrir dans un nouvel onglet**).
- Observez les différences de ces deux formats dans “Tool Outputs”, notez les différences de visualisation entre eux.

The screenshot shows the EBI ClustalW tool interface. On the left, a modal window displays a green header "YOUR JOB IS FINISHED" and a message: "Please note that results can only be retrieved for jobs submitted within the last seven days." Below this, the job ID is listed as "Job ID: clustalo-I20241011-125516-0751-81092089-p1m". At the bottom of this modal are two buttons: "View Results" and "Your Jobs". A red arrow points from the "Your Jobs" button to the right side of the screen. On the right, there is a table titled "Job Title (ID) [ALL]" showing two entries:

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

Below the table, it says "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

The screenshot shows the EBI alignments interface with the following elements:

- Tool Output**: Shows 'Alignments' is selected.
- Alignments**: The main content area displays a sequence alignment of 19 glycoside hydrolase sequences from various species (e.g., WP_198318972.1, WP_167520052.1, etc.).
- Guide Tree**: A phylogenetic tree showing the evolutionary relationship between the sequences.
- Phylogenetic Tree**: Another phylogenetic tree view.
- Results Viewers**: Options for viewing results.
- Result Files**: Options for managing result files.
- Submission Details**: Options for submission details.

Annotations on the interface:

- Schéma de coloration**: Points to the 'COLOR SCHEME' dropdown menu set to 'clustal2'. A red circle highlights this area.
- Nightingale**: Points to the name of the color scheme.
- Zoom**: Points to the zoom controls (magnifying glass icons) located below the sequence labels.
- Legend**: A color key for the sequence alignment, showing categories: A (green), R (red), N (blue), D (orange), C (yellow), Q (light blue), E (purple), H (pink), I (light green), L (light orange), K (dark blue), M (dark green), F (dark orange), P (light purple), S (light green), T (light orange), W (dark green), Y (dark orange), V (light green), B (light orange), X (grey), Z (black).
- Fenêtre d'affichage (cliquer et faire glisser)**: Points to the horizontal scroll bar at the bottom of the sequence viewer, which is highlighted with a red oval.

Sequence labels (partial list):

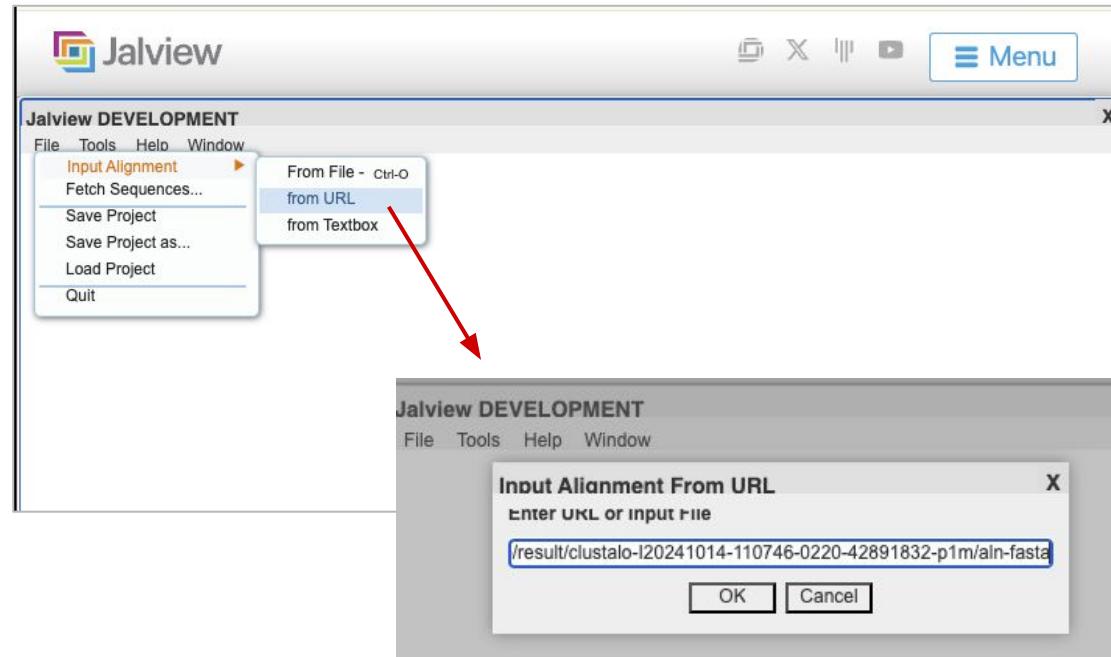
- WP_198318972.1 GLYCOSIDE HYDR
- WP_167520052.1 GLYCOSIDE HYDR
- WP_158592123.1 GLYCOSIDE HYDR
- WP_049623289.1 GLYCOSIDE HYDR
- WP_224722266.1 GLYCOSIDE HYDR
- WP_223067679.1 GLYCOSIDE HYDR
- WP_052702730.1 GLYCOSIDE HYDR
- WP_251460192.1 MULTISPECIES:
- WP_234969614.1 GLYCOSIDE HYDR
- WP_035322188.1 GLYCOSIDE HYDR
- WP_239984758.1 GLYCOSIDE HYDR
- WP_026830256.1 GLYCOSIDE HYDR
- WP_047390368.1 MULTISPECIES:
- WP_012371512.1 GLYCOSIDE HYDR
- WP_028105602.1 GLYCOSIDE HYDR
- WP_142092843.1 GLYCOSIDE HYDR
- WP_094364734.1 GLYCOSIDE HYDR
- WP_238594742.1 GLYCOSIDE HYDR

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.

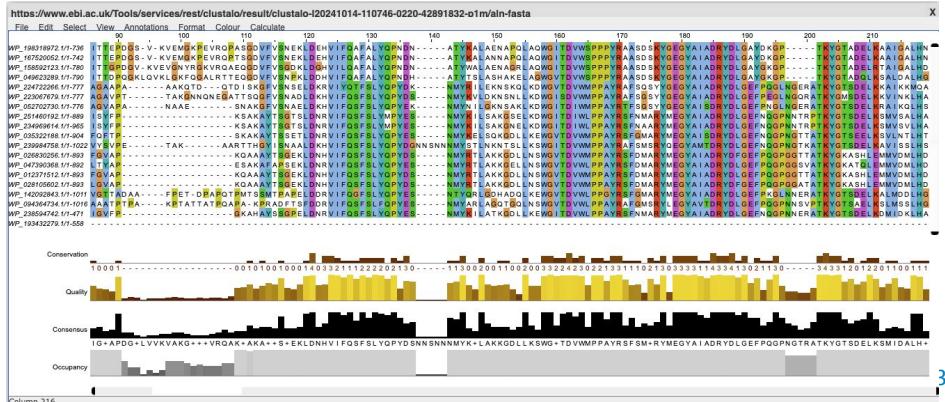
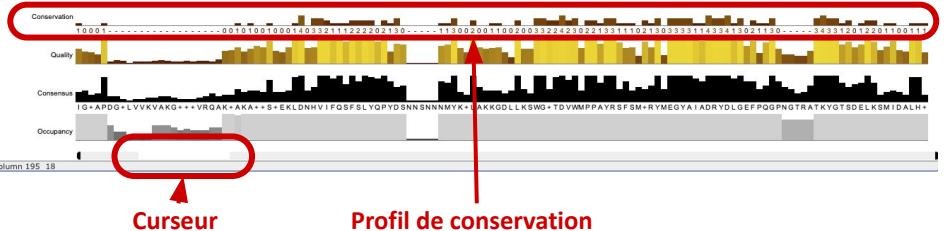
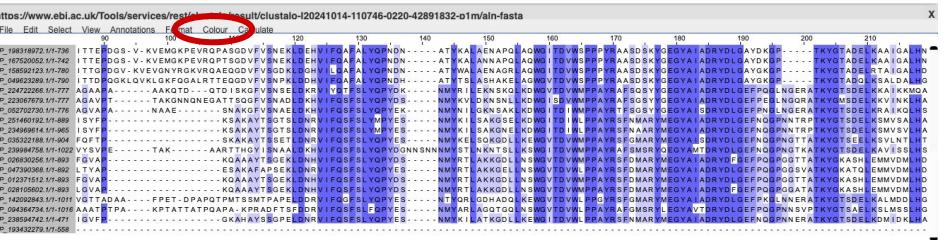
- A l'ouverture d'une page [JALVIEW](#), une fenêtre “**Jalview development**” dédiée à l'application apparaît devant celle du navigateur. Vous pouvez positionner cette fenêtre en haut à gauche puis la redimensionner vers le bas à droite pour qu'elle occupe bien votre écran.
- Dans cette fenêtre, cliquez 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**.



Jalview – Schémas de coloration des résidus

Une nouvelle fenêtre s'ouvre montrant l'alignement multiple.

- Déplacez le curseur blanc au bas de la fenêtre pour avancer vers le coeur de l'alignement (dépassez les gaps initiaux, généralement peu informatifs)
- Modifiez l'affichage de l'alignement en changeant la coloration (menu **Colour**) 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 déletion, 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 événement d'insertion ou de déletion.

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 procérons en deux temps

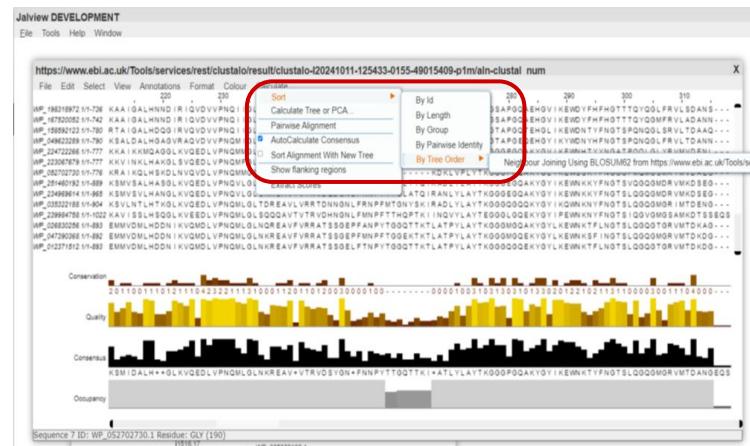
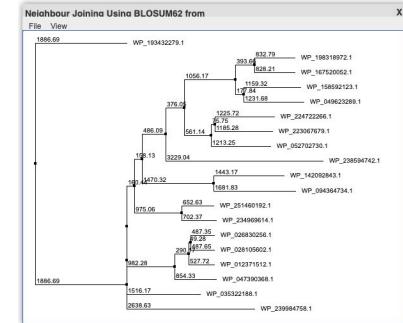
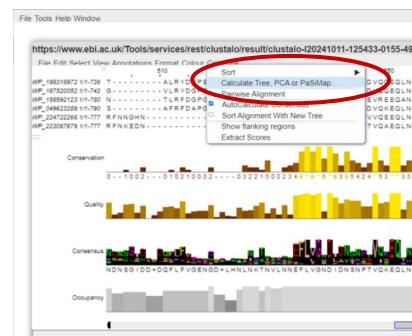
- Construction d'un arbre à partir de l'alignement multiple: menu **Calculate > Calculate Tree or PCA**, et laisser les paramètres par défaut.
- Ré-ordonnancement des séquences de l'alignement multiple en fonction de leur proximité dans l'arbre : **Calculate > Sort > By tree order**. Vous pouvez ensuite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

NE PAS OUBLIER DE CONTINUER LE TUTO POUR REPONDRE AUX QUESTIONS



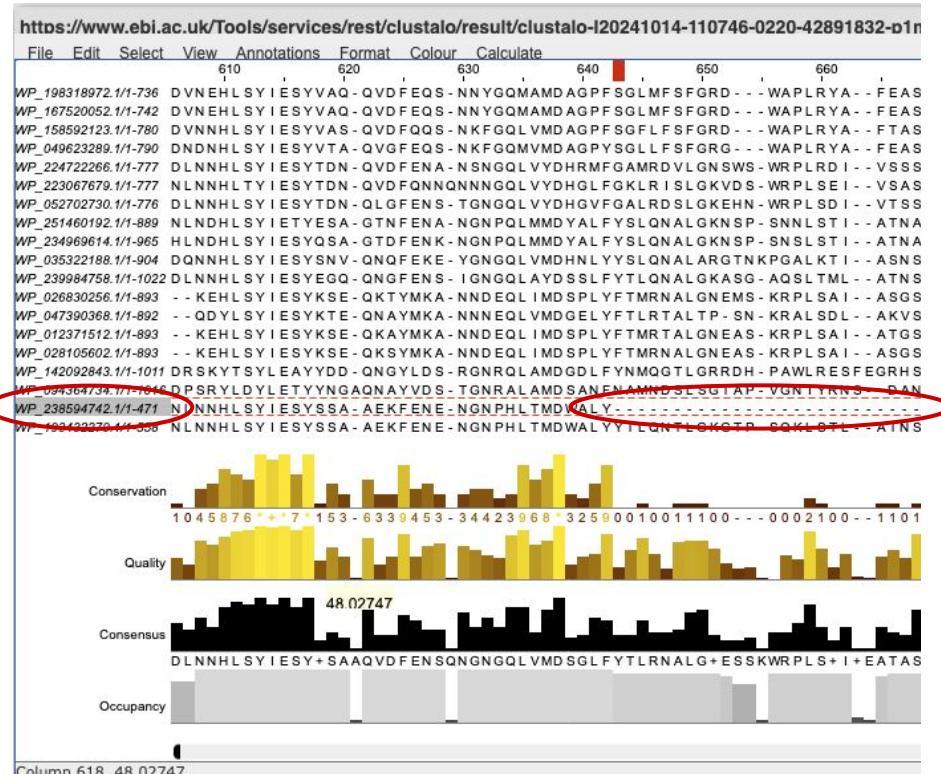
JalView - Edition du fichier d'alignement multiple

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.



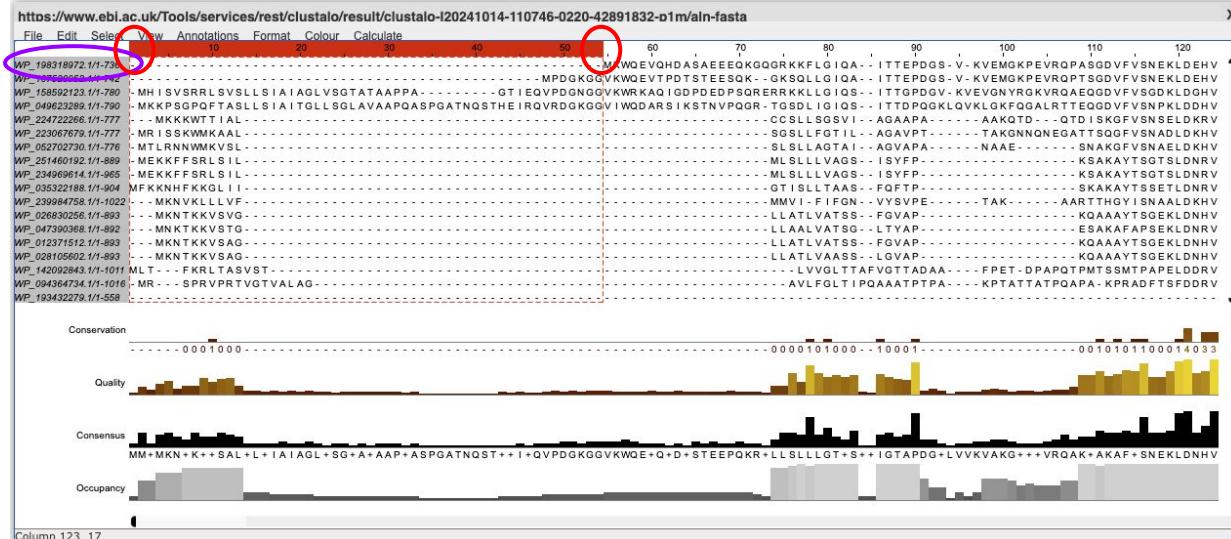
Suppression des séquences terminales excédentaires

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.

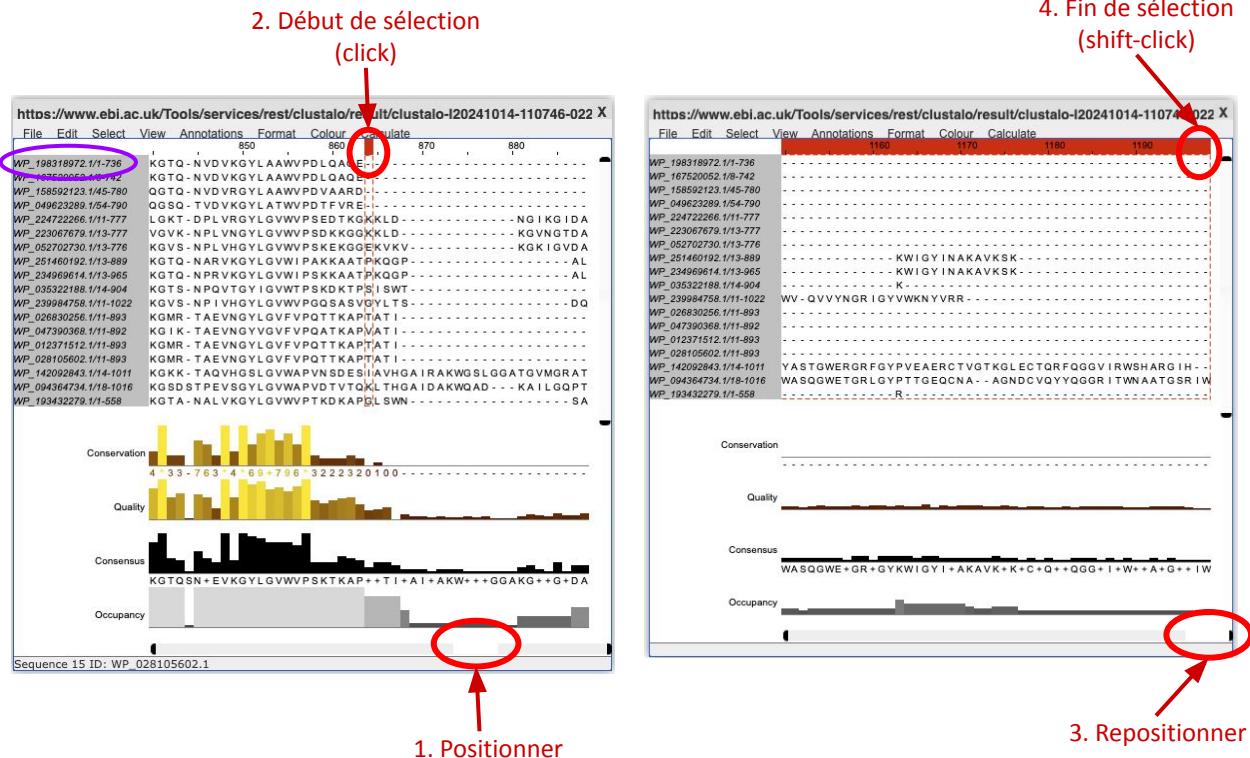


Suppression des séquences terminales excédentaires

- Fait 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 enfoncee

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



Analyse des séquences alignées avec les acides aminés catalytiques

Recherchez dans l'alignement, les acides aminés catalytiques D427, E469 et D528 de la séquence référence

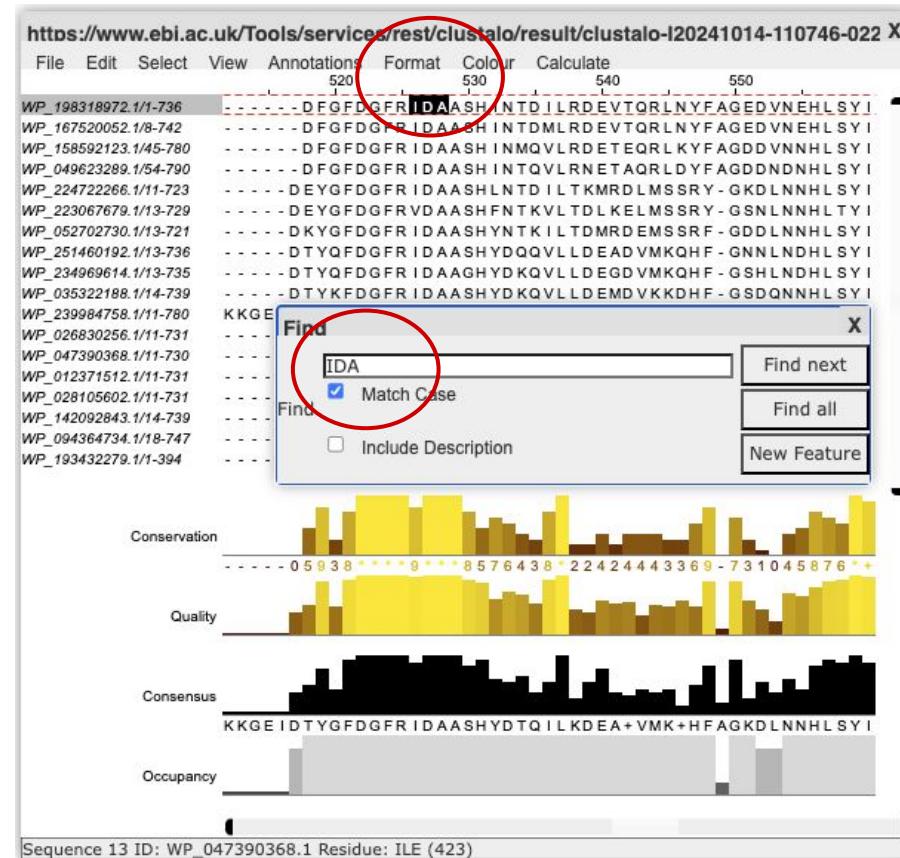
[WP_198318972.1](#). Pour cela, vous pouvez vous aider des acides aminés qui précèdent et qui suivent chacun de ces sites catalytique:

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

Attention: les positions des résidus dans une séquence donnée diffèrent des positions des colonnes de l'alignement, du fait de la présence de gaps. Il ne faut donc pas confondre l'acide aspartique à la position 427 de la séquence référence et l'acide aminé à la position 427 de l'alignement.

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. Par exemple, pour trouver le résidu D427, cherchez la chaîne de caractères "IDA".

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



Debriefing

- Acides aminés du site catalytique (figure du haut)
 - Les acides aminés qui forment le site catalytique ((D427, E469 et D528) sont fortement conservés (identiques dans toutes les protéines analysées ici)
 - Ils occupent des positions éloignées sur la séquence (427, 469 et 528), mais rapprochées dans la structure 3D.
- Permutation circulaire (figure du bas)
 - **A FAIRE**

C'est fini !