TP1 Bioanalyse L3 BCP, Année 2022_2023, durée 3H30
silico.biotoul.fr ou
Moodle / L3 BCP/ Bioanalyse ELSV6C1
(etudiantfsi mdptmp7 !)

OBJECTIFS

- Etre capable de retrouver une séquence dont on connait le numéro d'accession dans sa banque
- Savoir comment s'organisent les fiches des séquences, et où y chercher les informations
- Etre capable de trouver une ou des séquences à l'aide de mots clés ciblant des champs spécifiques
- Naviguer entre les banques, changer de format, télécharger des séquences
- Utiliser quelques outils de bioanalyse pour répondre à une question biologique

EXERCICE 1 : Recherche d'une séquence dans les banques via son numéro d'accession

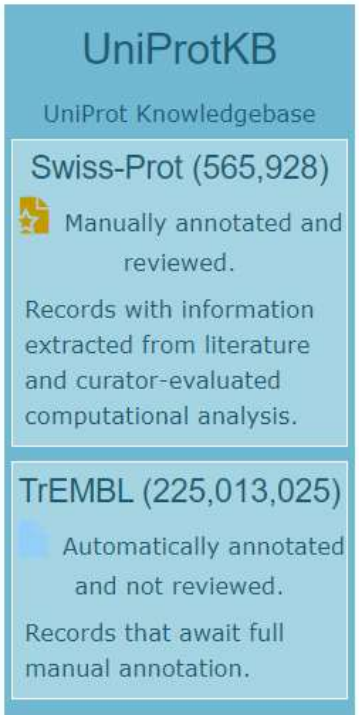
1/ Allez sur le site d' **UniProt :**

*Où êtes-vous localisés ? Qu'est-ce que Uniprot ?

Sur le site EBI-EMBL, centre de ressources en bioinformatique Européeen

Expliquer en effet que les données biologiques sont stockées en miroir dans 3 centres de ressources (EBI/EMBL pour l'Europe ; NCBI pour les USA et DBJ au Japon) en TP on restera sur EBI et NCBI
Dans ces centres de ressources il y a des 'banques de données' que l'on peut 'interroger' pour trouver de l'information (ie banques protéiques, nucléiques, structures de proteines…)

Ici Banque de données de protéines scindée en 2 sections TrEMBL et SwissProt
Expliquez la différence entre les 2 sections : TrEMBL ensemble des séquences protéiques provenant de la traduction automatique des CDS de EMBL/ SwissProt séquences de TrEMBL validées manuellement par un annotateur en se basant sur des données expérimentales (exemple : caractérisation mRNA, purif de la protéine correspondante…) donc beaucoup moins de sqces dans SwissProt



- Combien de séquences sont référencées dans la section SwissProt de UniProt ? dans la section TrEMBL de UniProt?

Sur la page de garde EBI EMBL, colonne de gauche, vous avez le nombre de séquences, cf résultats au 13/01/21 ci-contre

Chercher la séquence P01308 dans Uniprot

Indiquez que P01308 est un numéro d'accession, et qu'il est unique à une séquence, permet d'identifier les séquences dans les banques (comme un numero de référencement d'un livre à la BU, il est unique).

L'idée est ensuite qu'ils apprennent à lire les fiches des séquences et comprennent que l'information n'est pas en vrac mais structurée dans des 'Champs' (notés en couleur moutarde, ie Family Domains, Sequences et repris sur la colonne de gauche)

Ce classement en Champ va ensuite faciliter les requetes pour identifier l'information, notamment par utilisation de mots clés

A chaque question demander dans quel champ de la fiche ils ont trouvé l'info

- De quelle protéine s'agit-il ? chez quel organisme ? Insuline Humaine
- Quelle est la taille de cette séquence ? 2 isoformes (=issues épissage meme gene / eucaryote) : 110 AA et 200AA (Champ Sequences)
- Que sont les "VARIANT" ? zone où # AA possibles selon des familles associées à des pathologies (Champ Pathology Biotech) = polymorphisme avec des liens vers dbSNP
- Y a-t-il des preuves expérimentales de l'existence de cette protéine ? En haut de la page : Status Reviewed - - Experimental evidence at protein level
- Est-elle dans UniProt-trEMBL ou UniProt-SwissProt ? de fait elle devrait se trouver dans SwissProt, cela se confirme dans le Champ Status ou Entry Information
- Dans quels processus intervient cette protéine ? Cliquer sur des liens de la **GO** et regarder **Ancestor Chart**

Voir la fonction au début de la fiche, et la partie Ontology du Champ Function

Expliquer Gene Ontology =vocabulaire structuré pour décrire les produits des gènes de façon homogène quelque soit l'organisme. But de la GO= donner le même nom et recenser les synonymes.

3 structures dans GO : processus cellulaires, fonctions moléculaires et composants/compartiments cellulaires.

Cliquer sur un lien GO : regarder la page, description synonymes…

Et Ancestor Chart = montre la structure hiérarchique. Depuis processus, et de + en + précis vers le processus exact.

Intérêt : vue synthétique

MAIS surtout quand on travaille sur bcp de gènes : programmes qui extraient des processus communs ou sur-représentés

Afficher le format **UniProtKB** en cliquant sur Format => Onglet 'Text' en haut de la page

Expliquer ce format en disant que les Champ identifies en couleur moutarde initialement' sont sous forme de code a 2 lettres

Leur demander par exemple dans quel code on retrouve les infos

sur la fonction, maladies, etc CC (et les CC sont les commentaires mis par les experts de SwissProt)

la Gene Ontology DR

Faire afficher le format UniProtKB en cliquant sur Format => Text en haut de la page

2/ Sur le site du **NCBI** : chercher (*via* **ENTREZ**) la même séquence.

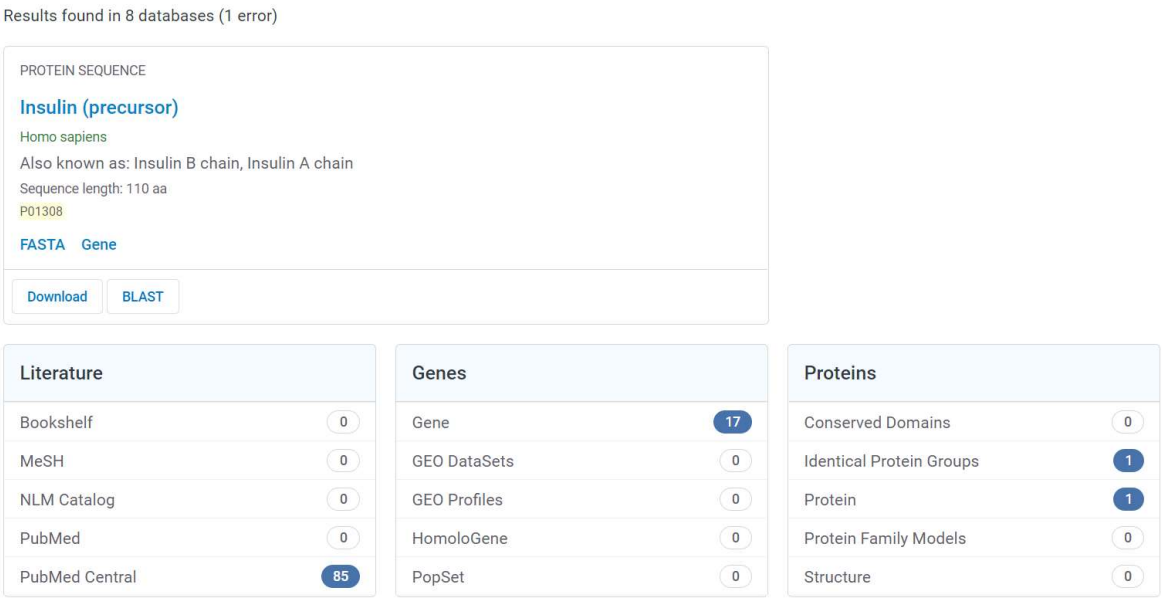
- Quels sont les résultats ?
- Cliquer sur Protein : la séquence est ici au format **GenPept**

Page intermédiaire de résultats listant toutes les banques dispo au NCBI classées par thème, litterature, gene, protein…. Les numéros indiquent le nombre d'occurrence par rapport à la requête (ie, 68 articles scientifiques stockés dans PubMed et concernant ce numero d'accession)

Indiquez que quand ils cliquent sur 'Protein' que cela les envoie dans la banque de données 'Protein' du NCBI (cf bandeau deroulant de la barre de requete 'protein') que chaque ligne en milieu de page correspond a une sequence et quand ils cliquent dessus cela ouvre la fiche descriptive correspondante au format GenPept

Expliquez le format GenpPept vs Uniprot, les champ sont des 'Mots Complets' sur Colonne de Gauche (Locus, Definition, Accession, Features…),

Faites remaqruer que la base de données initiale pour cette séquence est bien Uniprot/SwissProt (indiqué sous le titre Insulin)



Faire afficher le format 'FAsta' de la séquence. Qu'est ce que le format FASTA ?

Expliquez le format FAsta, important pour de nombreux logiciels de Bionalyse pour se 'repérer dans la séquence ' en effet le fasta commenece toujours par une ligne de titre identifiée entre le signe > et un retour à la ligne, généralement on y indique le numero d'accession

Après le retour à la ligne, la séquence (AA, bp) sans 'blanc' non numérotée souvent des lignes de 80 caractères puis un retour à la ligne (non visible)

- Revenez aux résultats et cliquer sur Gene : regarder l'entrée INS, en particulier la partie **NCBI Reference Sequences** : combien de variants d'épissage ? et dans **Related Sequences** : combien d'ARNm ?

Gene= banque propre au NCBI qui regroupe ttes les informations pour un gène donné.

Associée à Refseq

parler de RefSeq : numéro AC reconnaissable 2 lettres _ chiffres

RefSeq= valeur ajoutée du NCBI pour pallier à la redondance. 1 représentant ARNm et 1 prot

(ici on en a 4 car il y a des variants d'épissage)

*Dans **Related Sequences**, beaucoup d'ARNm (on les voyait aussi dans UniProt)*

Pb de redondance des banques (vu en cours) : séquençage pas fragment => on n'obtient pas tjs la séquence complète, donc x équipes séquencent le même gène mais obtiennent x séq #
+ le polymorphisme (variants) + épissage alternatif…

⇨ *Intérêt de RefSeq (avec info **Source sequences**)*

⇨

- *Regarder la séquence génomique RefSeq NG_007114 (format Genbank) : combien d'exons composent ce gène ? combien interrompent la séquence codante ?*

mRNA join(1..42,222..425,1213..1431) 3 exons

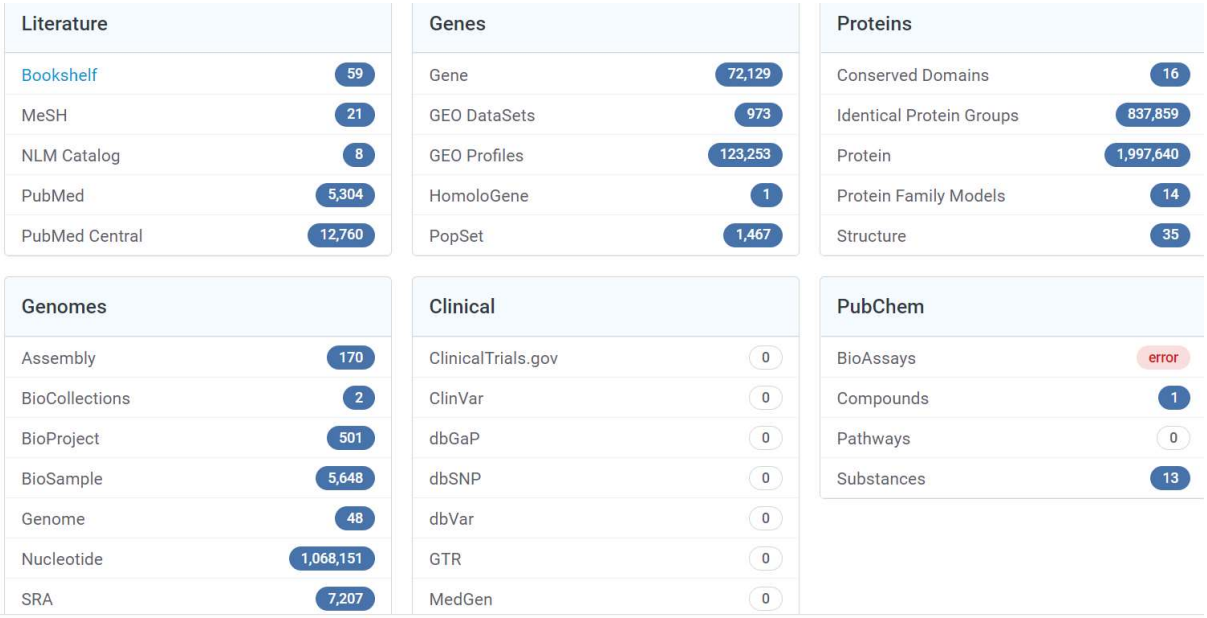
CDS join(239..425,1213..1358) 2 exons

EXERCICE 2 : Recherche dans les banques via l'utilisation de mots clés

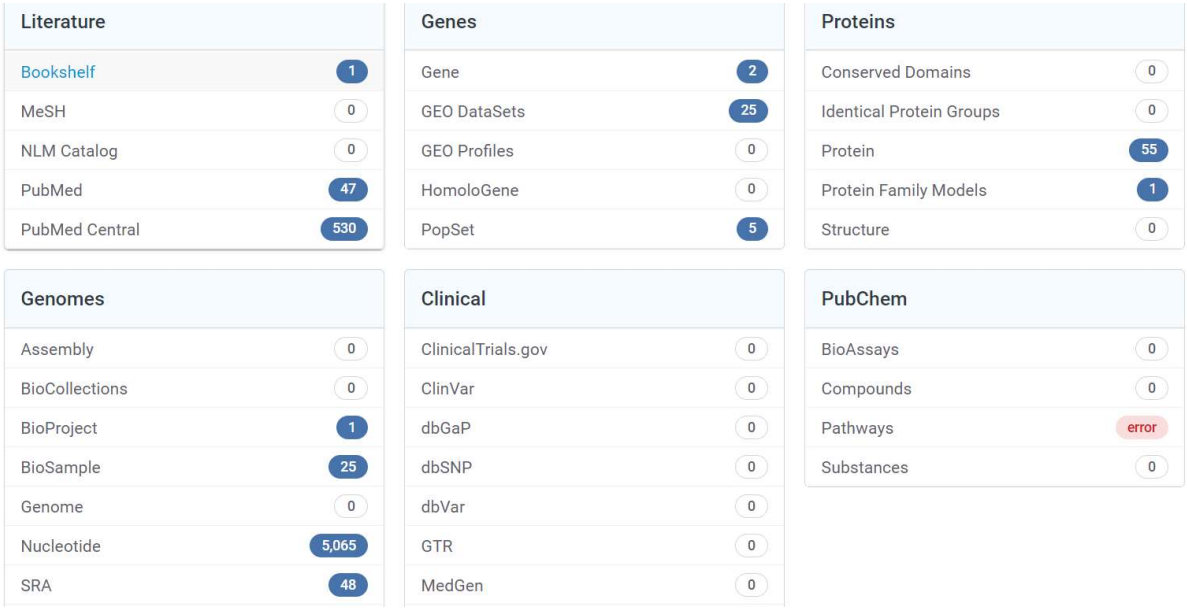
1/ Sur le site du NCBI, identifiez :

- toutes les séquences de l'oomycète Phytophthora (parasite de la pomme de terre) : combien sont-elles ? *Phytophthora*

  !! attention aux fautes d'orthographe, cela fait varier le nombre de résultats considérablement



- *Phytopthora results (manque le 2me h)*



- les séquences protéiques de Phytophthora parasitica correspondant à des éliciteurs (molécules capables d'induire les réponses immunitaires chez les végétaux)

Pour cela utiliser ENTREZ, et si vous ne voulez rechercher que dans la banque protéique, cliquer sur Protein, puis utiliser l'option Advanced.

A l'aide de l'outil Search builder préciser les champs (Organism, Title...) et conjuguer vos requêtes.

Par défaut lorsque plusieurs mots clés sont utilisés c'est l'opérateur AND qui s'applique entre les mots.

L'historique de vos requêtes est disponible en dessous et vous pouvez combinez des résultats de requêtes précédentes avec les mêmes opérateurs AND, OR et NOT.

"NB : L'utilisation de * permet de chercher une famille de mots. Par exemple, avec elicit*, vous pourrez trouver elicitor, elicitate, elicitin..."*L'idée ici est qu'ils arrivent à faire des recherches avancees avec des mots clés + des opérateurs (AND, OR) et le tout in english. Imporatnce des bons clés, exmple 'protéine' n'est pas un mot clé si on cherche dans une banque de protéine !*

NB : Janvier 2022: je n'ai pas mis a jour les chiffres des requetes

Phytophthora parasitica 274861

Phytophthora parasitica [Organism] 274819 (il y avait donc quelques faux positifs, le nom de phytophthora parasitica apparaissait ailleurs que dans le champs Organism

Phytophthora parasitica [Organism] AND elicit* 675

Cherche elicit partout sur une fiche de séquence
Leur faire ouvrir des fiches et chercher les mots
il faut préciser là où on veut chercher le mot. Tester des propositions de chmaps dans Advanced

NB : [protein name] => cherche dans les Features : Protein /product= « elicitin »

[Title] => ne cherche pas dans le titre de la publi mais dans le titre de la page càd Definition

Search	Add to builder	Query	Items found	Time
#9	Add	Search **((((Phytophthora parasitica[Organism] AND elicit*[protein name])) OR (Phytophthora parasitica[Organism] AND elicit*[Title]))) AND CAA65843[Accession]**	0	04:14:18
#8	Add	Search **((Phytophthora parasitica[Organism] AND elicit*[protein name])) OR (Phytophthora parasitica[Organism] AND elicit*[Title])**	23	04:11:57
#7	Add	Search **Phytophthora parasitica[Organism] AND elicit*[Title]**	22	04:11:08
#6	Add	Search **Phytophthora parasitica[Organism] AND elicit*[protein name]**	15	04:10:55
#5	Add	Search **Phytophthora parasitica[Organism] AND elicit***	675	04:06:44
#4	Add	Search **Phytophthora parasitica[Organism]**	274819	04:06:09
#3	Add	Search **Phytophthora parasitica**	274861	04:05:34
#2	Add	Search **Phytophthora**	525357	04:05:24
#1	Add	Search **Phytophthora paraistica**	525357	04:05:13

2/ On s'intéresse maintenant à la séquence dont le numéro d'accession est CAA65843

Regardez la fiche de la séquence correspondante :

- comment s'organise cette fiche ? Format GenPept : identifiant pour chaque partie. Description générale, puis les Features
- quel est le nom de cette protéine ? cellulose binding elicitor lectin (CBEL), formerly GP34, molecule capable d'induire des réponses de défense chez les végétaux. Peut se lier à la cellulose in vitro
- dans quel journal scientifique les travaux concernant cette protéine ont-ils été publiés ?Mol. Plant Microbe Interact
- sous quel numéro cette publication est-elle référencée dans PubMed ? 9390419
- de combien d'acides aminés est composée cette protéine ? 268 AA
  quels domaines sont présents dans la protéine ? Champ 'Feature' a lire de 'Haut en Bas' pour identifier tous les domaines et leurs positions sur la séquence : au final 2 PAN/APPLE et 2 fCBD (fungal Cellulose-Binding) + 1 peptide signal (1..18) mais ce n'est pas un domaine protéique juste une region d'AA hydrophobe + signal de clivage protéique en Nter des protéines

  Matérialisez l'organisation strcuturale de la proteine au tableau de la position 1 à 268 , demandez quel pourrait etre la fonction pour cet eucaryote filamenteux pathogène de plante au vu des domaines et probalement sa localisation extracellulaire → adhésion à de la cellulose, donc les feuilles de patate pour ensuite facilter l'infection de la plante ?

```
/db_xref="GOA:O42830"
/db_xref="InterPro:IPR000177"
/db_xref="InterPro:IPR000254"
/db_xref="InterPro:IPR003014"
/db_xref="InterPro:IPR003609"
/db_xref="UniProtKB/TrEMBL:O42830"
```

3/ On s'intéresse maintenant aux références croisées, notées "db_xref" sur la fiche

- à quoi correspondent ces différentes références croisées ?

liens vers banques spécialisées de domaines CDD = Conserved Domain Database = banque de domaines du NCBI, l'équivalent de InterPro (récupère les info de SMART, Pfam…) ou InterPro, lien vers GOA (GO Annotatipn) et vers UniProt

- quels domaines sont présents dans la protéine ?

2 PAN/APPLE et 2 fCBD (fungal Cellulose-Binding)

- quelle est la fonction du domaine "IPR000254" ? est-il spécifique des oomycètes/champignons ou de toute autre espèce ?

IPR000254 Cellulose-binding domain, fungal

The microbial degradation of cellulose and xylans requires several types of enzymes such as endoglucanases (EC:3.2.1.4), cellobiohydrolases (EC:3.2.1.91) (exoglucanases), or xylanases (EC:3.2.1.8) [PMID: 1886523]. Structurally, cellulases and xylanases generally consist of a catalytic domain joined to a cellulose-binding domain (CBD) by a short linker sequence rich in proline and/or hydroxy-amino acids. The CBD of a number of fungal cellulases has been shown to consist of 36 amino acid residues, and it is found either at the N-terminal or at the C-terminal extremity of the enzymes. As it is shown in the following schematic representation, there are four conserved cysteines in this type of CBD domain, all involved in disulphide bonds.

- 90% fungi mais aussi des algues et des oomycètes (organismes filamenteux ressemblant morphologiquement aux champignons mais phylogénétiquement plus proches des diatomées et algues brunes) cf Species, Evolution dans SMART et species dans Pfam

- ce domaine est-il référencé dans d'autres banques de domaines ? Si oui, lesquelles et avec quel numéro d'accession ?

InterPro rassemble les liens vers des banques de domaines/motifs donc on a des liens vers les contributing signatures :

SMART

- *SM00236 (fCBD)*

PROSITE profiles

- *PS51164 (CBM1_2)*

ProDom

- *PD001821 (CBD_fun)*

Pfam

- *PF00734 (CBM_1)*

SUPERFAMILY

- *SSF57180 (SSF57180)*

PROSITE patterns

- *PS00562 (CBM1_1)*

- aller sur le lien db_xref vers UniProt O42830 :

- à quelle section de UniProt appartient cette séquence ?

Entry status[i] Unreviewed (UniProtKB/TrEMBL)

  - que constatez-vous par rapport à la fiche GenPept ? + d'info, + de liens vers bcp de banques spécialisées
  - quelles sont les fonctions moléculaires correspondant à la Gene Ontology ?

- cellulose binding Source: InterPro
- hydrolase activity, hydrolyzing O-glycosyl compounds Source: InterPro
  - Donner les numéros des termes GO associés. GO:0030248 et GO:0004553
    en cliquant sur les liens GO ou en passant en format Text
    L'annotation GO est-elle présente dans la fiche GenPept ? non, lien vers GOA O42830
    /db_xref="GOA:O42830"

- que constatez-vous par rapport à la fiche GenPept ?

+ d'info, + de liens vers bcp de banques spécialisées

- quelles sont les fonctions moléculaires correspondant à la Gene Ontology ? Donner les numéros des termes GO associés. L'annotation GO est-elle dans GenPept ?

- cellulose binding Source: InterPro
- hydrolase activity, hydrolyzing O-glycosyl compounds Source: InterPro
  - Donner les numéros des termes GO associés. GO:0030248 et GO:0004553
    en cliquant sur les liens GO ou en passant en format Text
    L'annotation GO est-elle présente dans la fiche GenPept ? non, lien vers GOA O42830
    /db_xref="GOA:O42830"

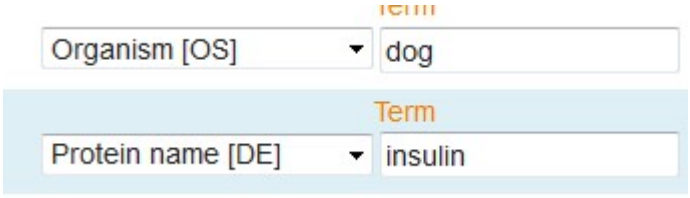
* sans cliquer sur le bouton 'retour en arrière de votre navigateur', depuis Uniprot (EBI-EMBL), trouver un moyen pour revenir à la fiche initiale au NCBI

Lien croise

EXERCICE 3 : Quelques exemples de recherches avancées sur Uniprot

1/ Chercher les séquences protéiques d'insuline chez le chien :

- combien sont référencées dans UniProt/SwissProt ?
- combien dans trEMBL ?
- afficher les séquences au format FASTA



4 dans SwissProt

49 dans trEMBL (mais aussi des dog tik et raccoon dog !)
sauf si on sélectionne dog[9615] ou si on met le nom taxonomique Canis lupus familaris
=> seulement 46

P33712	IGF1_CANLF		**Insulin-like growth factor I**	**IGF1** IGFIA	Canis lupus familiaris (Dog) (Canis familiaris)	153
P01321	INS_CANLF		**Insulin**	**INS**	Canis lupus familiaris (Dog) (Canis familiaris)	110
Q6X7V3	INSL3_CANLF		**Insulin-like 3**	**INSL3** RLF	Canis lupus familiaris (Dog) (Canis familiaris)	132
Q9XST2	GTR4_CANLF		**Solute carrier family 2, facilitate...**	**SLC2A4** GLUT4, B146	Canis lupus familiaris (Dog) (Canis familiaris)	162

Sur UniProt : ajoute * par defaut => trouve des insulin-like ! mais on ne peut pas enlever cette option !
Trouve aussi **Solute carrier family 2**
ouvrir la fiche : dans Alternative name : insulin-responsive (on le voit facilement au format text)
Avec NOT Protein name : insulin-responsive il ne sort pas

- afficher les séquences au format FASTA

Sur l'onglet du haut, faire un petit rappel du FASTA car ici plusieurs séquences à la suite (Sur UniProt : => Download)

2/ Toujours sur UniProt, chercher des séquences protéiques de dinosaures

Les laisser chercher, ils vont mettre dinosaur* comme organism et dire qu'il n'y en a pas...
Taxonomy : dinosaur* (ou dinosauria)
Plein de chicken !
Ouvrir 1 fiche=> chicken appartient au dinosaure. Tous les oiseaux (aves) en fait
Taxonomy : dinosaur*
NOT Taxonomy : aves

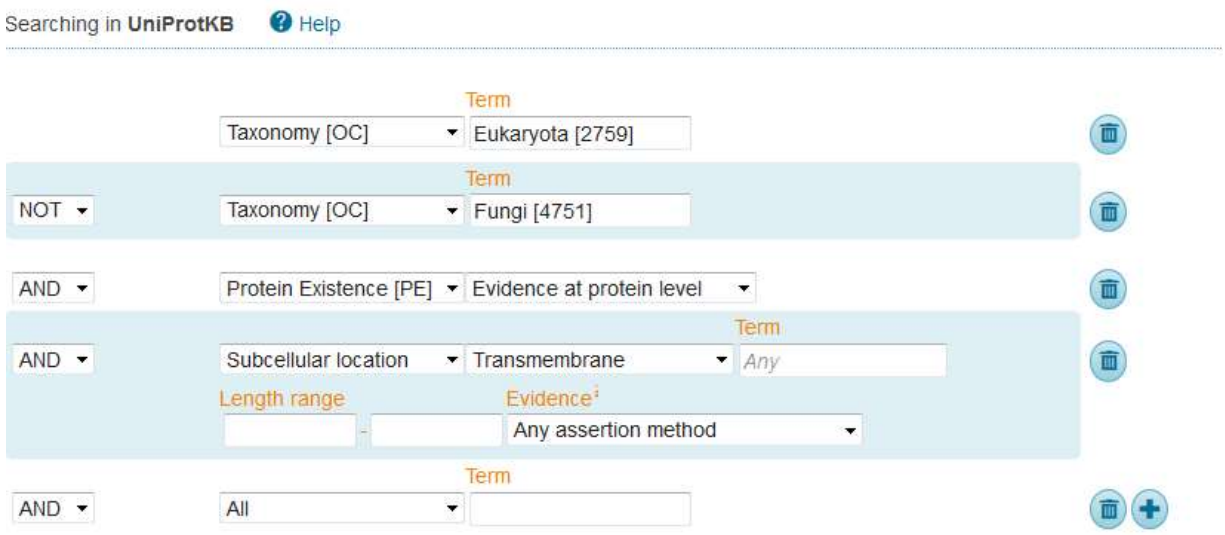
4 séquences dans UniProt

P0C2W2	CO1A1_TYREX		**Collagen alpha-1(I) chain**	**COL1A1**	Tyrannosaurus rex (Tyrant lizard king)
P86289	CO1A1_BRACN		**Collagen alpha-1(I) chain**	**COL1A1**	Brachylophosaurus canadensis (Campanian hadrosaur)
P0C2W4	CO1A2_TYREX		**Collagen alpha-2(I) chain**	**COL1A2**	Tyrannosaurus rex (Tyrant lizard king)
P86290	CO1A2_BRACN		**Collagen alpha-2(I) chain**	**COL1A2**	Brachylophosaurus canadensis (Campanian hadrosaur)

3/ Trouver le nombre de séquences de trEMBL avec des preuves au niveau protéique, chez des organismes eucaryotes n'appartenant pas aux champignons.

- combien sont transmembranaires ?

Janvier 22 : je n'ai pas mis à jour les chiffres !



Reviewed (10,868)
Swiss-Prot

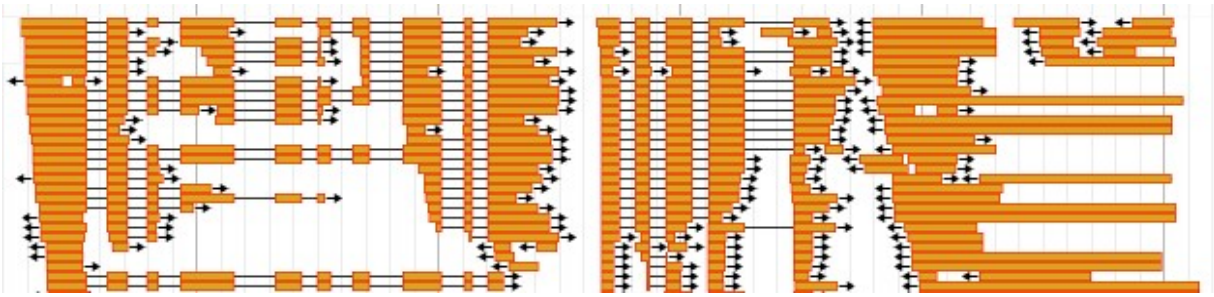
Unreviewed (15,873)
TrEMBL

EXERCICE 4 : Recherche dans des banques spécialisées

1/ Allez sur le site ARAPORT

Portail d'information sur Arabidopsis

- Allez sur JBROWSE : utilisez le zoom pour afficher des gènes sur une région. Demandez à voir les alignements avec EST/Full length cDNAs
  JBROWSE = navigateur sur un génome, permet de voir les annotations par position à gauche, dans Community data, EST and profile alignments, cocher EST/full length cDNA
  On voit ce genre de chose :



- Allez dans THALEMINE : dans l'onglet Regions, chercher les CDS du chromosome 4 entre la position 20 000 et 60 000
  THALEMINE = système de requête par mots clés, positions…
  cochez CDS. Faire afficher exemple pour connaitre le format
  Chr4:20000..60000

FEATURE	FEATURE TYPE	LOCATION
AT4G00060.1	CDS	Chr4:21437..27996
AT4G00060.2	CDS	Chr4:21437..25248
AT4G00060.3	CDS	Chr4:21437..27996
AT4G00060.4	CDS	Chr4:21437..27996
AT4G00070.1	CDS	Chr4:29672..31426
AT4G00070.2	CDS	Chr4:29774..31426
AT4G00080.1	CDS	Chr4:32946..33575
AT4G00090.1	CDS	Chr4:34234..36594
AT4G00100.1	CDS	Chr4:37172..38123
AT4G00110.1	CDS	Chr4:38702..39994
AT4G00120.1	CDS	Chr4:42601..43197
AT4G00130.1	CDS	Chr4:48414..49068
AT4G00140.1	CDS	Chr4:51166..53449
AT4G00150.1	CDS	Chr4:57429..59105

2/ Recherches dans les banques de domaines :

- interrogez PFAM par mots-clés pour chercher les domaines cytochrome b5

  Keyword search : cytochrome b5 => lien PF00173 Cyt-b5

- interrogez InterPro avec InterProScan à l'EBI (Onglet Services => InterProScan) pour chercher s'il y a des domaines connus sur la séquence P00174 (que vous devez d'abord récupérer au format FASTA sur UniProt ou au NCBI).

En cas de problème, lien vers ancienne version InterProScan : http://www.ebi.ac.uk/interpro/legacy