Interrogation de banques via Gquery (Entrez)

Vous allez utiliser GQuery (anciennement nommée Entrez) du NCBI, qui est l'interface d'interrogation rapide de l'ensemble des banques de données développée par le NCBI. Vous allez explorer la manière d'interroger efficacement une banque de données de séquences en progressant dans la difficulté.
La copie d'écran suivante indique les fonctionnalités que nous allons utiliser. Il est conseillé de la garder ouverte afin de pouvoir s'y référer.

Lecture d'une entrée de séquence nucléique

Vous arrivez d'abord sur la page générale permettant d'interroger toutes les banques en une seule fois. Lorsqu'une requête est lancée, il apparaît à côté de chaque banque le nombre d'entrées correspondant à la requête dans la banque. Nous allons rechercher l'entrée de la banque "Nucleotide" qui porte le numéro d'accession `Y10810`. Pour cela, vous devez saisir les mots à rechercher (ici le numéro d'accession) dans la zone de saisie des requêtes.

> Question 1
>
> Combien d'entrées trouve-t-on dans la banque Nucleotide ?
> Y a-t-il d'autres banques avec une correspondance ?

Consultez l'entrée de la banque Nucleotide pour répondre aux questions suivantes (vous devez cliquer sur le nom de la banque, puis sur le numéro de l'entrée) :

> Question 2
>
> Est-ce une séquence ADN ou ARN ?
> De quel organisme provient la séquence de l'entrée ?
> Est-ce un eucaryote ou un procaryote ?
> Quelle est la taille de la séquence contenue dans l'entrée ?
> Combien de séquences codantes (CDS) compte cette entrée ?
> Cela vous semble-t-il normal étant donné la nature moléculaire de la séquence (ADN/ARN) et son origine (eucaryote/procaryote) ?

La première séquence codante est annotée comme étant une uORF. C'est un élément qui intervient dans la régulation de la traduction des ARNm. C'est pourquoi il y a 2 CDS sur un ARNm eucaryote.
Pour avoir une description plus précise de ce qu'est une uORF, vous pouvez faire une recherche dans la banque Books qui contient un ensemble de livres dont le texte est disponible sur le web. Il suffit de saisir `u0RF` dans la zone de saisie, après avoir choisi la banque *Books*.

> Question 3
>
> Que signifie le 'u' de uORF ?
> Comment fonctionne une uORF ?

Retournons maintenant à l'entrée de notre ARNm dans la banque Nucleotide pour déterminer le peptide codé par l'uORF.

> Question 4
>
> Quelle est la position de l'uORf sur l'ARNm ?
> Consultez la séquence de l'entrée pour déterminer la séquence de l'uORF.
> A l'aide du code génétique traduisez la séquence de l'uORF.
> Vérifiez votre traduction à l'aide de la séquence protéique donnée dans l'entrée (ligne *translation*).

Changement du format d'affichage et manipulation de fichiers

En haut à gauche de la page figurent différentes options pour changer le format d'affichage. On y trouve *FASTA*, *Graphics* et le menu *Display Settings*. Le format par défaut d'une entrée est *GenBank*. Un autre menu, *Send*, qui se trouve à droite de la page, permet d'enregistrer les données soit dans un fichier (*File*), soit dans une mémoire temporaire sur le site du NCBI (*Clipboard*), soit de façon permanente un compte MyNCBI que vous devez créer.

Changez le format d'affichage de l'entrée en *Graphics*. Localisez la représentation du CDS `CPRF4b` (vert), de la protéine associée (rouge), ainsi que leurs informations respectives.

> Question 5

Quelle est la longueur du CDS et de la protéine produite ?
Combien de motif protéiques (parmi les éléments en noir) sont décrits sur cette protéine ? à quoi correspondent-ils ?
Y a-t-il d'autres informations visibles ? Lesquelles ?

Repassez du format d'affichage *Graphics* au format *GenBank*.
Enregistrez ensuite l'entrée GenBank complète à l'aide du menu *Send* → *Complete Record* + *File* sous le format *GenBank (full)*, ainsi que sous le format *Fasta* (deux fichiers).

Question 6

Que se passe t-il si vous essayez d'ouvrir les fichiers précédemment enregistrés au format *Fasta* ou *GenBank* par double clic ?

Pour lire ces fichiers, vous devez ouvrir le Bloc-notes (ou tout autre éditeur de texte), puis utiliser le menu *Fichier* → *Ouvrir* de cet éditeur pour aller chercher le fichier. Il est également possible :

de changer le nom du fichier (par exemple en ajoutant l'extension ".txt" aux fichiers "sequence.gb" et "sequence.fasta"),

sinon (si vous n'avez par exemple pas de droits d'accès suffisants), d'enregistrer le fichier précédemment ouvert par l'éditeur de texte en y ajoutant une extension ".txt" qui facilitera sa réouverture sous Windows.

Notez qu'il est toujours possible de procéder par copier-coller dans un fichier, en utilisant les formats proposés dans le menu *Display Settings*.

Affichez la séquence au format *GenBank* à l'aide du menu *Display Settings* : vous noterez qu'il est possible de Copier-Coller cette séquence si l'on fait attention à ne sélectionner que la partie associée au fichier GenBank.

Affichez ensuite la séquence au format *FASTA* puis au format *FASTA (text)*.

Question 7

Lequel de ces deux affichages est le plus aisé pour procéder à un copier-coller ?

Interrogation ciblée

Lancez la requête `Bacillus subtilis` sur la banque *Nucleotide*.

Question 8

Est-ce que les séquences trouvées proviennent toutes de cette bactérie ?
Pour les entrées qui ne proviennent pas de *B. subtilis*, où le nom de cette bactérie apparaît-il (ne consulter que quelques entrées) ?

Par défaut, les termes d'une requête sont recherchés dans l'ensemble de l'entrée. Pour faire des recherches plus pertinentes, il faut préciser le champ dans lequel les termes sont recherchés ...

Lancez la requête `Bacillus subtilis [organism]` sur la banque *Nucleotide*.

Question 9

Est-ce que les séquences trouvées proviennent désormais toutes de cette bactérie ?
Devinez et interprétez alors le sens de la requête complémentaire `Bacillus subtilis NOT Bacillus subtilis [organism]` et valider son résultat avec le nombre d'entrées trouvées par les deux précédentes requêtes.

Pour les banques du NCBI, l'interroger d'un *champ spécifique* se fait à l'aide du *nom du champ entre crochets* après le ou les termes recherchés.
Ex : `Bacillus subtilis [organism]`.

Construire des requêtes "manuellement" en indiquant les champs (comme par exemple `(Bacillus subtilis [organism] OR Yersinia pestis [organism]) AND MOTB [gene]`) peut s'avérer assez rapidement fastidieux ... Fort heureusement, en dessous de la zone de saisie de la requête, le lien *Advanced* donne accès à un outil de construction des requêtes nommé *Builder*.

Builder

Dans l'outil *Builder*, rechercher dans la liste des champs disponibles, celui qui permet de limiter les entrées à celle qui proviennent de l'**organisme** *Bacillus subtilis*. Puis, saisissez `Bacillus subtilis` dans la zone de saisie qui se trouve à côté de la liste des champs.

Question 10

Quelle est alors la requête automatiquement créée ?
La requête automatiquement construite est-elle identique à la précédente ?
Vérifiez en la cohérence à l'aide du nombre de résulats obtenus.

Recherchez maintenant la séquence du gène `MAKORIN1`, chez le poisson *Seriola quinqueradiata*. Si vous ne précisez pas de nom de champ pour le nom de gène, vous devez obtenir 4 entrées : 2 entrées d'ARNm (un complet et un partiel) et 2 entrées de séquences génomiques (une avec le gène complet, l'autre avec le gène incomplet).

Consultez ces 4 entrées et notez derrière quel "qualifier" est indiqué `MAKORIN1`. Vous trouverez quel champ interroger en consultant la liste des champs disponibles du formulaire de recherche avancé.

Question 11

Quelle est la requête alors automatiquement construite par *Builder* si le champ `[gene name]` est utilisé pour `MAKORIN1` en plus du champ `[organism]` pour `Seriola quinqueradiata`?
Combien d'entrées sont alors trouvées ?
Pourquoi les autres entrées sont perdues ?

Utilisation d'opérateurs booléens

Lorsque plusieurs termes sont recherchés, il est possible de les combiner à l'aide des opérateurs booléens :

ET : les deux termes sont tous les deux dans les entrées.

OU : au moins un des deux termes est dans l'entrée.

MAIS PAS : le premier terme doit être présent dans les entrées et les entrées qui contiennent le deuxième sont exclues.

Pour les banques du NCBI, les opérateurs booléens sont représentés par les mots suivants écrits en majuscules :

`AND` pour le ET,

`OR` pour le OU,

`NOT` pour le MAIS PAS

Nous allons travailler à la recherche de **protéines** possédant soit la fonction enzymatique qui porte le numéro EC 5.3.1.24, soit la fonction enzymatique 4.1.1.48, soit les deux. Il faut donc interroger la banque *"Protein"*. Pour commencer, recherchez s'il existe un champ qui correspond aux numéros EC. Puis, répondez aux questions suivantes en choisissant les opérateurs appropriés :

Question 12

Combien d'entrées décrivent une protéine qui possède la fonction 4.1.1.48 ?
Combien d'entrées décrivent une protéine qui possède au moins une des deux fonctions 4.1.1.48, 5.3.1.24 ?
Combien d'entrées possèdent les deux fonctions 4.1.1.48 et 5.3.1.24 ?

Historique des requêtes

Souvent les systèmes d'interrogation de banques de données mémorisent les requêtes effectuées par votre ordinateur depuis le début de votre connexion. Cet ensemble de requête est appelé l'historique.

L'historique est accessible soit dans la partie *History* qui se trouve à droite de l'écran vers le bas, soit à l'aide du lien *Advanced* qui permet de manipuler les requêtes). Vous pouvez ainsi afficher de nouveau les résultats d'une requête.

Question 13

Relancez la requête qui donne le gène **et** l'ARNm du gène MAKORIN1, chez le poisson *Seriola quinqueradiata*.

Liens entre banques

GQuery permet d'interroger de nombreuses banques de données (Nucleotides, Protein, PubMed, ...). Les données d'une banque peuvent être liées à celle d'une autre banque. Par exemple, des séquences nucléiques contenant un gène peuvent être liées aux protéines codées par ces gènes. Les liens sont

obtenus à l'aide de la section *Find related data* (voir bandeau à droite de la page).

Lien vers la banque Proteins

> Question 14
>
> A partir de la requête précédente, faites un lien vers la banque *Protein* pour obtenir les protéines codées par les 4 entrées.
> Combien d'entrées obtenez-vous ?
> Est-ce que les séquences des protéines obtenues vous semblent différentes ?

En fait, vous trouvez 2 protéines de 418 aa et 2 protéines de 435 aa puisque à chaque fois il y a un gène et un ARNm qui codent la même protéine. Les deux paires d'entrées protéiques sont générées automatiquement à partir des séquences nucléiques, il n'y a pas de suppression des séquences redondantes qui ont des numéro d'accession différents.

Lien vers la banque Genes

> Question 15
>
> Recherchez les gènes ou ARNm qui codent pour une protéine ayant une fonction *"selenophosphate synthetase"* (avec les guillemets), chez l'organisme *"Homo Sapiens"*.
> Quelle banque faut-il interroger ?
> Est-ce que les entrées trouvées sont redondantes ou correspondent à des gènes différents ?

Il y a, en plus des chromosomes et ARNm, des fragments de chromosomes (*genomic contig/scaffold*) issus du séquençage car ils ont été séquencés avant que le génome complet ne soit disponible. De plus, la liste de résultats contient de nombreux ARNm aux épissages multiples. Il est donc difficile de savoir combien de gènes sont associés à cette fonction et quel ARNm correspond à quel gène.

> Question 16
>
> Suivez le lien vers la banque *Gene*. Combien d'entrées obtenez-vous ?

Le nombre d'entrée après suivi du lien vers *Gene* est très important, car les (fragments de) chromosomes parasitent le résultat : en effet **tous** les gènes qui y sont contenus sont recherchés ...

Il est possible (par exemple) de sélectionner un certain nombre d'ARNm (en cliquant sur les "cases à cocher" à gauche de chaque entrée)

> Question 17
>
> Sélectionnez désormais les quelques ARNm intitulés "Homo sapiens selenophosphate synthetase 1 (SEPHS1)", et suivez le lien vers la banque *Gene*
> Combien d'entrées obtenez vous désormais ? Faites en la consultation.

Filtres

Pour éviter un travail fastidieux de sélection des ARNm, il est possible, après avoir lancé la requête précédente (`"selenophosphate synthetase" AND ("Homo Sapiens"[organism])`), d'utiliser dans le **bandeau gauche** les filtres *Molecule types* pour ne conserver que les ARNm.

> Question 18
>
> Lancez à nouveau la requête précédente.
> Dans le bandeau gauche, précisez `Molecule Type` et cliquez sur `mRNA` pour cocher ce type.
> Faites enfin le lien vers la banque *Genes*
> Combien d'entrées obtenez vous alors ?

Page réalisée par les membres de l'équipe bonsai