TD GÉNÉTIQUE : Recherche d'information dans les génomes

Sujet : Recherche de séquences codantes dans le génome de la levure *Candida glabrata*

Candida glabrata est un champignon unicellulaire pathogène opportuniste qui provoque, au niveau du tractus urogénital, des infections, chez les individus immunodéprimés (HIV positifs, transplantés, patients soumis à une chimiothérapie…).

Le génome de *Candida Glabrata* a été séquencé pour la première fois en 2004. Ce génome reste encore mal annoté puisque, pour plus de 90% des séquences codantes identifiées par des méthodes informatiques, la fonction de la protéine correspondante n'est pas connue. En comparaison, 80% des séquences codantes de la levure de boulangerie *Saccharomyces cerevisiae*, qui est un modèle très étudié, ont une fonction connue.

Il reste donc encore beaucoup de travail aux biologistes et aux bioinformaticiens pour pleinement comprendre la biologie du génome de *Candida glabrata*.

Dans ce TD, nous vous proposons d'explorer ce génome avec des méthodes classiques de recherche de gènes codants.

1) Base de données NCBI sur le génome de *Candida glabrata*

Base de données génomiques NCBI : https://www.ncbi.nlm.nih.gov/genome

a) Découverte des informations sur le génome de *Candida glabrata*

Utiliser la barre de recherche pour accéder aux données du génome de *Candida glabrata*

Questions/Discussions :
1- Combien de molécules d'ADN constituent le génome de *Candida glabrata*

Le génome de Candida glabrata est constitué de 13 chromosomes.

2- Analyse des données pour le chromosome A : Combien de gènes sont présents sur le chromosome A ? Donner le nombre de gènes codants et de gènes non codants présents sur ce chromosome. Calculer le nombre de gènes codants présents par kb (kiloBase) sur le chromosome A.

Il y a 209 gènes sur le chromosome A, 200 gènes codant et 9 gènes non codant. On peut voir qu'il y a 490 kB pour 200 gènes, ce qui fait 2.45 gènes codant/kB.

b) Format de la séquence chromosome A de Candida glabrata

Cliquer sur le lien RefSeq du chromosome A puis afficher la séquence de ce chromosome au format FASTA.

Questions/Discussions :
1- Quelles sont les caractéristiques du format FASTA ?

On peut voir des informations au début du fichier au format txt, un ID, un nom, le chromosome étudié etc… Ensuite, on a la séquence des nucléotides au format txt.

2 - Expliquez pourquoi une unique chaine de caractères est suffisante pour décrire la séquence d'ADN du chromosome A

Les nucléotides A, T, G, C ordonnés forment la séquence ADN, c'est dans cette séquence qu'est codé l'information génétique. Cette séquence est donc modélisable par une suite ordonnées de caractères.

3 - Copier les 10 premiers nucléotides de la séquence du chromosome A en indiquant l'orientation de cette séquence.

5' -> TCAAAGGTAT -> 3'

4 - Ecrire la séquence des 10 premiers nucléotides du chromosome sous forme double brin en conservant l'orientation du brin 1 donnée par la base de données. La séquence obtenue pour le brin 2 est dite **complémentaire** par rapport au brin 1.

5' -> TCAAAGGTAT -> 3'
3' <- AGTTTCCATA -< 5'

5 - Ecrire au format Fasta la séquence du brin 2. Cette séquence est dite **reverse-complémentaire**.

3' <- AGTTTCCATA <- 5'
5' -> ATACCTTTGA -> 3'

6 - Utiliser l'outil en ligne de convertion de séquence pour vérifier votre travail sur la séquence des 10 premières bases du chromosome A. A quoi correspondent les opérations « reverse », « complément » et « reverse complément » proposées par cet outil.

Reverse : change l'ordre de la séquence de 5' 3' à 3' 5' ou inversement.
Complément : créer une nouvelle séquence en remplaçant les A par des T, les G par des C et inversement.
Reverse complément : composition des deux opérations reverse et complément.

2) Analyse du chromosome A de *Candida glabrata*

a. Recherche des séquences codantes sur le chromosome A de Candida glabrata.

Outil de recherche de séquence : https://www.ncbi.nlm.nih.gov/orffinder/

Lien pour les informations sur le code génétique utilisé par l'outil ORF finder.

Utiliser l'outil de ORF FINDER de NCBI pour rechercher des séquences codantes sur le chromosome A de *Candida glabrata*. Vous pouvez soit donner en entrée le numéro accession de ce chromosome (NC_005967.2), soit copier la séquence Fasta associée.

Pour cette recherche vous limiterez les coordonnées à analyser à la région située entre les coordonnées 50 000 et 75 000 du chromosome A.

Dans un premier temps lancer cette analyse avec les paramètres par défaut.

Cliquer sur le lien **Six-Frame Translations> Add six-frame translation track** situé en bas de la figure montrant la détection des ORF.

Questions/Discussions :

1- Observez la représentation Six-Frame Translations dans laquelle l'analyse de la distribution des codons initiateurs (barres vertes) et Stop a été effectuée sur la séquence. Pourquoi cette analyse est-elle faite sur 6 cadres de lecture ? Comment détecter des séquences codantes (CDS) à partir de cette représentation ?

On peut étudier une séquence de nucléotides en n, n+1, n+2 car les codons sont constitués de 3 nucléotides. Il y a deux brins, on a donc 2x3 = 6 cadres de lecture possible.

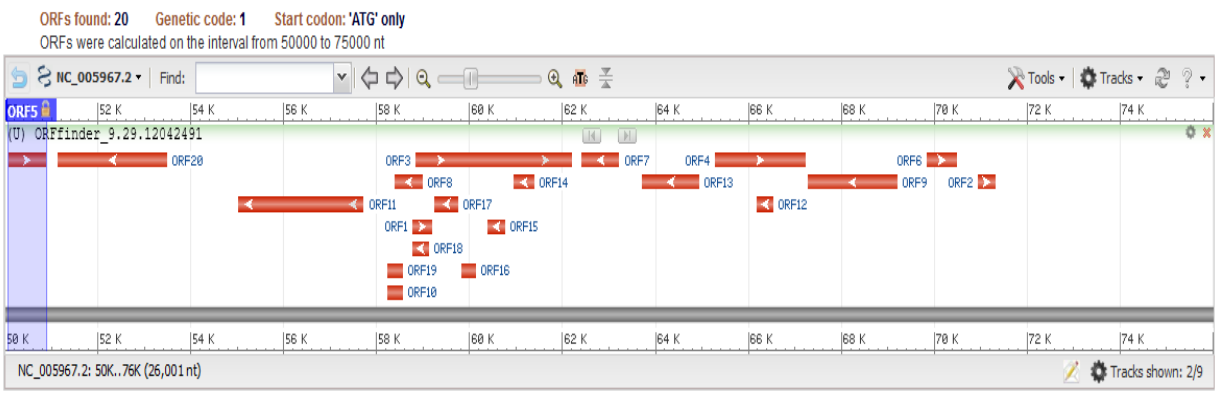
2- Observer le résultat de l'analyse ORF FINDER. Combien de CDS ont été détectées avec les paramètres par défaut ? Que pensez-vous de cette analyse compte tenu du nombre moyen d'ORF par kb calculé précédemment ?

ORF FINDER détecte 139 CDS.
On avait 2.45 CDS/kB.
Ici on utilise ORF FINDER sur une séquence de 25 000 nucléotides.
Sachant que 25 000 nucléotides = 25 kB.
On a donc 139/25 = 5.56 CDS/kB.
Il y a donc environ 2 fois plus de CDS avec cette analyse.

3- Quels paramètres pourraient être modifiés pour augmenter l'efficacité de détection des CDS ? Refaire l'analyse en modifiant les paramètres de la recherche qui vous semblent importants. Donnez les paramètres choisis pour obtenir la table de détection ci-dessous.

On peut modifier le choix des codons initiateurs, par exemple et prenant en compte d'autres codons initiateur que ATG. On peut aussi modifier le nombre minimum de nucléotides par ORF. Avec les paramètres ATG Only et > 300 nucléotides, on retrouve le tableau ci-dessous.

Numéro	Brin	Cadre de lecture	Début	Fin	Taille (nt |aa)
ORF5	+	1	50068	50880	813 | 270
ORF20	-	3	53497	51137	2361 | 786
ORF11	-	2	57722	55014	2709 | 902
ORF19	-	3	58555	58220	336 | 111
ORF10	-	2	58562	58230	333 | 110
ORF1	+	2	58775	59191	417 | 138
ORF3	+	3	58830	62198	3369 | 1122
ORF8	-	1	58995	58378	618 | 205
ORF18	-	3	59119	58760	360 | 119
ORF17	-	3	59749	59249	501 | 166
ORF16	-	3	60127	59813	315 | 104
ORF15	-	3	60766	60386	381 | 126
ORF14	-	3	61393	60947	447 | 148
ORF7	-	1	63225	62416	810 | 269
ORF13	-	3	64948	63710	1239 | 412
ORF4	+	3	65286	67226	1941 | 646
ORF12	-	3	66538	66188	351 | 116
ORF9	-	2	69203	67281	1923 | 640
ORF6	+	1	69832	70485	654 | 217
ORF2	+	2	70937	71320	384 | 127



4- A partir de cette détection, tester l'option « Ignored Nested ORFs ». Sur la figure et la table précédente, indiquez les ORF impactées par cette option.

b. Comparaison de la prédiction obtenue avec ORF FINDER avec l'annotation du génome.

Site de visualisation d'annotation de génome : https://www.ncbi.nlm.nih.gov/projects/sviewer/

Visualisation centrée sur la région [50000-75000] du chromosome A de Candida Glabrata

Questions/Discussions :

Le site de visualisation des données d'annotation montre les CDS pour lesquelles un ARN a pu être détecté expérimentalement. Le cadre six frames translation montre en grisée les CDS de plus de 300 nucléotides détectés de manière informatique.

1- Comparez l'annotation présentée à la détection effectuée précédemment. Annoter la figure ci-dessus pour mettre en valeur : (1) les CDS correctement détectées, (2) les CDS détectées mais qui n'ont pas été confirmées par l'annotation, (3) Les CDS non détectées, (4) les gènes non codants annotés.

(1): Barres vertes
(2): Barres grises n'étant pas présente aussi en barre verte
(3): Carré verts, ce sont des séquences où on sait qu'il y a une CDS car on a observé par l'expérimentation qu'une protéine existe pour ce gène. Le logiciel ne détecte peut être pas la CDS à cause des paramètres choisis.
(4): On peut voir des gènes non codant pour le tRNA, ce sont les petits chromosomes violets sur l'image.

2- Faire un bilan des différences observées entre la prédiction obtenue avec ORF FINDER et l'annotation du génome à partir de données expérimentales ? Comment expliquer ces différences ?

Il y a des faux positifs et des faux négatifs suivant les paramètres que l'on choisit pour effectuer notre analyse. Il n'y a pas de paramètres optimaux, il faut souvent faire une analyse en utilisant plusieurs paramètres pour ne pas omettre certains détails.

3) Analyse du chromosome mitochondrial de *Candida glabrata*

a. Observations de gènes annotés sur le chromosome mitochondrial

Annotation du chromosome mitochondrial de *Candida glabrata* sur NCBI viewer : Accession NC_004691.1

Questions/Discussions :

1 -Combien de gènes codants sont présents sur le génome mitochondrial de *C. glabrata* ?

Il y a 11 gènes codants sur le génome mitochondrial de Candida glabatra.

2 -Le gène COX1 est formé d'une succession d'exons et d'introns. Cliquer sur le gène COX1 pour faire apparaitre la limite des exons et des introns dans le cadre « Six-Frame Translation ». Les exons de COX1 sont-ils tous codés dans le même cadre de lecture ? Comment l'élimination d'un intron peut-il permettre de changer de cadre de lecture ?

Les exons de COX1 ne sont pas tous codés dans le même cadre de lecture, ils sont en fait présents sur 2 cadres de lecture. Un intron n'a pas forcément un nombre de nucléotides multiple de 3, il est donc possible que la suppression d'un intron fasse changer le cadre de lecture.

3- Le génome mitochondrial de C. glabrata contient de très nombreux gènes non codants. A quoi correspondent ces gènes non codants ?

Ces gènes non codants codent des ARN de transfert, des ARN ribosomiques et de la ribonucléase P.

b. Utilisation du logiciel ORF FINDER pour détecter les CDS du chromosome mitochondrial de *Candida Glabrata*.

Utiliser l'outil ORF FINDER en utilisant l'identifiant du chromosome mitochondrial de *Candida Glabrata* en entrée (NC_004691.1).

Questions/Discussions :

1-Combien de séquences codantes sont détectées avec les paramètres par défaut ? Cette détection est-elle en cohérence avec l'annotation du génome mitochondrial ? Donner une explication au résultat obtenu ?

En utilisant les paramètres par défaut, on a 58 ORF. Cette détection n'est pas cohérente avec l'annotation du génome mitochondrial car on peut avoir des faux négatifs et faux positifs suivant les paramètres choisis. Les introns peuvent aussi contenir des codons stop, ce qui change le nombre

d'ORF détectés . COX1 n'est pas présent sur cette analyse. Cependant on peut voir des ORF sur plusieurs cadres de lecture à l'endroit où COX1 était présent.

2-Quels paramètres modifier pour obtenir la détection la plus efficace des CDS mitochondriales de *Candida Glabrata* ?

On peut augmenter le nombre minimum de nucléotides dans notre recherche. On peut aussi ignorer les ORF imbriqués, cela augmentera l'efficacité de notre recherche. Ces changements fonctionnerait car il y a beaucoup de faux positifs lorsque l'on étudie les ORF constitués de peu de codons et les ORF inclus dans de plus grands ORF. Ce dernier point s'explique par le fait qu'il peut y avoir plusieurs codons codant pour la méthionine dans un grand ORF sans forcément être des codons start.

3-Retourner sur la page d'annotation du génome mitochondrial de *C. Glabrata* (Accession NC_004691.1). Comparer la détection informatique qui apparait dans le cadre six-frame translations (CDS de plus de 300 nucléotides grisées) à l'annotation basée sur des données expérimentales.

Quelles sont les limites de la méthode ORF FINDER mises en évidence par cette comparaison ? Donner des exemples de séquences codantes non détectées et expliquer pourquoi elles n'ont pas été correctement détectées.

ORF FINDER ne trouve que 2 ORF constitués de plus de 300 nucléotides. Lorsque l'on regarde l'analyse, on voit que la première ORF correspond à i-Cgill, qui n'est pas une protéine. C'est donc un gène non codant qui correspond à i-Cgill. Ensuite d'autres CDS n'ont pas été détectés, c'est le cas de COX1. Cette CDS n'apparaît pas car il y avait probablement des codons stop dans les introns (qui ne sont plus là après epissage). Ces codons stop n'étant plus présents, ORF FINDER ne trouve pas les ORF associés. Une partie des ORF est aussi supprimée car on prend un minimum de 100 codons pour notre recherche.

4-Le génome mitochondrial de *C. glabrata* contient de nombreux gènes d'ARNt. Pourquoi ces gènes ne sont-ils pas détectés par ORF Finder ?

Les gènes codant les ARN de transfert ne sont pas détectés car ils ne commencent pas par des ATG et ne finissent pas par des codons stop comme pour les CDS.

5- Le génome mitochondrial est dérivé du génome d'une bactérie ancestrale capable d'effectuer la respiration et qui est entrée en symbiose avec l'ancêtre de la cellule eucaryote. Au cours de l'évolution, les gènes mitochondriaux redondants avec les gènes de la cellule hôte ont été perdus, ce qui fait que la mitochondrie a perdu son autonomie. Votre analyse de ce génome vous permet-elle de comprendre pourquoi le génome mitochondrial a néanmoins conservé ses gènes d'ARNt ?

Pour que la mitochondrie puisse produire ses propres protéines afin d'effectuer la respiration, il fallait qu'elle garde les éléments permettant d'effectuer ces opérations. Les ARN étant utiles à la traduction ont donc été conservés dans le génome mitochondrial.