TP BIOLOGIE : 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. Lors de votre projet, vous serez amenés à développer des outils pour découvrir des motifs permettant le contrôle de l'expression des séquences codantes de *Candida glabrata*.

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*
14 molécules d'ADN

2- Combien de gènes contiennent chaque chromosome ? Quelles sont les différentes catégories de gènes indiqués dans la table ? A quoi correspondent ces catégories ?
A = 209, B = 222, C = 237, D = 301, E = 287, F = 398, G = 451, H = 474, I = 489, J = 533, K = 581, L = 605, M = 634, MT = 37

A = Protein; tRNA, B = Protein; tRNA; Other RNA, C = Protein; tRNA, D = Protein; tRNA, E = Protein; tRNA, F = Protein; tRNA, G = Protein; tRNA, H = Protein; tRNA, I = Protein; tRNA; Other RNA, J = Protein; tRNA; Other RNA, K = Protein; tRNA; Other RNA, L = Protein; rRNA; tRNA; Other RNA, M = Protein; tRNA; Other RNA, MT = Protein; rRNA; tRNA; Other RNA,

Protein = formés de chaine polypeptidique
tRNA = acides ribonucléiques de transfert
RNA = acides ribonucléiques
rRNA = acides ribonucléiques ribosomique

b) Format de la séquence chromosome 1 de Candida glabrata

Cliquer sur le lien RefSeq du chromosome 1 puis afficher la séquence du chromosome 1 au format FASTA.

Questions/Discussions :

1- Quelles sont les caractéristiques du format FASTA ?
Une séquence commence par « > » suivi de RefSeq et de la description de la séquence. Il y a un retour a la ligne puis la séquence de nucléotides complète.

2 - Expliquez pourquoi une unique chaine de caractères est suffisante pour décrire la séquence d'ADN du chromosome A
Une unique chaine de caractère est suffisante un brin suffit pour déduire le second.

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.
3'-AGTTTCCATA-5'

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

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 inverse l'orientation de la séquence
Complément donne le complément de la séquence
Reverse Complément combine les deux

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/

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.

Configurer l'affichage du résultat, pour afficher l'analyse ORF MAP en cliquant **TRACKS>Configure TRACKS>Sequence>Six-Frame Translations**

Questions/Discussions :

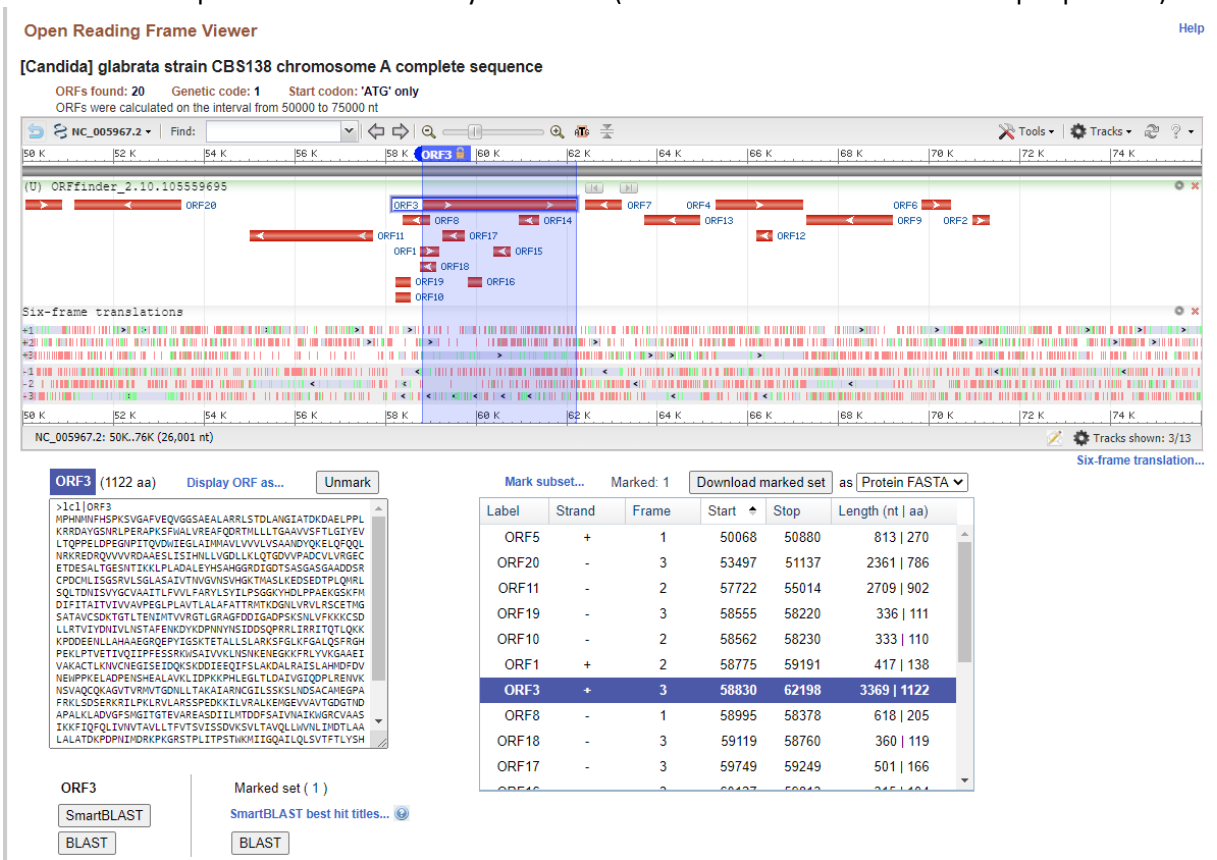
1- Observez la représentation Six-Frame Translations dans laquelle l'analyse de la distribution des codons initiateurs (barres vertes) et Stop a été effectuée sur la séquence. Pourquoi cette analyse est-elle faite sur 6 cadres de lecture ? Comment détecter des séquences codantes (CDS) à partir de cette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

On a besoin d'une partie positive pour le brin 1 et d'une partie negative pour le brin deux. Ces deux parties sont composés de 3 cadres de lectures car un acide aminé est constitué de 3 nucléotides. Les 3 cadres sont donc 3k, 3k + 1 et 3k + 2. Cela nous permet de couvrir l'ensemble des possibilités.
On les détecte grâce à la présence d'un trait vert (codon start) suivi d'un trait rouge (codon stop) avec une distance minimale de 75.

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 ?
139 CDS ont été détectés.
De toutes petites séquences ont été détectées elles ne sont peut-être pas indispensables.

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 et le nombre de CDS détectées dans ces conditions.
Le paramètre : « Minimal ORF length »
On peut l'augmenter à 300.
20 CDS sont maintenant détectés

4- Coller ici la capture d'écran de l'analyse retenue (schéma + table des ORF classées par position)



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

Site de visualisation d'annotation de génome : h NC_004691.1ttps://www.ncbi.nlm.nih.gov/projects/sviewer/

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

Le site de visualisation des données d'annotation montre les CDS pour lesquelles un ARN a pu être détecté expérimentalement.

Questions/Discussions :

1- Combien de gènes sont annotés dans la région du chromosome A étudiée ?
10 gènes

2- Quelles sont les différences observées entre la prédiction obtenue avec ORF FINDER et l'annotation du génome à partir de données expérimentales ? Comment expliquer ces différences ?
Le nombre de gènes que l'on compte est diffèrent donc on peut supposer que nos paramètres sont responsables de ces différences
De plus les ORFs ne correspondent pas forcément a des gènes puisqu'un ORF est généré par un codon start et un codon stop. Or ces codons sont présents un peu partout dans le génome.

c. Analyse détaillée de CDS : de la séquence d'ADN à la protéine

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

Nous vous proposons d'étudier deux des CDS détectées par l'outil ORF FINDER dont les coordonnées sont [50068-50880] et [53497-51137].

Pour chacun de ces deux CDS, cliquer sur le bouton ATG pour zoomer au niveau de la séquence puis puis utiliser le cadre FIND pour vous placer soit au début soit à la fin de la séquence. Observer la séquence nucléotidique et celle de la protéine produite.

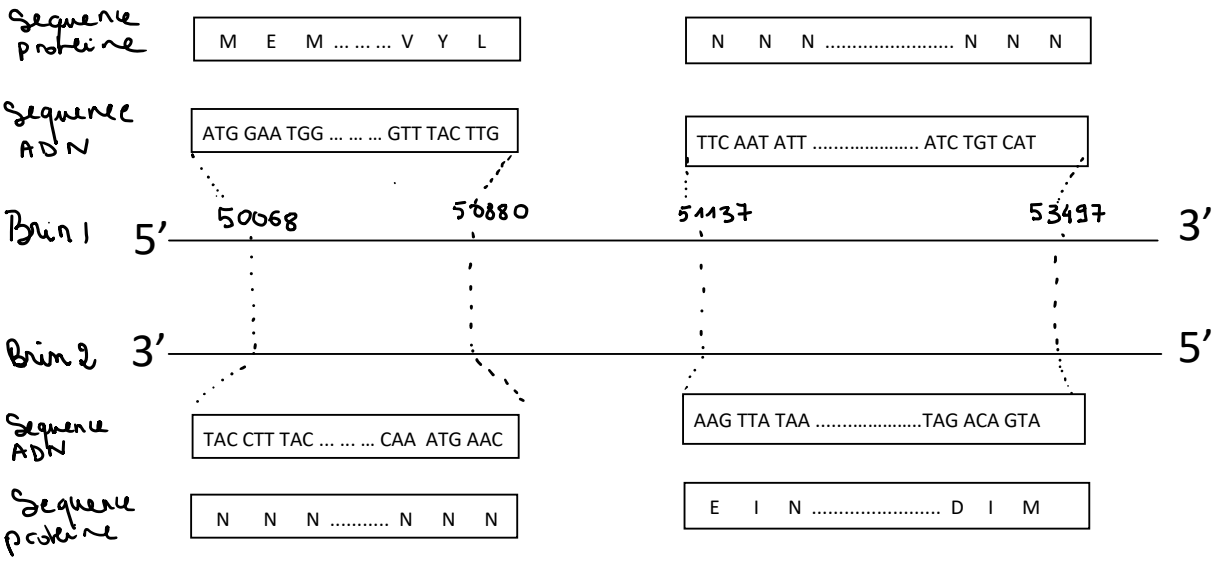
Questions/Discussions :

Compléter le schéma suivant à partir de votre analyse :

1- Indiquer l'orientation des deux brins (remplacer les ? par 5 ou 3)

2- Donner la séquence correspondant aux 3 premiers et derniers codons de chaque CDS pour les 2 brins : Remplacer les N par les lettres appropriées, cette séquence sera surlignée pour le brin codant

3- Donner les séquences en acides aminés correspondantes : Remplacer les N par les lettres appropriées uniquement dans le cadre correspondant au brin traduit en protéine (effacer l'autre cadre)



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* ?
37 gènes

2 -Le gène COX1 est formé d'une succession d'exons et d'introns. Après transcription, un ARNm (m : messager) prématuré est produit qui contient à la fois les exons et les introns. Cet ARNm est ensuite maturé de manière à éliminer les séquences des 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 ?
Exon 1 = +2
Exon 2 = +1
Non les exons ne sont pas dans le même cadre de lecture.
Un intron n'est pas traduit en protéine, son élimination peut permettre le décalage des fenêtres de lecture pour la traduction des triplets de nucléotides constitutifs des exons en acides aminés.

3- Le génome de C. glabrata contient de très nombreux gènes non codants. A quoi correspondent ces gènes non codants ?
Ces gènes non codants correspondent aux ARNs ribosomiques et aux ARNs de transfert.

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 :

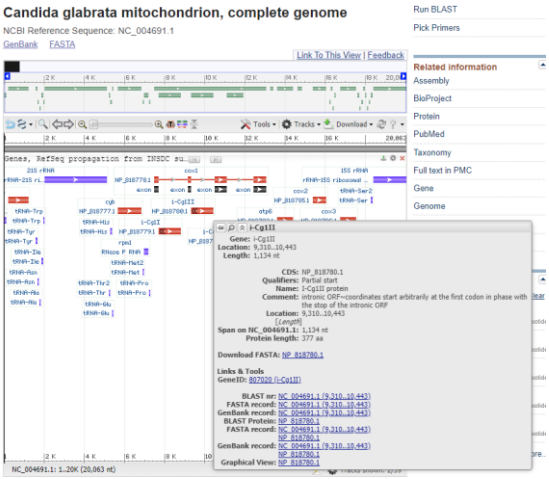
1-Combien de séquences codantes sont détectées avec les paramètres par défaut ? Cette détection est-elle en cohérence avec l'annotation du génome mitochondrial ? Donner une explication au résultat obtenu ?
58 seq codantes
Non absolument pas, 58 > 37

2-Quels paramètres modifier pour obtenir la détection la plus efficace des CDS mitochondriales de *Candida Glabrata* ?
Min ORF length : 150
Genetic code : Yeast Mitochondrial

ORF start codon : Any sense codon
Ignore nested ORF's : YES
Found : 38 plutôt satisfaisant

3- Quelles sont les limites de la méthode ORF FINDER mises en évidence par l'analyse du génome mitochondrial de *Candida Glabrata*. Donner des exemples de séquences codantes non détectées et expliquer pourquoi elles n'ont pas été correctement détectées.
- cytochrome-c oxidase subunit n'est pas détecté on peut penser que c'est parceque l'ORF est trop longue. (>5594 nt) Peut être aussi que les introns sont trop long.

- i-Cg1II n'est pas détecté, il est indiqué « Partial Start »



4-Le génome mitochondrial de *C. glabrata* contient de nombreux gènes d'ARNt. Pourquoi ces gènes ne sont-ils pas détectés par ORF Finder ?
Nos réglages ne nous permettent pas de détecter les ARNs qui ont une taille < 150nt.

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 ?
Malgré que la mitochondrie ait perdu son autonomie elle a toutefois conservé son propre génome ainsi la conservation de ses ARNt lui permet d'assurer la traduction des ARNm en protéine pour le bon fonctionnement des organites présents en son sein.