**Tutoriel annotation manuelle d’un gène connu d’intérêt dans une séquence de génome non annotée**

Ce tutoriel présente la manière d’annoter un gène d’intérêt lorsqu’il n’est pas annoté dans un génome. Il peut aussi servir pour vérifier et éventuellement corriger des annotations si elles sont erronées. Il combine une annotation par similarité de séquence couplée à la reconnaissance *ab initio* de motifs (démarrage de traduction et jonctions intron/exon). Il requiert la possibilité de faire des BLAST et l’utilisation du logiciel ARTEMIS (https://www.sanger.ac.uk/tool/artemis/).

- Déclarer la séquence inconnue comme une banque de BLAST si elle n’est pas disponible sur un serveur BLAST.

Sous PC utiliser Bioedit pour faire des BLAST en local (<https://bioedit.software.informer.com/T%C3%A9l%C3%A9charger/>). Pour déclarer la banque, faire accessory application -> BLAST-> create a local nucleotide database et charger votre fichier de sequence.

- utiliser la séquence du gène connu pour interroger par tBLASTn la banque nouvellement créée. Pour l’interprétation des données, il faut connaître (1) la distance phylogénétique entre l’espèce d’où provient le gène connu et celle d’où provient la séquence à annoter, (2) savoir si le gène évolue lentement ou vite. Dans le cas des gènes des codes-barres fongiques, les séquences évoluent lentement et il n’y a donc pas de problème pour reconnaitre la position de la séquence de l’orthologue dans le génome à annoter. Voici par exemple, le résultat du BLAST du premier code barre de *Podospora anserina* sur le génome de *Cercophora samala*. Le premier hit montre la très bonne conservation des séquences entre les deux espèces (qui sont très proches) ainsi que la présence d’un intron (visualisé par la présence d’une séquence contenant des codons stop et qui ne matche pas avec la séquence de *Podospora anserina*). Notez que sur ce hit la similarité de séquence commence à l’acide aminé n°98. Ce qui montre qu’il y a au moins un exon supplémentaire qui code pour la portion N-terminale de la protéine.



intron

Et le hit suivant est effectivement un autre exon qui est localisé sur le même contig que les deux exons précédents et à quelques dizaines de paires de bases (le résultat de blast donne les coordonnées sur le contig). Il code pour une partie de la séquence N-terminale de l’acide aminé 12 au 97 :



Par contre les suivants ne sont pas des bons hits car trop divergents et localisés sur d’autres contigs. Ils correspondent soit à des paralogues éloignés soit à des faux positifs :



En regardant les coordonnées des hits positifs assurés, on constate qu’il semble manquer les 10 à 12 premiers acides aminés. C’est dû à un score BLAST de cette partie trop faible (c’est-à-dire une e-value trop elevée) pour être retenu. Notez qu’il est possible de baisser le seuil de détection BLAST (il faut augmenter la e-value dans les paramètres), mais dans ce cas le nombre de faux positifs augmente… C’est dû au fait que les portions N et C terminales des protéines évoluent généralement rapidement et que BLAST les aligne donc mal. Il est aussi possible que ce soit dû à la présence d’un intron qui génère un exon trop petit pour être détecté par BLAST. Dans ce cas, il faudra faire une évaluation de la séquence au cas par cas pour savoir ce qu’il se passe.

Maintenant que les scores en BLAST sont définis, il faut ouvrir la séquence du génome avec artemis et rechercher les régions de similarité (pour ouvrir le navigateur, il faut faire ctrl-G) et définir avec ctrl-C (pour create) les portions de CDS. Dans l’exemple ci-dessus, il faut donc définir les 3 exons et les fusionner en utilisant la commande ctrl-M. Les jonctions exon/exon sont alors mal définies et la CDS n’aura pas son codon start.

Pour définir les jonctions exon/exon, il faut rechercher les consensus qui sont chez les dikarya :

Jonction exon/5’-intron : exon/GTaaGt (surlignés en jaune les premiers nucléotides de l’intron) avec les nucléotides en majuscules étant les plus importants, en particulier les deux premiers nucléotides sont quasiment toujours GT. Attention, il existe cependant un variant qui sera obligatoirement GCAAGT (les introns de ce type représentent moins de 1% des introns et sont épissés par un variant particulier du spliceosome). Il peut y avoir jusqu’à 3 différences avec ce consensus, ce qui rend la détection de cette jonction pas toujours simple…

Jonction 3’-intron/exon : PyAG/exon (surlignés en jaune les derniers nucléotides de l’intron).

Point de branchement du lariat : CTNA (le A étant le point de branchement) 15 à 20 nucléotides en amont de la jonction 3’

Voici deux introns correctement annotés avec leur point de branchement entourés en rouge :





Pour la recherche du codon start, regarder s’il n’y a pas un start (ATG) dans la première ORF de la CDS qui pourrait correspondre au codon start. En général, en position -3 d’un codon start on doit avoir une purine : PuXXATG. Sinon, voir la possibilité de la présence d’un intron supplémentaire.

Attention, la taille des introns est variable. Souvent entre 50 et 70 pb, elle peut monter jusqu’à 2 000 pb. Si vous avez des difficultés à trouver les introns et/ou le codon start, vous pouvez utiliser FGENESH (<http://www.softberry.com/berry.phtml?topic=fgenesh&group=programs&subgroup=gfind>) ou augustus (https://bioinf.uni-greifswald.de/webaugustus/prediction/create). Notez que si vous disposez de données de RNAseq de votre espèce à annoter, vous pouvez mapper les RNAseq avec TopHat ou TopHat2 et charger les fichiers BAM pour vous aider à trouver les jonction précises et les codons start !