**Annotation d'un fragment génomique**

**Plate-forme d'annotation**

L'annotation d'une séquence génomique fait généralement appel à différentes méthodes de prédictions. Ces méthodes prédisent soit le même type d'objet biologique (gènes codant pour une protéine, gènes d'ARN non codant...), soit des objets complémentaires (sites d'initiation de la traduction, promoteurs et terminateurs de la transcription...). Afin de faciliter l'interprétation de ces différentes sources d'information par l'annotateur, il est nécessaire de les visualiser simultanément en regard de la séquence d'ADN à annoter. C'est l'objectif des plate-formes d'annotation, qui implémente une stratégie d'annotation en intégrant dans un environnement graphique commun les méthodes de prédiction (ou au moins les résultats de ces méthodes). Une interface conviviale permet à l'utilisateur de procéder à l'annotation de la séquence d'ADN et de produire un fichier formaté prêt pour la soumission de la séquence à l'une des banques publiques. De nombreuses plate-formes ont été développées, principalement par des groupes engagés dans le séquençage de génomes complets. Ces logiciels complexes nécessitent généralement l'installation de logiciels et paquets distribués indépendamment et une configuration personnalisée dont ils tirent leur puissance. Afin de ne pas nous noyer dans cet aspect système du problème, nous allons utiliser dans ce TD un logiciel simple à installer, très largement utilisé dans la communauté mais qui n'est pas nécessairement le plus puissant.

**Artemis**

[Artemis](http://sanger-pathogens.github.io/Artemis/Artemis/) est une plate-forme de visualisation et d'annotation des séquences d'ADN développée au [Sanger Institute](http://www.sanger.ac.uk/) par le groupe de K. Rutherford. Le logiciel est distribué librement sous licence publique GNU. Son installation nécessite une version récente de java. Le manuel d'utilisation d'artemis est disponible au téléchargement sur ce même site.

**Premiers pas**

La première étape consiste à lire la séquence à annoter dans artemis. Pour cela lancer la commande art et dans File utiliser le File Manager pour sélectionner le fichier. On obtient une fenêtre découpée horizontalement en trois parties: une partie supérieure avec la distribution des codons stops dans les six phases de lecture de la séquence, une partie inférieure sans rien pour le moment mais destinée aux annotations en format textuel et entre les deux la séquence ADN sur le brin direct et complémentaire inversé avec les traductions en acides aminés pour les six phases de lecture. Les échelles sont différentes pour les différentes parties et on peut naviguer indépendamment dans l'une et l'autre, avec le curseur horizontal. Les curseurs verticaux, placés à droite, permettent de réduire/agrandir l'échelle (zoom). Les différentes parties peuvent être mises en correspondance par un clic gauche de la souris. Le bouton du milieu permet de sélectionner une région entre deux codons stops. Le bouton droit de la souris donne accès à différentes fonctions mais permet également de contrôler l'apparence graphique.

Avec l'option **Mark Open Reading Frames** de **Create**, on peut annoter les phases ouvertes de lecture d'une longueur supérieure à un seuil choisi. Ces ORF sont définis entre deux codons stop. On peut demander à ce que le début des ORF soit fixé au premier codon codant pour Met avec l'option **Trim selected Features to Met** dans le menu **Edit**.

**Annotation**

Artemis ne dispose pas de méthode de prédiction des gènes performante. On peut, pour s'en convaincre, utiliser la méthode du Chi2 disponible dans l'onglet Graph. Nous allons utiliser des méthodes externes pour notre annotation. Pour cela, il est nécessaire de re-formater les sorties de ces programmes de façon à ce qu'elles soient compatibles avec artemis. Outre le format FASTA que nous avons déjà vu, artemis peut lire les formats EMBL et GENBANK (séquences déjà annotées) ou le format [GFF](http://www.sanger.ac.uk/Software/formats/GFF/) (General Feature Format).

Exemple:

SEQ1 EMBL exon 103 172 . + 0

SEQ1 EMBL splice5 172 173 . + .
SEQ1 netgene splice5 172 173 0.94 + .
SEQ1 genie sp5-20 163 182 2.3 + .
SEQ1 genie sp5-10 168 177 2.1 + .
SEQ2 grail ATG 17 19 2.1 - 0

Ce format associe à la séquence SEQ1 différentes features (exon, splice5, ATG...) provenant de différentes sources (annotation EMBL, méthodes de prédiction) avec le score associé, positionnées sur la séquence (avec indication du brin +/- et de la phase de lecture) et une liste optionnelle d'attributs avec un format tag value séparés par des ';'.

Si nous reprenons les résultats de GenMark :

List of Open reading frames predicted as CDSs, shown with alternate starts
(regions from start to stop codon w/ coding function >0.50)

Left Right DNA Coding Avg Start
end end Strand Frame Prob Prob
-------- -------- ---------- ----- ---- ----
 1 276 direct fr 1 0.68 ....
 28 276 direct fr 1 0.68 ....
 112 276 direct fr 1 0.58 ....

 1191 3587 direct fr 3 0.51 0.32
 2397 3587 direct fr 3 0.51 0.00
 2589 3587 direct fr 3 0.59 0.97
 2619 3587 direct fr 3 0.57 0.02

Nous allons devoir les re-formater pour obtenir un fichier du type :

BS09819 GenMark CDS 1 276 0.68 + . gene GM\_CDS\_1.1
BS09819 GenMark CDS 28 276 0.68 + . gene GM\_CDS\_1.2
BS09819 GenMark CDS 112 276 0.58 + . gene GM\_CDS\_1.3
BS09819 GenMark CDS 1191 3587 0.51 + . gene GM\_CDS\_2.1
BS09819 GenMark ATG 1191 1193 0.32 + .
BS09819 GenMark CDS 2397 3587 0.51 + . gene GM\_CDS\_2.2
BS09819 GenMark ATG 2397 2399 0.51 + .
BS09819 GenMark CDS 2589 3587 0.59 + . gene GM\_CDS\_2.3
BS09819 GenMark ATG 2589 2590 0.97 + .
BS09819 GenMark CDS 2619 3587 0.57 + . gene GM\_CDS\_2.3
BS09819 GenMark ATG 2619 2621 0.02 + .
...

Pour cela, vous pouvez utiliser le langage de programmation qui a votre préférence et/ou le Perl ([introduction](http://sylvain.lhullier.org/publications/perl.html) à la programmation en Perl par **Sylvain Lhullier).**