Titre : Développement d'une interface pour l'analyse et la classification de séquences d'anticorps scFv

Contexte :
Dans le cadre du projet ABCardionostics, l'objectif est de concevoir de nouveaux outils pour la détection précoce des plaques d'athérome instables, à l'origine de pathologies cardiovasculaires majeures comme l'infarctus ou les AVC. Ce projet vise à offrir des solutions thérapeutiques personnalisées, reposant notamment sur l'ingénierie d'anticorps humains.
Les anticorps au format scFv (combinaison des régions variables VH et VL des chaînes lourdes et légères des anticorps) ont été sélectionnés in vivo et séquencés grâce à la technologie NGS PacBio. Ces régions variables, VH (Variable Heavy) et VL (Variable Light), sont essentielles pour la reconnaissance des antigènes. Elles se distinguent par leur organisation structurale en segments spécifiques, définis selon les standards d'annotation d'IMGT (International ImMunoGeneTics Information System).
Le stage s'inscrit dans cette démarche, en se concentrant sur l'analyse des séquences obtenues afin d'identifier les anticorps pertinents, spécifiques des plaques d'athérome.

Le travail inclura :

Une analyse statistique des données VH, VL et combinaisons VH-VL, visant à évaluer leur abondance selon deux axes :
- Fréquence des séquences VH, VL ou de leurs combinaisons.
- Analyse des variantes d'un anticorps monoclonal identifié par des techniques biologiques classiques.
- Une classification des séquences les plus représentées.
- Une sélection des 100 séquences scFv uniques les plus enrichies.

Objectifs du projet :
Développer un outil client pour l'analyse des séquences d'anticorps scFv (~1000 paires de bases par séquence) intégrant plusieurs fonctionnalités basées sur les standards IMGT (https://www.imgt.org/HighV-QUEST/ (définition des CDR, FR, mutations ...). Cet outil permettra une analyse fine des séquences d'anticorps en facilitant leur exploration, leur classification, et leur mise en relation avec des données biologiques ou cliniques. Il contribuera ainsi à l'avancée des thérapies ciblées et personnalisées pour les pathologies cardiovasculaires.

Cet outil intégrera les fonctionnalités suivantes :

Préparation des données et décomposition des séquences :
- Écrire un script permettant de décomposer les fichiers en séquences VH et VL, en identifiant les motifs spécifiques pour extraire ces séquences.
- Assigner des identifiants uniques aux séquences VH (VH1, VH2, etc.) et VL (VL1, VL2, etc.), y compris après alignement pour reconnaître les séquences supplémentaires dans les bases de données.

Alignement et comparaison des séquences :
- Comparer chaque séquence Query (VH-VL, VH, VL) avec les séquences Target (~300 000 séquences) en utilisant l'algorithme Smith-Waterman.
- Développer un module pour filtrer et spécifier les paramètres d'alignement.
- Renommer les combinaisons VH-VL en fonction des alignements identifiés.

Analyse des résultats via une interface utilisateur :

- Concevoir une interface pour explorer qualitativement et quantitativement les résultats d'alignement :
- Qualitatif : Visualiser les correspondances entre les différentes séquences alignées.
- Quantitatif : Quantifier le nombre de correspondances par Query (VH-VL, VH, ou VL).
- Intégrer une vue tabulaire où chaque ligne correspond à une Query, permettant d'afficher et de filtrer les résultats par VH-VL, VH ou VL.
- Ajouter des visualisations permettant de vérifier les alignements spécifiques (par exemple, pour chaque VH, identifier combien de fois il est associé à un VL particulier dans les Target).
- Explorer la combinatoire des résultats grâce aux identifiants des séquences alignées.

Analyse statistique et sélection :

- Réaliser une analyse descriptive des séquences les plus enrichies (fréquence, abondance).
- Identifier les 100 scFv les plus pertinents pour une exploration approfondie.