Proposition TER M1 Informatique
Algorithmes de Reconstruction de Séquences de Peptides

Encadrants : Guillaume FERTIN, Géraldine JEAN, Emile BENOIST
Équipe ComBi (Combinatoire et Bio-Informatique)
LS2N, Faculté des Sciences et Techniques de Nantes Université
E-mail: {guillaume.fertin, geraldine.jean, emile.benoist}@univ-nantes.fr

Mots-Clés: Séquences, Algorithmes, Fusion, Acides Aminés, Peptides.

Nombre d'étudiants: 1 ou 2.

Description du sujet. Suite à des travaux réalisés dans l'équipe ComBi et à l'INRAe de Nantes sur la spectrométrie de masse[1], nous cherchons à résoudre un problème de reconstruction de séquences de peptides[2] à partir d'informations partielles et parfois erronées, fournies en entrée sous la forme de séquences qu'on appelle *baitModels*.

Les *baitModels* sont des séquences composées de trois types d'éléments: (1) des caractères, (2) des valeurs numériques entre crochets et (3) des caractères entre crochets.
Par exemple, les 4 séquences suivantes sont des *baitModels*[3]:

- IV[251,10]IVEEDR
- IVHNI[357,15]DR
- IVH[114,04]IVEE[76,99]VI et
- IVHN[212,15]EEDR

Les valeurs numériques entre crochets représentent des masses, les caractères représentent des acides aminés[4]. Les crochets indiquent qu'il y a eu une modification (insertion, suppression ou substitution) d'un ou plusieurs acide(s) aminé(s) (chaque acide aminé étant vu soit comme un caractère soit comme une masse).

L'exemple ci-dessus montre quatre *baitModels* pouvant représenter une même séquence S d'acides aminés (la séquence S s'appelle un peptide ; ici, S=IVHNIVEEDR), potentiellement à quelques erreurs près. On aimerait, sur la base de ces quatre *baitModels*, être capable de reconstruire le peptide S. Pour cela, on cherche à fusionner les *baitModels*, en tirant parti des informations qu'ils portent, mais aussi possiblement en décidant d'en ignorer certaines parties, considérées alors comme fausses.

Dans ce TER, il s'agira de concevoir et d'implémenter un ou plusieurs algorithme(s) de fusion de *baitModels*, qui, à partir d'un ensemble de *baitModels* censés représenter le même peptide, permet(tent) de reconstruire la séquence de ce peptide.
On analysera les avantages et inconvénients de cet ou ces algorithme(s), et on le(s) testera sur un jeu de données issues du protéome humain (ce jeu de données sera fourni). Ainsi, les résultats pourront être comparés et analysés.

Compétences requises. Même si ce sujet fait suite à des travaux sur la spectrométrie de masse, *il n'est pas nécessaire d'avoir des compétences en biologie* pour pouvoir l'aborder. Il sera en revanche préférable d'avoir une certaine appétence pour l'algorithmique des séquences.

Précision. Tout.e candidat.e intéressé.e doit impérativement prendre contact avec les encadrants, qui procèderont à une sélection.

---

[1] voir `https://doi.org/10.1101/2022.05.31.494131`

[2] un peptide peut être vu comme une suite de caractères, chaque caractère représentant un acide aminé

[3] pas de caractères entre crochets dans ces exemples, mais ça peut exister.
Exemple: G[746,31]M[T]T[-101,05][V]A[59,00]DFFQGTK

[4] Il existe 20 acides aminés, chacun ayant une masse connue