# Proposition Stage M2 Informatique
# Reconstruction de Séquences et Analyse du Graphe des Peptides

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

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

**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 deux types de problèmes complémentaires, brièvement décrits ci-dessous.

**1. Fusion de séquences *baitModels*.** Le premier problème consiste à reconstruire des 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 ; IVHN[212,15]EEDR.

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

L'exemple ci-dessus montre quatre *baitModels* représentant une même séquence $S$ d'acides aminés (la séquence $S$ s'appelle un peptide), 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, alors considérées comme fausses.

**2. Extraction d'informations du Graphe des Peptides.** Nous avons en notre possession un jeu de données (tiré du protéome humain), pour lequel on a plusieurs dizaines de milliers de peptides, que l'on est capable de comparer deux à deux. A chaque comparaison de deux peptides $p_1$ et $p_2$, on associe un score $s_{p_1,p_2}$. Si ce score dépasse un certain seuil, alors $p_1$ et $p_2$ sont suffisamment proches pour que cette information soit utile. On peut alors construire un graphe $G$ (appelé "graphe des peptides"), où chaque sommet est un peptide, et deux sommets $p_1$ et $p_2$ sont reliés par une arête si $s_{p_1,p_2}$ est supérieur ou égal au seuil ; l'arête $(p_1, p_2)$ est alors pondérée par $s_{p_1,p_2}$.

Le graphe $G$ pourrait contenir plus d'informations, comme par exemple les *baitModels*, ou même le résultat de la fusion des *baitModels*, comme évoqué ci-dessus. L'ajout de ces informations dans $G$ devrait pouvoir nous permettre d'en savoir plus sur la façon dont les séquences des peptides changent quand on passe d'un sommet à l'autre, en suivant une ou plusieurs arêtes. Cela peut par exemple être

[1] voir deux articles récents de notre équipe: `https://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-021-03963-6` et `https://doi.org/10.1101/2022.05.31.494131`

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

[3] pas de lettres 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

fait (1) soit en propageant l'information le long d'un chemin (dans $G$) entre deux peptides, (2) soit en analysant des sous-graphes de $G$ très fortement connectés (et donc sans doute porteurs de "signal").

Dans ce stage, il s'agira de traiter les deux problématiques évoquées ci-dessus.

En ce qui concerne la fusion des *baitModels*, on concevra et implémentera 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. Ainsi, les résultats pourront être comparés et analysés.

En ce qui concerne le graphe des peptides $G$, il s'agit d'une partie plus exploratoire. D'abord, on réalisera "l'enrichissement" de $G$ en y ajoutant les informations évoquées ci-avant, et qui n'y sont pas encore. On procèdera ensuite à une analyse de $G$, et en particulier on identifiera les questions précises qu'il est intéressant de se poser concernant ce graphe. Enfin, on cherchera à répondre à ces questions, et à évaluer la pertinence des informations que ces réponses nous fournissent.

**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 appétence pour l'algorithmique des séquences et des graphes.