Debrief du vendredi 13/01

1 Récapitulatif de la journée

G. Fertin, en présence de E. Benoist, a présenté les travaux qui ont été fait en amont du TER par A. Lysiaq [1].

Pour l'instant, deux stratégies pour attaquer le problème sont à considérer :

- la première, proposée par G. Fertin, consiste à balayer les baitModels de gauche à droite et, quand un acide aminé est bien placé (c'est-à-dire que l'on a ni une lettre entre crochets ni une masse entre crochets), en sélectionnant l'acide aminé le plus fréquent parmis les acides aminés de même position.
- la seconde, que j'ai proposée, consiste à ignorer les masses entre crochets des baitModels et appliquer un algorithme d'alignement de séquences (e.g. Needleman-Wunsch, Smith-Waterman) pour ensuite déterminer si dans les sections de décalage d'un baitModel on retrouve bien la même masse que dans un autre baitModel.

Considérons un exemple où le peptide est IVHNIVEEDR avec 4 baitModels potentiellement représentatifs de cette séquence :

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

1.1 Première méthode

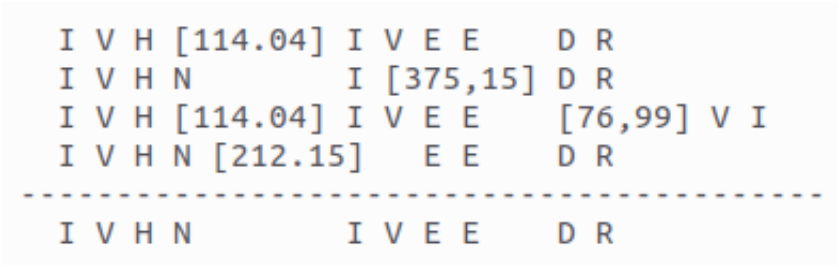
Sur cet exemple, la première méthode peut être illustrée comme suit :

```
  I V [251.10]   I V E E    D R
  I V H N        I [375,15] D R
  I V H [114.04] I V E E    [76,99] V I
  I V H N [212.15]   E E    D R
-----------------------------------------
  I V H N        I V E E    D R
```

En balayant les baitModels de gauche à droite, nous sélectionnons I comme étant le premier acide aminé de la séquence (les 4 baitModels ont I en première position), ensuite V est sélectionné pour la même raison, H n'est présent que dans 3 baitModels (mais c'est tout de même la majorité donc H peut être sélectionné). Les masses sont les éléments constituant une difficulté pour cette méthode. En effet, pour l'élément en 4e position, nous trouvons N dans deux baitModels mais aussi 2 masses indéterminées. Une option pour enlever les ambiguïtés serait, par exemple, de soustraire la masse de H de la masse entre crochets du

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

premier baitModels lorsqu'on sélectionne H. Dans le premier baitModel, après sélection de H, on aurait 251.10 - m(H) = 114.04 Da la masse entre crochet.
Voici l'étape intermédiaire :



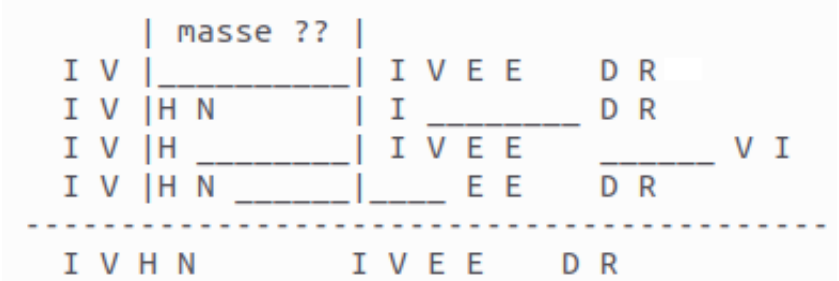
On peut aussi simplifié les masses au fur et à mesure notamment en remarquant que 114.04 = m(N) donc N est bien le quatrième acide aminé de la séquence. Pour le 5 élément de la séquence on sélectionne I et on soustrait la masse de I de la masse entre crochets du quatrième baitModel (il reste donc 212.15 - m(I) = 99.06 Da). Pour le 6 élément on trouve V dans deux baitModels, une masse égale à 99.06 Da dans le quatrième baitModel et une masse de 375.15 Da dans le premier baitModel. On remarque ici que m(V) = 99.06 Da donc on peut simplifié la masse du quatrième baitModel et placé V. Enfin, on soustrait m(V) à 375.15 et on trouve 276.08 Da.

Bien que cette méthode soit pertinente on peut remarquer deux choses : la masse de 76.99 Da présente avant VI dans le quatrième baitModel ne correspond à aucun acide aminé ; après simplification de la masse du deuxième baitModel il restera une masse résiduelle de 17.99 Da. De plus, sur quelques exemples cette méthode ne fonctionnera pas parce que si tous les acides aminés ont des masses entre crochets à la même position il est impossible de déterminer quel acide aminé il faut placé. Par exemple, pour la séquence GTFQIVYK avec deux baitModels :

- G[89.09]YVI[287.09]K
- G[376.17]IVYK

La masse de 89.09 Da dans le premier baitModel ne correspond à aucun acide aminé et il n'y a aucune information dans le second baitModel pouvant débloquer la situation. Cependant, retenons que cette méthode semble simple et efficace pour les instances "faciles".

1.2 Deuxième méthode



La deuxième méthode consisterait à aligner les séquences sans prendre en compte les masses. Ensuite, il s'agirait de choisir les acides aminés de sorte que la masse d'un décalage dans un baitModel corresponde à la masse présente dans une autre baitModel. Ici, la masse entre V et I pour le premier baitModel devrait être au moins égale à m(H). Le dossier "SOTA" (pour "State Of The Art") contient quelques articles abordant le problème d'alignement de séquences. La plupart des algorithme font appel à des notions de programmation dynamique. Le dossier CombOpt contient trois articles où les auteurs utilisent plutôt des notions d'optimisations combinatoires et de graphes pour résoudre le problème d'alignement de séquences.

Une étude plus approfondie de l'état de l'art doit encore être faite pour déterminer si cette méthode est réalisable. De plus, cette méthode serait beaucoup plus compliqué en terme de complexité.

2 Quelques remarques

2.1 Fichier de données

Dans le répertoire "instances/csv", un fichier de données au format csv avec quatre colonnes. La colonne 1 et 2 sont des peptides : le bait et le hit associés par le logiciel SpecOMS car ils ont un nombre de pics en commun (SPC) supérieur ou égal au seuil (ici 7). La colonne 3 est le hitModified renvoyé par le logiciel SpecGlob. Enfin, la colonne 4 est le baitModel qui est trouvé grâce à un prétraitement du hitModified. Chaque bait a un ou plusieurs hit(s). Le fichier contient un échantillon de 10 000 baits.

Étant donné que seuls les baitModels sont utilisés dans nos méthodes (le bait peut éventuellement être utilisé pour confirmer si la fusion des baitModels donne un bon résultat), j'ai implémenté un script pour convertir le fichier csv dans un autre format de fichier pour faciliter le chargement des données. Ce nouveau fichier ("instances/bait10000.txt") est un fichier texte tel que :

- la première ligne est le nombre de baits
- pour chaque bait, on a :
  - le bait (la séquence à représenter)
  - le nombre n de baitModels associés au bait
  - les n prochaines lignes sont les baitModels associés au bait

2.2 Table des masses

Dans le dossier "mass tables" se trouvent deux scripts (une version en Python et une autre en Julia mais pouvant faire usage de plusieurs threads) permettant de générer une table

où la première colonne correspond à une masse en Dalton et les colonnes suivantes correspondent aux séquences d'acides aminés (composés de 8 acides aminés maximum) ayant cette masse. Ainsi, en cherchant une masse de 114.04 Da dans la première colonne, on trouvera parmis les colonnes suivantes (sur la même ligne) l'acide aminé N. Ou encore, en cherchant la masse 212.15 on trouvera la séquence HN car m(H) + m(N) = 212.15. On stocke systématiquement les séquences dans l'ordre lexicographique sinon le nombre de colonnes dans la table serait beaucoup plus important. Cette table est nommée "mass_table_8" et est au format csv. D'autres versions considérant des séquences de plus petites tailles (de 3 à 5 acides aminés maximum) sont aussi fournies.

2.3 Méthodes de simplification

En analysant les baitModels fournies dans le fichier de données, je pense avoir trouver une méthode permettant de simplifier quelques baitModels en utilisant la table mentionnée ci-dessus. La méthode viendrait compléter le pré-traitement et permettrait d'éliminer des masses entre crochets.

Considérons quelques exemples. Soit la séquence SHSIEAPGK avec un seul baitModel [224.09]SIEAPGK. Grâce à la table "mass_table_3", on trouve que la masse de 224.09 Da correspond à la séquence HS. On sait donc que H et S sont potentiellement dans la séquence à trouver. Cependant, nous ne savons pas dans quel ordre les placer donc on propose deux solutions : HSSIEAPGK et SHSIEAPGK. Jusque là, l'utilisation de la table est triviale.

Bien évidemment, si la masse entre crochets est grande il est possible qu'elle corresponde à une séquence assez longue et, pire encore, plusieurs séquences peuvent avoir la même masse. De plus, si l'on suppose que les séquences sont constituées au plus de 8 acides aminés alors pour chaque séquence il faudra renvoyer au plus 8! solutions par séquence. Par exemple, la séquence GHIQSVTAPMGITMK avec un baitModel GHI[771.36]ITMK est impossible à simplifier en utilisant la table "mass_table_3". Parcontre, lors de la recherche d'une masse, on peut prendre en compte un élément déjà présent dans le baitModel pour élaguer certaines séquences dans la table. Ainsi, au lieu de considérer la masse de 771.36 Da pour effectuer la recherche, on pourrait chercher 771.36 + m(I) et ne prendre en compte que les séquences possédant au moins un acide aminé I.

Enfin, la simplification la plus intéressante est la suivante. Considérons la séquence GTFQIVYK avec deux baitModels G[376.17]IVYK et G[89.09]YVI[287.09]K. C'est sur ce même genre d'exemple que la première méthode proposée par G. Fertin est difficile à appliquée (voir plus haut). On remarque que la masse 89.09 ne correspond à aucune séquence d'acides aminés dans la table. On va alors considérer la somme de cette masse avec la masse suivante. Ainsi, au lieu de chercher une séquence de masse 89.09 Da, on cherchera une masse de 376,18 Da en considérant une marge d'erreur de 0.01 Da. On trouve alors dans la table les séquences FQT et NVY de masse 376.17. La séquence trouvée devra alors être placée avant la séquence entre crochets quand cela est possible et la séquence entre crochets devra aussi être inversée (donc on aura IVY au lieu de YVI). On renvoie toutes les solutions possibles (ici on a deux séquences de 3 acides aminés donc il y aura $2 * 3!$ solutions, c'est-à-dire 12) :

- GNVYIVYK

- GNYVIVYK
- GVNYIVYK
- ...
- GYVNIVYK
- GFQTIVYK
- GFTQIVYK
- ...
- GQFTIVYK

Parmis les 12 solutions proposées, 6 ont des similarités avec le bait (les solutions avec F, Q et T) et une solution correspond exactement au bait. Le but est de comparer ces baitModels "simplifiés" avec les baitsModels présents au départ en utilisant la méthode de G. Fertin. Malheuresement, sur cet exemple il reste difficile de départager les acides aminés à partir du second baitModel donc on serait contraint de renvoyer les 12 solutions trouvées.

L'inconvénient de ces simplifications sont :

- l'imprécision des masses : il faut souvent prévoir une marge d'erreur lorsque l'on vérifie si une masse est bien dans la table ou non. De plus, il n'est pas impossible d'avoir des décalages de masses assez important correspondant à du bruit
- plus on considère les tables complexes (e.g. "mass_table_8") plus on aura de séquences possibles à considérer d'où l'utilité d'inclure la séquence entre crochets dans la masse à rechercher pour élaguer les séquences. L'incovénient de cet élagage est que les acides aminés entre crochets ne seront pas forcément retrouvés dans le bait (e.g. PIPFPVIAPFSNPEHSAPAK)
- les masses précalculées correspondent à des séquences d'au plus 8 acides aminés dans le cas de "mass_table_8" donc si les masses entre crochets sont trop grandes ou correspondent à une séquence de plus de 8 acides aminés alors on obtiendra des solutions moins similaires au bait (e.g. GHIQSVTAPMGITMK)
- lorsque la dernière méthode de simplification est utilisée, il est possible que la masse corresponde à une séquence contenant R ou K alors que la séquence trouvée devrait être placée au début ou au milieu du baitModel (e.g. FSMPGFK, HSSVGSVIAK). Il faudrait alors prendre un choix entre éliminer les acides aminés gênants (ici R ou K) ou les placer à la fin du baitModel (exemples ci-après).

Soit la séquence FSMPGFK et le baitModel [146.11]FGPMSF[-18.02]. Ni 146.11 ni -18.02 correspondent à des masses dans la table. On somme alors les deux masses et on obtient une masse de 128.09 Da. Or, 128.09 = m(K). On ne peut pas placer K au début donc on le place à la fin et on oublie pas d'inverser la séquence entre crochets. On obtient FSMPGFK.

Soit la séquence HSSVGSVIAK et le baitModel H[9.05]AIV[408.14]K. On applique la dernière méthode de simplification parce que 9.05 ne correspond à aucune masse de séquence d'acides aminés. On trouve 417.19 Da la masse à rechercher. Or 417.19 = m(RSSS). Il est impossible, a priori, d'avoir R après H. On continue donc à chercher (avec une marge d'erreur de 0.01 Da) et on trouve 417.18 = m(GSSSV) en utilisant "mass_table_5". Le problème c'est que l'on trouve 7 autres séquences de même masse, élevant le nombre de solutions à 540. Parmis ces solutions, on trouve HSSVGSVIAK et au moins 120 solutions similaires au bait.