Travail de fin d'études

Analyse protéomique du repas sanguin de phlébotomes pour identification des hôtes

Marchand Lucien

Bachelier en Biotechnique option Bio-informatique

HEH – Département des Sciences et technologies

HEPH – Condorcet

Année académique 2023-2024

Travail de fin d'études

Analyse protéomique du repas sanguin de phlébotomes pour identification des hôtes

Marchand Lucien

Bachelier en Biotechnique option Bio-informatique
HEH – Département des Sciences et technologies
HEPH – Condorcet
Année académique 2023-2024

Remerciements

J'aimerais remercier le corps enseignant de la HEH, en particulier Monsieur Coornaert et Madame Léonet pour leur accompagnement durant ces mois de stage et TFE. Merci également à Sabrina, Sofia, Redouane et Esra pour m'avoir accompagné et aidé tout au long de ce travail.

Merci à tous mes amis qui ont été là dans les bons ou dans les mauvais moments, les Biogosses, les JSiens, les camarades de classes et les autres rencontrés par hasard mais devenus inoubliables.

Enfin merci à mes parents et à ma famille pour leurs encouragements et leur patience.

Merci à Loïc qui m'a aidé à trouver mon chemin et m'a accompagné jusque là.

Résumé

Ce travail de fin d'études repose sur l'identification par une aproche protéomique des organismes hôtes sur lesquels les phlébotomes, aussi appelés mouches des sables, peuvent se nourrir. Ces diptères hématophages sont impliqués dans le cycle de développement de protistes responsables de la Leishmaniose dans plusieurs régions du monde. Comprendre les hôtes préférentiels permettra non seulement de contrôler les foyer infectueux mais également de trouver des méthodes de lutte à long terme contre la maladie.

L'objectif principal de ce travail est l'élaboration d'une base de données de protéines sanguines permettant l'identification des protéines contenues dans l'intestin des phlébotomes via le séquençage par spectrométrie de masse. L'utilisation d'une banque assez petite et contenant uniquement les protéines d'intérêt permet de réduire le temps d'analyse des données de séquençage et de d'augmenter la sensibilité de l'analyse permettant l'identification de peptides dont le spectre est de faible intensité.

Une base de données basée sur la banque Swissprot a été créée et digérée *in silico* afin de trouver les protéines pouvant servir de marqueur pour différencier les espèces entre-elles.

Ensuite, de nombreuses analyses en shutgun ont permis de révéler les peptides les plus abondamment identifiés par spectrométrie de masse, et de les comparer aux résultats théoriques afin d'identifier les protéines les plus pertinantes à cibler lors de l'analyse de données par spectrométrie de masse.

Abstract

This thesis focuses on the identification of host organisms on which phlebotomines, also known as sand flies, may feed, using a proteomic approach. These hematophagous dipterans are involved in the development cycle of protists responsible for Leishmaniasis in various regions of the world. Understanding their preferred hosts will not only help control infectious hotspots but also aid in developing long-term strategies to combat the disease.

The primary objective of this work is to develop a blood protein database that enables the identification of proteins within the intestines of phlebotomines through mass spectrometry sequencing. Utilizing a small database that contains only the proteins of interest reduces the time required for data analysis and increases the sensitivity of the analysis, allowing the identification of peptides with low-intensity spectra.

A database based on the SwissProt repository was created and digested in silico to find proteins that could serve as markers to differentiate species. Subsequently, numerous shotgun analyses were conducted to reveal the most abundantly identified peptides by mass spectrometry, which were then compared to theoretical results to identify the most relevant proteins to target during data analysis by mass spectrometry.

Table des matières

Table des illustrations

Table des tableaux

Lexique

- Ancien Monde : Europe, Afrique et Asie.
- Nouveau Monde : Nom donné aux Amériques.
- Heteroxène : Désigne un parasite possédant deux hôtes lors de son cycle.
- Ergostérol : Lipide membrannaire (stérol) présent dans les membrannes de fungis majoritairement (et certains autres organismes).
- Parseur : programme qui analyse des données dans un format et les restructure pour un autre format.

1. Introduction

Figure 1 : Photographie de Rhodnius prolixus [1].

Ce travail de fin d'études est la suite de mon stage d'insertion professionnelle que j'ai réalisé dans le laboratoire de Biologie des vecteurs de parasites de l'ULB, au sein de l'Institut de Biologie et de Médecine Moléculaire (IBMM) à Gosselies, sous la supervision du Professeur Sabrina Bousbata. Le laboratoire étudie principalement l'interaction entre les vecteurs de pathogènes et leurs hôtes du point de vue protéomique, et notamment sur l'insecte Rhodnius prolixus, vecteur de la Maladie de Chagas. Le sujet traité lors de ce TFE est issu de la collaboration entre le laboratoire de Biologie des vecteurs de pathogènes et de l'Institut Pasteur au Maroc afin d'étudier la biologie des phlébotomes, vecteurs de la leishmaniose, pour élaborer des statégies de contrôle des deux vecteurs dominants au Maroc : *Phlebotomus papatasi* et *Phlebotomus sergenti*.

1.1. Introduction à la leishmaniose

La leishmaniose est une maladie parasitaire due à des protozoaires intracellulaires du genre *Leishmania* et transmise à l'homme par plus de 90 espèces de phlébotomes [2], principalement du genre *Phlebotomus* dans l'Ancien Monde et du genre *Lutzomyia* dans le Nouveau Monde [3]. La leishmaniose se présente principalement sous quatre formes : la leishmaniose vicérale (LV ou kala-azar) ; la leishmaniose dermique post-kala-azar (LDPK) ; la leishmaniose cutanée (LC) ; et la leishmaniose cutanéomuqueuse [5].

A) Epidémiologie

Elle est répendue mondialement dans près de 90 pays, réparis pripalement en Asie, Afrique, dans les Amériques et dans le bassin Méditerranéen (Figure 2, Figure 3)[3][5]. On estime le nombre de personnes infectées entre 12 et 15 millions, avec entre 1,5 et 2 millions de nouveaux cas chaques années, plus de 70.000 morts par an et 350 millions de personnes exposés au risque de contracter la maladie [3,4].

Certains facteurs augmentent le risque de contamination, tels que la pauvreté, la malnutrition, le manque d'hygiène, les migrations de populations ou l'immunodépression [5].

B) Diversité clinique

Les leishmanioses peuvent se présenter sous plusieurs formes [2] (Figure 4) :

- La leishmaniose cutanée : Forme la plus commune, elle entraîne des lésions cutanées, notamment des ulcères, pouvant laisser des cicatrices à vie. Ces cicatrices peuvent entrainer une stigmatisation des personnes concernées.
- La leishmaniose cutanéomuqueuse : Conduit à une destruction partielle ou totale des muqueuses dans le nez, la bouche ou la gorge.
- La leishmaniose vicérale (kala-azar) : Forme la plus mortelle (95% des cas sans traitement), elle est caractérisée par des poussées de fièvre, une perte de poids ainsi qu'un grossissement du foie et de la rate.
- La leishmaniose dermique post-kala-azar : Elle apparait chez 5 à 10% des patients atteints de LV, majoritairement en Afrique de l'Est et dans le sous-continent indien. Elle se développe entre 6 et 1 an (ou plus) après la guérison apparente de la LV.

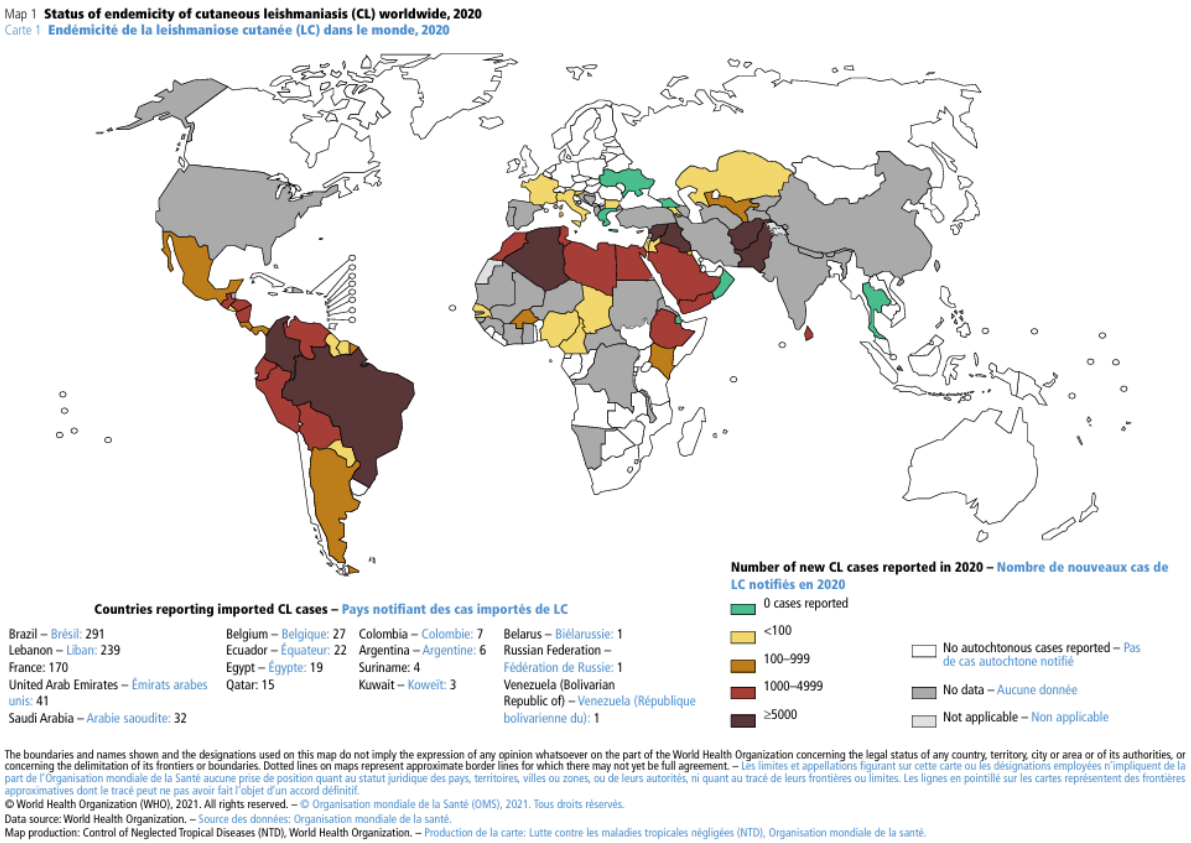


Figure 2 : Répartition au niveau mondial de la leishmaniose cutanée [5].

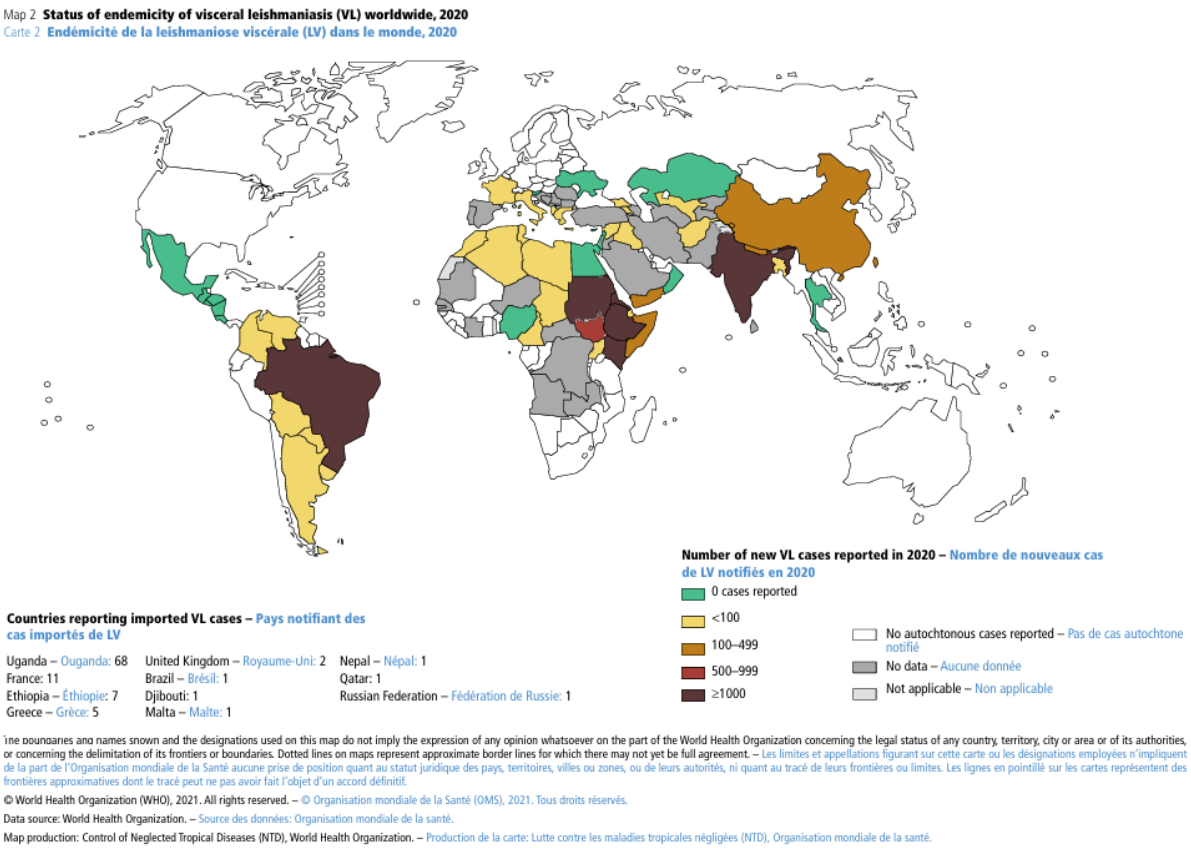


Figure 3 : Répartition de la leishmaniose viscérale au niveau mondial [5].

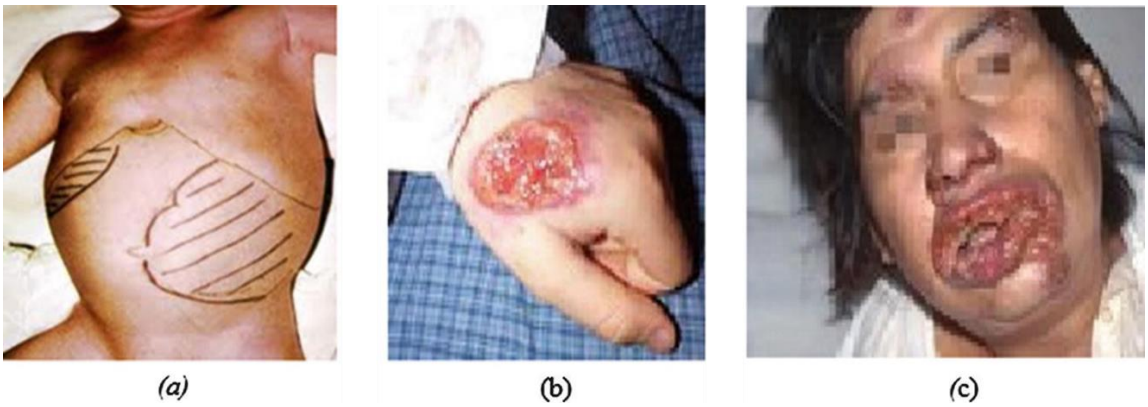


Figure 4 : Conséquences de la leishmaniose viscérale (a), cutanée (b) et cutanéo-muqueuse (c)[6].

C) Diagnostic

Le diagnostique de la leishmaniose vicérale est posé sur base d'un examen clinique associé à des tests parasitologiques ou sérologiques. On y retrouve :

- Un examen au microscope provenant d'une ponction tissulaire, permettant une grande spécificité, mais une sensibilité variable (93 à 99% pour une ponction de rate contre 53 à 65% pour une ponction de ganglion lymphatique)[4].
- La recherche d'ADN du parasite par PCR dans le sang ou dans la moelle osseuse, offrant une meilleure spécificité que le microscope, mais réservé pour l'instant aux hôpitaux et centres de recherche [4], en attente d'un équipement de pointe pour une utilisation plus étendue. La PCR révèle d'avantages d'infections asymptomatiques, mais peut donner des résultats positifs en cas de splénomégalie provoquée par une autre maladie.
- Un examen sérologique par la méthode ELISA (non-adapté pour le terrain) ou par le test immunochromatographique basé sur l'antogène rK39 (plus adapté pour le terrain), ce dernier étant rapide, peu coûteux et reproductible [4]. Cependant ces examens détectent des anticorps spécifiques plusieurs années après guérison, ne permettant pas le diagnostic d'une récidive avec certitude. Ils détectent également les anticorps présents chez des individus eposés mais asymptomatiques.

Les formes cutanées se détectent par les manifestations cliniques. Cependant elles sont semblables à d'autres infections comme celles causées par staphylocoques ou streptocoques, une mycose, … [4]. Elles doivent donc être confirmées par des tests parasitologiques [2] comme l'examen au microscope ou la PCR. Les diagnostics immunologiques sont limités en raison de leur faible sensibilité et d'une spécificité variable. Cependant, aucun test ne permet de distinguer une infection en cours d'une infection passée [5].

D) Traitements

De nombreux traitements existent, mais ont tous des effets secondaires plut ou moins importants. De plus, ils nécessitent souvent un traitement de soutient comme une réhydratation ou des suppléments alimentaires avant le début de la thérapie [4]. Ils nécessitent également une surveillance médicale et tenir compte du risque de pharmacorésistance. C'est pourquoi l'utilisation d'associations médicamenteuse semble la meilleure stratégie afin de lutter contre.

Dérivés de l'antimoine pentavalent

Deux molécules sont actuellement utilisées : l'antimoniate de méglumine et le stibogluconate de sodium. Ces deux molécules sont similaires au point de vue chimique, mais le stibogluconate de sodium est légèrement plus toxique en raison de sa concentration plus importante en animoine (10% contre 8,5% pour l'antimoniate de méglumine) [4]. Ils peuvent être

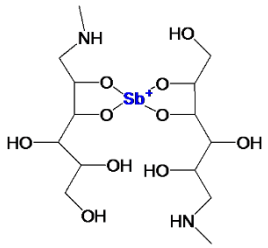


Figure 6 : Formule chimique de l'Antimoniate de méglumine

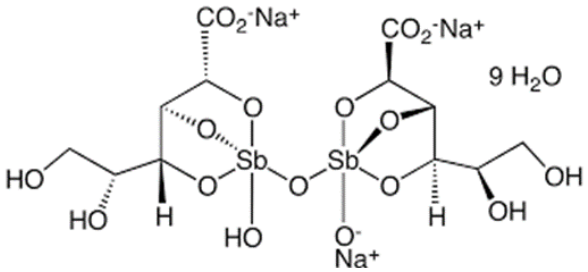


Figure 5 : Formule chimique du Stibogluconate de sodium [7].

injectés par intraveineuse ou intramusculaire, ou également par voie intralésionnelle dans le cas d'une leishmaniose cutanée. L'antimoine qu'ils contiennent pénètre dans les macrophages (où se trouvent les leishmanies). Une fois absorbé par ces dernières, il est réduit par leur métabolisme en antimoine trivalent qui va inhiber la topoisomérase I (réplication de l'ADN)[10].

Ils offrent un bon taux de guérison (>90%), mais la pharmacorésistance est déjà constatée dans plueisurs régions d'Asie.

Leurs effets secondaires sont principalement l'anorexie, des nausées et vomissements, des douleurs abdominales, un goût métallique ou encore la léthargie. Cependant ils peuvent également induire des effets secondaires graves comme la cardiotoxicité ou l'hépatotoxicité, mais ils sont rares [4].

Désoxycholate d'amphotéricine B

L'amphotéricine B est un antibiotique de la famille des polyènes. Il se lie à l'ergostérol présent dans les membranes cellulaires des fungis, que l'on retrouve également chez les leishmanies [11]. Cette interraciton rend la membrane perméable, conduisant à terme à la destruction de la cellule [12]. Ce traitement nécessite une surveillance permanante et doit donc être effectuer en milieu hospitalier, en raison d'une néphrotoxicité et d'un besoin d'une bonne idratation et d'un apport de potassium.

Ce traitement s'est révélé efficace à 99% en Inde mais les données dans les autres régions ne sont pas encore assez nombreuses [4].

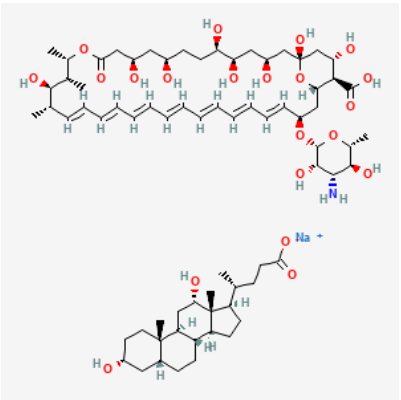


Figure 7 : Formule chimique du Désoxycholate d'amphotéricine B [9].

Formes galéniques lipidiques d'amphotéricine B

Ce traitement est d'une efficacité analogue au Désoxycholate d'amphotéricine B, mais sa toxicité est sensiblement moindre [4]. Une néphrotoxicité ou une thrombocytopénie transitoire peuvent se produire occasionnellement.

C'est le traitement de référence pour une leishmaniose vicérale dans le bassin méditerranéen. Son efficacité varie entre 90 et 98% en Europe du Sud et est suppérieure à 95% en Inde [4].

1.2. Les parasites du genre *Leishmania*

Le genre Leishmania regroupe 53 espèces reconnues de parasites heteroxènes[13]. Ce sont des parasites intracellulaires obligatoire, vivant dans les le tractus digestif de leur vecteur ou dans les cellules phagocytaires de leurs hôtes vertébrés.

Plus communément appelés leishmanies, ces organismes peuvent provoquer l'apparition d'une forme de leishmaniose, souvent dépendante de l'espèce [13].

A) Classification

Tableau 1 : Classification des parasites du genre Leishmania.

Domaine	Eukaryota
Règne	Protista
Phylum	Euglenozoa
Classe	Kinetoplastea
Ordre	Trypanosomatida
Famille	Trypanosomatidae
Genre	Leishmania
Espèce(s)	L. major, L. donovani, L. infantum, L.tropica, …

Les leishmanies sont des protistes de la famille des Trypanosomatidae, famille regroupant plusieurs parasites comme les trypanosomes responsables de la maladie du sommeil ou de la maladie de Chagas [14]. Il en existe 4 sous-genres et un total de 53 espèces décrites dont 31 sont connues comme parasites de mammifères et 20 sont pathogènes pour l'Homme [15].

B) Cycle

Les leishmanies ont deux phases distinctes dans leur cycle de vie : une phase promastigote, mobile, à l'intérieur du tractus digestif du vecteur, et une phase amastigote à l'intérieur de l'hôte vertébré [16]. Le cycle complet d'une leishmanie est décrit en Figure 10. Lorsqu'un phlébotome infecté se nourrit sur un hôte vertébré (**1**), les formes promastigotes vont être injectées dans la peau et phagocytées par les cellules mononucléées du système immunitaire de l'hôte (**2**), à l'intérieur desquelles les leishmanies vont se transormer en forme amasitgote (perte du flagelle) (**3**). Le parasite va alors se multiplier dans les cellules infectées, provoquant la lyse de ces dernières, lui permettant d'infecter d'autres tissus (**4**). Les amastigotes peuvent voyager à travers la circulation sanguine ou le système lymphatique, pouvant infecter d'autres organes (foie, rate) et causer une forme viscérale de la maladie. Ensuite, lorsqu'un autre phlébotome viendra se nourrir du sang de l'hôte (**5**), les macrophages infectés qui s'y trouvent vont se retrouver dans l'intestin de l'insecte (**6**), où les amastigotes vont se transformer en promastigotes (**7**). Ces promastigotes vont ensuite se reproduire par division cellulaire et migrer jusqu'au proposcis afin d'infecter de nouveaux hôtes lors d'un repas sanguin.

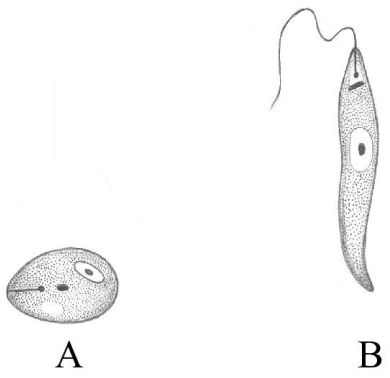


Figure 9 : Dessin des formes amastigotes (A) et promastigotes (B) des leishmanies [18].

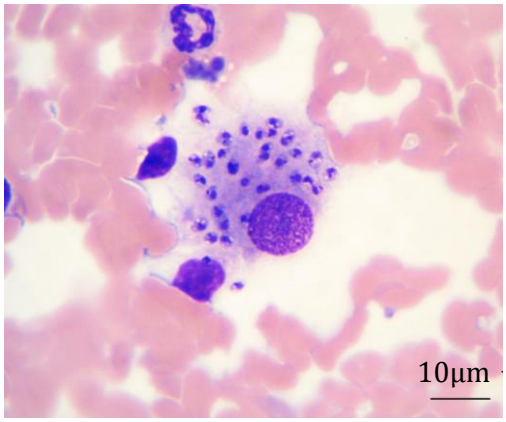


Figure 8 : Cliché microscopique de la forme amastigote des leishmanies dans un macrophage chez le chien [19].

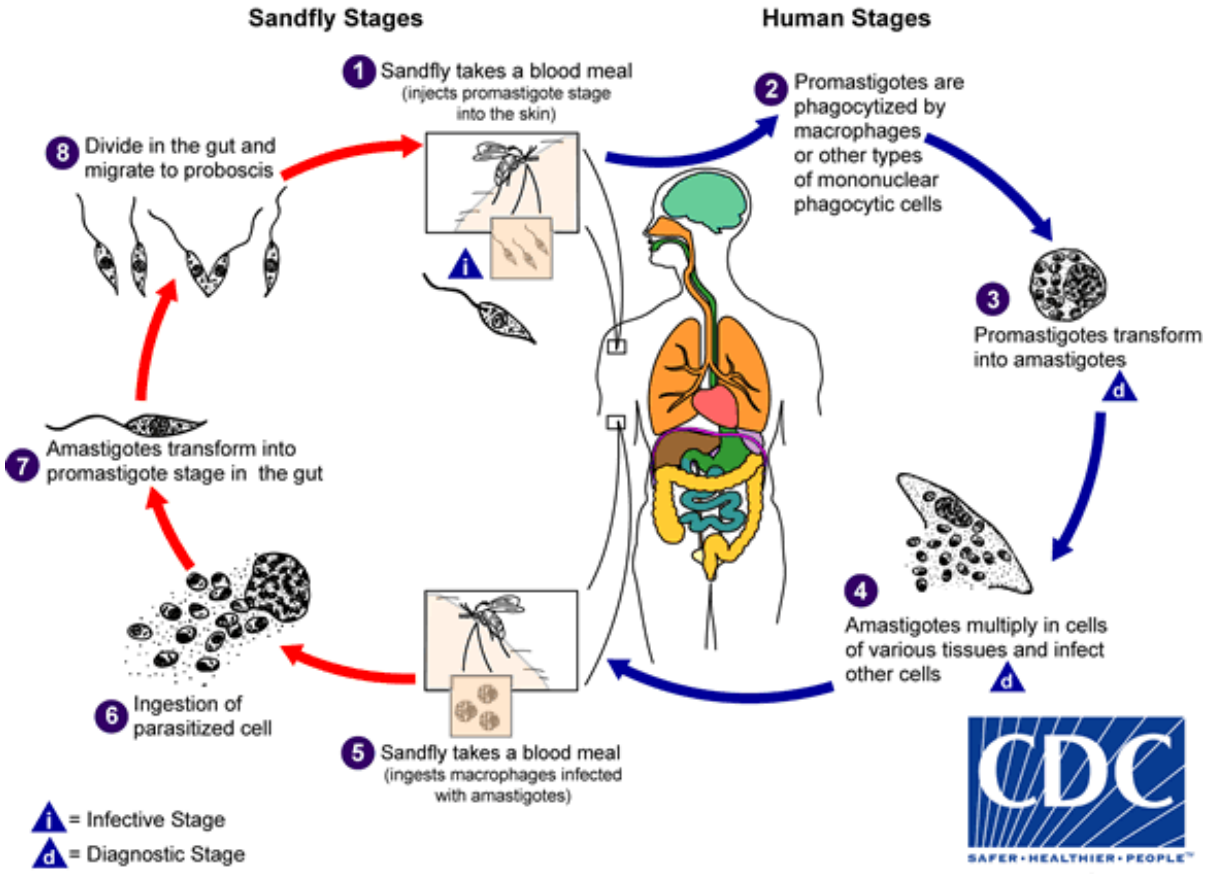


Figure 10 : Cycle des parasites du genre Leishmania [17].

1.3. Le phlébotome, vecteur de la maladie

Comme expliqué précédemment, la leishmaniose se transmet par l'intermédiaire d'un vecteur, le phlébotome. Également appelé « mouche des sables », ce nom désigne plus de 500 espèces de la famille des Phlebotomidae, regroupant 3 genres principaux : *Phlebotomus* et *Sergentomyia* dans l'Ancien Monde et *Lutzomyia* dans le Nouveau Monde [20].

A) Classification

Tableau 2 : Classification des espèces de phlébotome

Domaine	Eukaryota
Règne	Animalia
Phylum	Arthropoda
Classe	Insecta
Ordre	Diptera
Famille	Psychodidae
Sous-famille	Phlebotominae
Genre(s)	Phlebotomus, Sergentomyia, Lutzomyia
Espèce(s)	P. papatasi, P. sergenti, L. longipalpis, …

B) Cycle de vie

Les phlébotomes passent par 4 phases lors de leur développement : [21]

- Œuf
- Stade larvaire (comprenant lui-même 4 stades)
- Pupe
- Stade adulte

Les stades précédent le stade adulte ont besoin d'un environnement chaud et humide pour terminer leur développement. Les femelles pondent leurs œufs après la prise d'un repas sanguin. Les larves de premier stade se développent entre 12 et 19 jours et les pupes de 25 à 59 jours. Les adultes terminent leur développement après 35 à 69 jours. [21]

Les phlébotomes adultes mesurent environ 3,5mm de longueur. Ils possèdent un duvet dense et leurs ailes ont une forme de « V » caractéristique lorsqu'ils sont au repos, ainsi que des pattes lontues et fines [21]. Les mâles se différencient des femelles via la taille de leurs organes génitaux. Chez les mâles, ils sont proéminents et portent des épines sclérifiées, absentes chez les femelles (Figure 11 et Figure 12).

L'acouplement se fait généralement à proximité de l'hôte. Les mâles se rassemblent près de ce dernier en émettant des phéromones sexuelles, ainsi qu'en faisant vibrer leurs ailes pour encourager l'accouplement [21].

Figure 11 : Phlébotome adulte femelle (P. papatasi) [21].

Figure 12 : Phlébotome mâle sauvage (P. papatasi) [24].

C) Alimentation

Les phlébotomes adultes se nourissent principalement de nectar de plantes et de sécrétions sucrées qu'elles produisent [22], mais également de miellat provenant de pucerons. Cependant les femelles ont besoin d'au moins un repas sanguin pour terminer le développement de leurs œufs [21]. Les femelles sont donc les seules à se nourrir sur un hôte vertébré, induisant la possibilité de transmettre la leishmaniose.

Le stade de digestion d'un repas sanguin chez le phlébotome femelle engorgé est visible à l'œil nu, en regardant la forme et la couleur de l'abdoment. Un phlébotome qui s'est nourri récemment a un abdomen gonflé et rouge tandis qu'un état avancé de digestion sera visible par l'abdomen noir (Figure 14 et Figure 13).

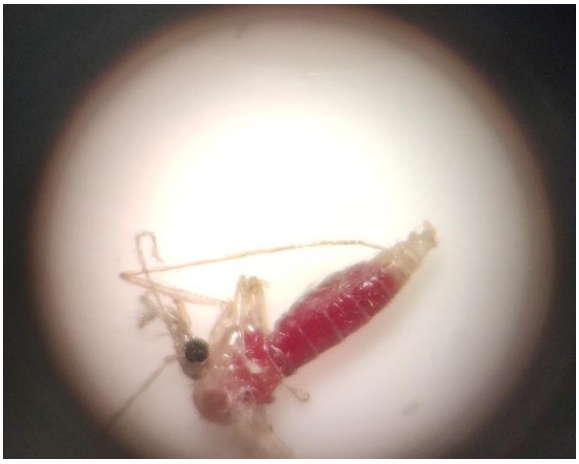


Figure 14 : Phlébotome femelle ayant eu un repas sanguin peu avant sa capture [24].

Figure 13 : Phlébotome femelle ayant un état de digestion avancé du repas sanguin [24].

D) Hôtes

Les hôtes vertébrés chez les phlébotomes sont nombreux et variés [22]. On y retrouve notamment les humains et les chiens pour qui les cas de leishmanioses sont connus et décrits, mais également des animaux de rente tels que les moutons et les vaches. On retrouve aussi dans leurs hôtes des rongeurs sauvages et domestiques, ainsi que des équidés et des oiseaux [23]. Bongiorno et al. ont montré que plus de 95 % des spécimens recueillis à l'intérieur d'une écurie, d'un chenil, d'une bergerie et d'un poulailler avaient consommé le sang des animaux hébergés dans ces abris respectifs. Leur étude conclut que les hôtes retrouvés dans le repas sanguin varient en fonction de la disponibilité et de l'abondance plutôt qu'en fonction de la spécificité à une espèce [23]. Les zones urbaines et périurbaines favorisent donc la transmission de parasites du genre Leishmania à l'homme ou au chien [23].

1.4. Les protéines sanguines

Le sang est un fluide riche en protéines. Il est constitué à 62.5% d'eau, à 34% d'hémoglobine, et à 2.5% d'autres protéines, parmis lesquelles on retrouve l'albumine à 50%, les immunogloulines G à environ 20%, de la trasferrine, du fibrinogène, des macroglobulines, … Près de 90% du contenu en protéines dans le sang est dû uniquement à 22 protéines [29].

2. Objectifs du travail

Identification des hôtes par MS/MS du repas sanguin

Ce travail s'inscrit dans un projet de recherche impliquant plusieurs parties visant à comprendre la biologie des phlébotomes. Il s'axe sur différents points :

- L'identification des hôtes par spectrométrie de masse permettant d'investiguer les habitudes alimentaires du vecteur et d'identifier les réservoirs du parasite
- La compréhension des mécanismes de digestion du repas sanguin afin d'identifier des protéines qui pourraient constituer une cible antivectorielle

Le but premier de ce travail est la création d'une base de données légère et fiable permettant l'identification de peptides issus du repas sanguin du phlébotome en utilisant la spectrométrie de masse en tandem (MS/MS). En effet, l'analyse de données via MS/MS peut être très longue en fonction de la taille des données de séquençage ainsi que la taille de la banque utilisée. Une banque plus petite permettra une analyse plus rapide.

Bien que les phlébotomes digèrent le sang dans leur intestin, il est possible de retrouver des peptides par spectrométrie de masse lorsque l'état de digestion n'est pas trop avancé. Une étude menée par Hlavackova et al. [25] a montré que la spectrométrie de masse identifie avec certitude les hôtes jusqu'à 36h après le repas sanguin, et cette identification reste correcte à 80% jusqu'à 48h après le repas sanguin.

Dans un second temps, ce travail mettra en lumière les différents peptides signature permettant l'identification d'espèces à partir du contenu en sang dans l'intestin des phlébotomes, et ce afin de pouvoir déterminer les hôtes majoritaires et de trouver des moyens de lutte contre la maladie provoquée par les protistes via ces hôtes cibles.

3. Matériel et méthodes

3.1. Origine des insectes

Les phlébotomes du terrain (Tableau 3) ont été récoltés dans la région de Taza, au Maroc, par le Dr. Sofia El Kacem du laboratoire du Dr. M. Lemrani de l'Institut Pasteur du Maroc de Casablanca et le laboratoire du professeur S. Boussaa des Instituts supérieurs des professions infermières et techniques de santé de Rabat.

Les phlébotomes ont été capturés à l'aide de pièges lumineux CDC (Center for Disease Control). Les pièges ont été placés la journée et laissés actifs durant la nuit, puis récoltés le lendemain matin.

Les insectes récoltés ont ensuite été observés au binoculaire pour décrire leurs caractéristiques (espèce, sexe, stade de digestion, etc…) puis conservés dans des tubes annotés à -20°C.

Figure 15 : Récolte des insectes capturés via le dispositif CDC [24].

Figure 16 : Observation, identification et caractérisation des insectes récoltés [24].

Afin de valider la base de données créée, des insectes d'élevage (Tableau 3) provenant de l'insectarium du département de parasitologie de l'Université Charles à Prague (en collaboration avec le Prof. Petr Volf), nourris spécifiquement sur une espèce d'hôte définie, ont également fait l'object de ce travail.

3.2. Préparation des échantillons

Les femelles engorgées sélectionnées ont été broyées individuellement dans 50 µL de tampon DLA (7M Urea, 2M Thiourea, 4% CHAPS, 30 mM Tris) à l'aide d'un pilon adapté. Les échantillons ont ensuite été soumis à trois cycles de sonication de 5 minutes chacun (Ultra Sonic

Cleaner VWR), avec des intervalles de 30 secondes sur glace entre chaque cycle. Après la sonication, les échantillons ont été centrifugés à 13 000 g pendant 15 minutes à 4 °C. Le surnageant, contenant les protéines solubilisées, a été récupéré dans un nouveau tube.

La concentration en protéines dans le surnageant a été déterminée par la méthode Pierce 660 nm Protein Assay (Thermo Scientific Inc., Rockford, IL, USA). Ensuite, les échantillons ont été analysés par électrophorèse sur gel SDS-PAGE pour évaluer le profil protéique de chaque extrait. Un total de 10 µg de protéines extraites a été digéré à la trypsine (Promega, Madison, WI, USA) selon le protocole FASP [28]. Les peptides obtenus ont été reconstitués dans 10 µL de solvant d'injection (ACN 2 % / FA 0,1 %) avant d'être soumis à un séquençage par spectrométrie de masse à l'UNamur.

Tableau 3 : Liste des échantillons séquencés par spectrométrie de masse

Nom de l'échantillon	Nombre d'insectes	Remarques	Date de séquençage
FG10	1	*P. sergenti* sauvage capturé en août 2015	23-01-2024
TG02	1	*P. sergenti* sauvage capturé en août 2015	23-01-2024
TG16	1	*P. sergenti* sauvage capturé en août 2015	23-01-2023
TG05	1	*P. sergenti* sauvage capturé en août 2015	22-04-2024
TG14	1	*P. sergenti* sauvage capturé en août 2015	22-04-2024
TG22	1	*P. sergenti* sauvage capturé en août 2015	22-04-2024
PA2A	1	*P. papatasi* d'élevage nourri avec du sang de rongeur (*Acomys cahirinus)*	22-03-2024
PA2B	1	*P. papatasi* d'élevage nourri avec du sang de rongeur (*Acomys cahirinus)*	22-03-2024
PA2C	1	*P. papatasi* d'élevage nourri avec du sang de rongeur (*Acomys cahirinus)*, digestion selon le protocole C18	22-03-2024
PA2D	1	*P. papatasi* d'élevage nourri avec du sang de rongeur (*Acomys cahirinus)*, digestion selon le protocole C18	22-03-2024
PH4A	2	*P. papatasi* d'élevage nourris avec du sang humain	13-05-2024
PO1B	2	*P. papatasi* d'élevage nourris avec du sang humain	13-05-2024
SO3A	2	*P. sergenti* d'élevage nourris avec du sang de mouton	13-05-2024
SH4	2	*P. sergenti* d'élevage nourris avec du sang de mouton	13-05-2024

3.3. Analyse par spectrométrie de masse

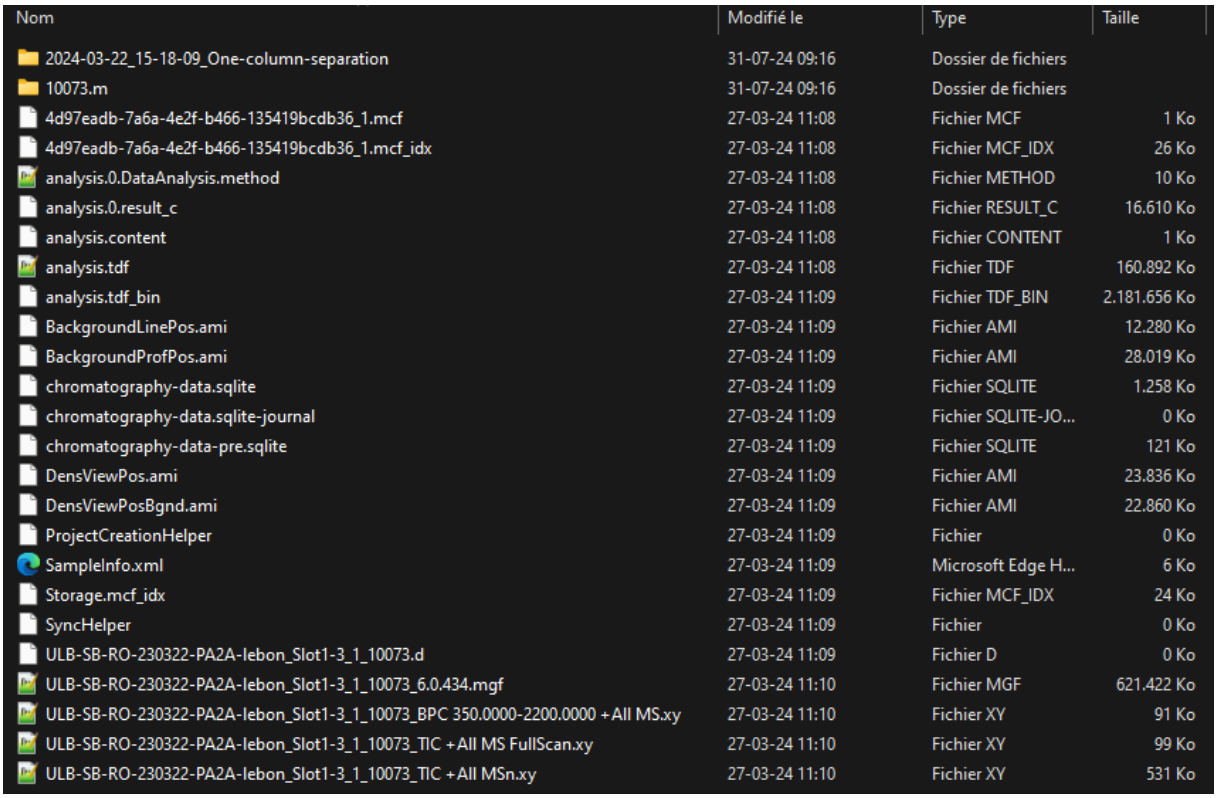
A) Paramètres de séquençage

Les peptides résultants de la digestion à la trypsine ont été analysés à l'aide d'un nano-LC-ESI-MS/MS timsTOF Pro (Bruker, Billerica, MA, USA) associé à un système de chromatographie liquide ultra-haute performance nanoElute (Bruker). Les spectres de masse en tandem ont été extraits, déconvultés pour les états de charge, et désisotopés à l'aide du logiciel Data Analysis version 5.3 (Bruker).

B) Informations sur les fichiers de spectrométrie de masse

Le but de ce travail étant l'identification par spectrométrie de masse des protéines sanguines des hôtes présentes dans l'intestin des insectes, nous avons d'abord récupéré les données brutes issues du spectromètre de masse. Ces données sont sous la forme de dossiers « .d » (Figure 17). La taille de ces dossiers varie selon le nombre de protéines séquencées, généralement entre 1Go et 3Go de données. Le contenu de ces dossiers est présenté à la Figure 17.

Le dossier « *_One-column-separation » contient les paramètres utilisés lors de la chromatographie liquide. Le dossier « *.m » contient le protocole les paramètres liés à l'analyseur timsTOF.



Nom	Modifié le	Type	Taille
2024-03-22_15-18-09_One-column-separation	31-07-24 09:16	Dossier de fichiers	
10073.m	31-07-24 09:16	Dossier de fichiers	
4d97eadb-7a6a-4e2f-b466-135419bcdb36_1.mcf	27-03-24 11:08	Fichier MCF	1 Ko
4d97eadb-7a6a-4e2f-b466-135419bcdb36_1.mcf_idx	27-03-24 11:08	Fichier MCF_IDX	26 Ko
analysis.0.DataAnalysis.method	27-03-24 11:08	Fichier METHOD	10 Ko
analysis.0.result_c	27-03-24 11:08	Fichier RESULT_C	16.610 Ko
analysis.content	27-03-24 11:08	Fichier CONTENT	1 Ko
analysis.tdf	27-03-24 11:08	Fichier TDF	160.892 Ko
analysis.tdf_bin	27-03-24 11:09	Fichier TDF_BIN	2.181.656 Ko
BackgroundLinePos.ami	27-03-24 11:09	Fichier AMI	12.280 Ko
BackgroundProfPos.ami	27-03-24 11:09	Fichier AMI	28.019 Ko
chromatography-data.sqlite	27-03-24 11:09	Fichier SQLITE	1.258 Ko
chromatography-data.sqlite-journal	27-03-24 11:09	Fichier SQLITE-JO...	0 Ko
chromatography-data-pre.sqlite	27-03-24 11:09	Fichier SQLITE	121 Ko
DensViewPos.ami	27-03-24 11:09	Fichier AMI	23.836 Ko
DensViewPosBgnd.ami	27-03-24 11:09	Fichier AMI	22.860 Ko
ProjectCreationHelper	27-03-24 11:09	Fichier	0 Ko
SampleInfo.xml	27-03-24 11:09	Microsoft Edge H...	6 Ko
Storage.mcf_idx	27-03-24 11:09	Fichier MCF_IDX	24 Ko
SyncHelper	27-03-24 11:09	Fichier	0 Ko
ULB-SB-RO-230322-PA2A-lebon_Slot1-3_1_10073.d	27-03-24 11:09	Fichier D	0 Ko
ULB-SB-RO-230322-PA2A-lebon_Slot1-3_1_10073_6.0.434.mgf	27-03-24 11:10	Fichier MGF	621.422 Ko
ULB-SB-RO-230322-PA2A-lebon_Slot1-3_1_10073_BPC 350.0000-2200.0000 +All MS.xy	27-03-24 11:10	Fichier XY	91 Ko
ULB-SB-RO-230322-PA2A-lebon_Slot1-3_1_10073_TIC +All MS FullScan.xy	27-03-24 11:10	Fichier XY	99 Ko
ULB-SB-RO-230322-PA2A-lebon_Slot1-3_1_10073_TIC +All MSn.xy	27-03-24 11:10	Fichier XY	531 Ko

Figure 17 : Structure d'un dossier ".d" issu des instruments Brucker timsTOF Pro.

La plupart des fichiers sont des fichiers cryptés contenant des métadonnées. Les plus gros fichiers sont les fichiers TDF et TDF_BIN. Le fichier TDF_BIN contient toutes les informations sur les spectres, les rapports m/z, etc… Cependant, ces données sont au format binaire et nécessitent un programme dédié pour les ouvrir. Le fichier TDF quant à lui est un fichier au format SQLite qui contient les métadonnées et les informations relatives à l'expérience, ainsi que des liens vers le fichier binaire des données brutes. On peut également trouver un fichier « SampleInfo.xml » qui contient les données utilisées lors du séquençage, ainsi que le fichier du protocole spécifique au laboratoire utilisé pour l'analyse.

Enfin, on trouve un fichier « .mgf », format utilisé notamment par le logiciel Mascot (un standard en analyses protéomiques mais sous liscence), et qui contient des informations sur les spectres, notamment le rapport masse sur charge du peptide séquencé, sa charge, sa mobilité et les ions générés après le passage dans la cellule de collision. C'est le deuxième plus gros fichier des dossiers « .d », avec une taille variant de 100ko pour un petit jeu de données à 600ko pour les plus grandes données de séquençage.

3.4. Programmes utilisés

A) EMBOSS

Les bases de données ont été créées et indexées en utilisant la suite de logiciels EMBOSS dans sa version 6.6.0.0 (https://emboss.sourceforge.net). Cette suite contient notamment les programmes :

- dbxflat : Indexation des bases de données ;
- seqret : récupération des séquences au format « fasta » ;
- entret : récupération des entrées d'une base de données (ici au format swiss) ;
- infoseq : montre les informations d'une séquence en fonction des champs utilisés dans l'indexation de la base de données ;
- pepdigest : digestion *in silico* de séquences protéiques.

Il contient également de nombreux programmes utilisés pour l'analyse de données biologiques (nucléotidiques ou protéiques).

B) Maxquant

Maxquant (https://www.maxquant.org) est un logicel open source d'analyse protéomique conçu pour l'analyse de grands ensembles de données issus de la spectrométrie de masse. Il est basé sur le moteur de recherche Andromeda, dont les résultats sont très proches de ceux du moteur de recherche Mascot [27]. Maxquant peut etre utilisé via une interface utilisateur ou a la ligne de commande, en introduisant les paramètres de charge via un fichier « mqpar.xml ». Les fichiers de sortie de maxquant sont des fichiers texte pouvant être interprétés comme des fichiers TSV (Tabulation-Separated Values), chacun ayant ses spécificités selon le type d'analyse de données souhaité. Le fichier le plus utilisé dans le cadre de ce travail est le fichier « peptide.txt ». C'est le logiciel utilisé par le laboratoire dans lequel ce travail a été réalisé. Les analyses ont été réalisées sur la version 2.5.0.0 de Maxquant.

Définition des paramètres

Au lancement, Maxquant ouvre une fenêtre d'interface graphique utilsateur permettant son utilisation simplifiée pour les personnes n'ayant pas l'habitude de travailler à la ligne de commande (Figure 18). On y trouve plusieurs onglets permettant de spécifier les paramètres nécessaires à l'analyse des données.

Les dossiers « .d » peuvent être chargés en utilisant le bouton « Load folder » de l'onglet « Données brutes » (Figure 18.1). Le format spécifique des données est automatiquement détecté par Maxquant (Figure 19). Pour effectuer une analyse comparative des différentes données, il faut ajouter un nom d'expérience (Figure 18.2,Figure 19).

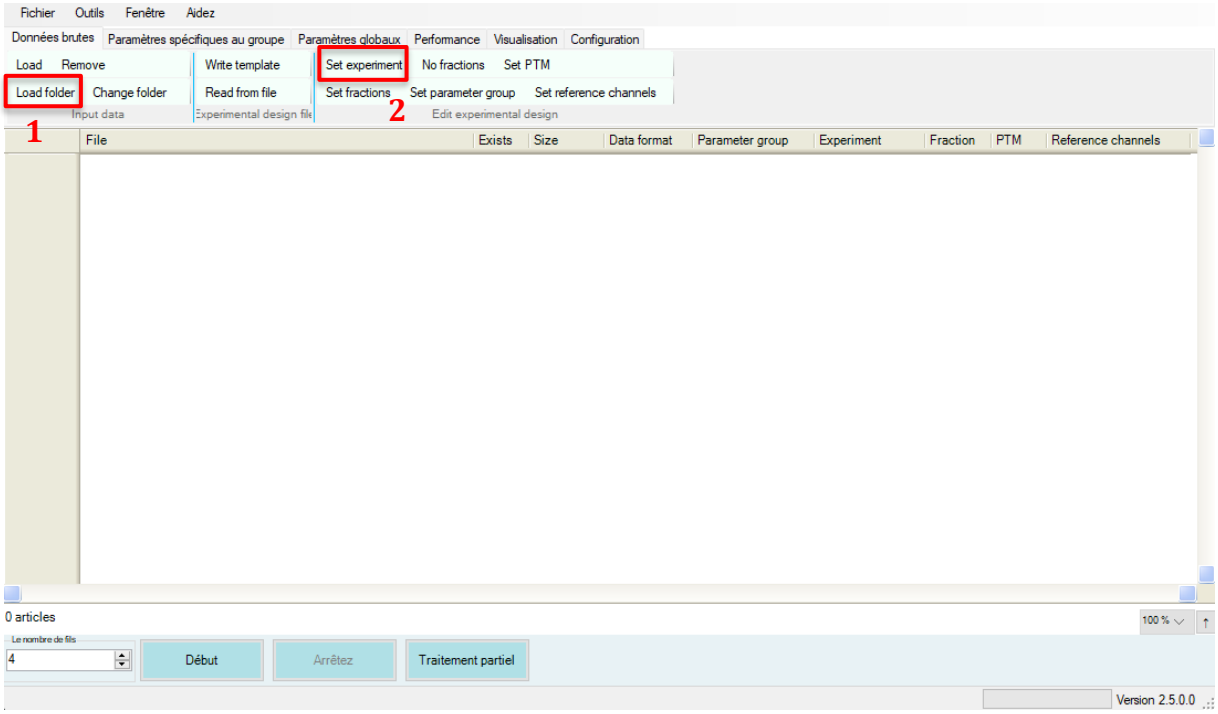


Figure 18 : Interface graphique utilisateur de Maxquant.

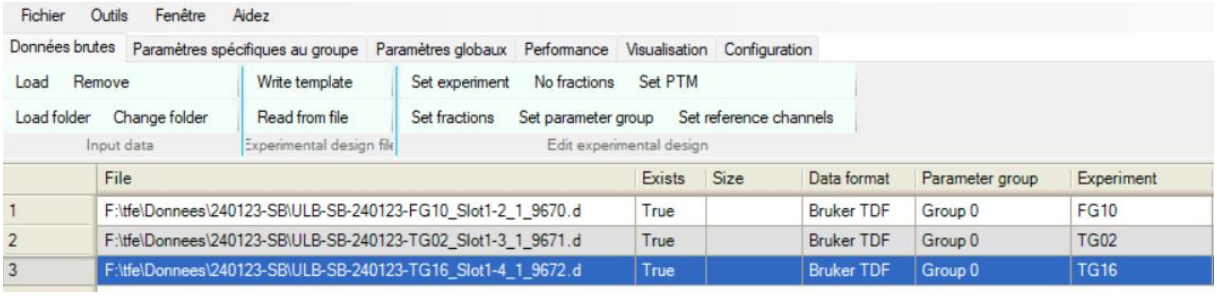


Figure 19 : Ajout des données et définition du nom de l'expérience.

L'onglet « Paramètres spécifiques au groupe » permet d'affiner la recherche en indiquant les modifications post-traductionnelles potentielles subies par les peptides séquencés (comme l'oxydation des méthionines ou l'acétylation du côté N-terminal du peptide)(Figure 20). Les modifications par défaut seront appliquées lors de ces analyses (les deux cités précédemment et la méthylation du côté C-terminal).

L'enzyme de restriction peut également être modifiée dans le sous-onglet « Digestion ». Par défaut, la trypsine est sélectionnée car elle est l'enzyme la plus couramment utilisée dans la digestion des protéines en spectométrie de masse. C'est également l'enzyme utilisée lors de ces expériences (voir Préparation des échantillons). Le sous-onglet « Instrument » permet de spécifier l'appareil utilisé pour l'aquisition des spectres, bien qu'il soit détecté automatiquement via les métadonnées contenues dans les dossiers « .d ». Enfin, le type d'analyse et la méthode de quantification peuvent être spécifiées, mais ne seront pas utilisées dans le cadre de ce travail.

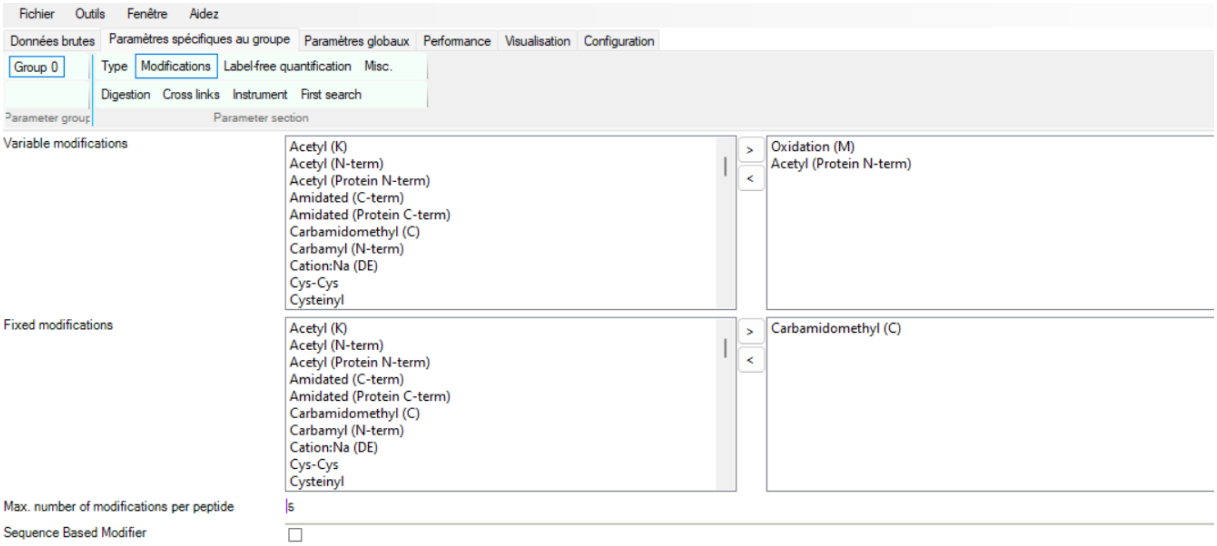


Figure 20 : Modification des paramètres de recherche.

L'onglet « Paramètres globaux » est utiliser pour fournir la base de données de référence utilisée pour la recherche des pics et la reconstruction de la séquence. Plusieurs bases de données peuvent être fournies. Les colonnes « Identifier rule » et « Description rule » utilisent des expressions régulières pour déterminer l'identifiant de la séquence et l'entête, et peuvent être modifiées si les lignes d'identifiants ont une structure différente.

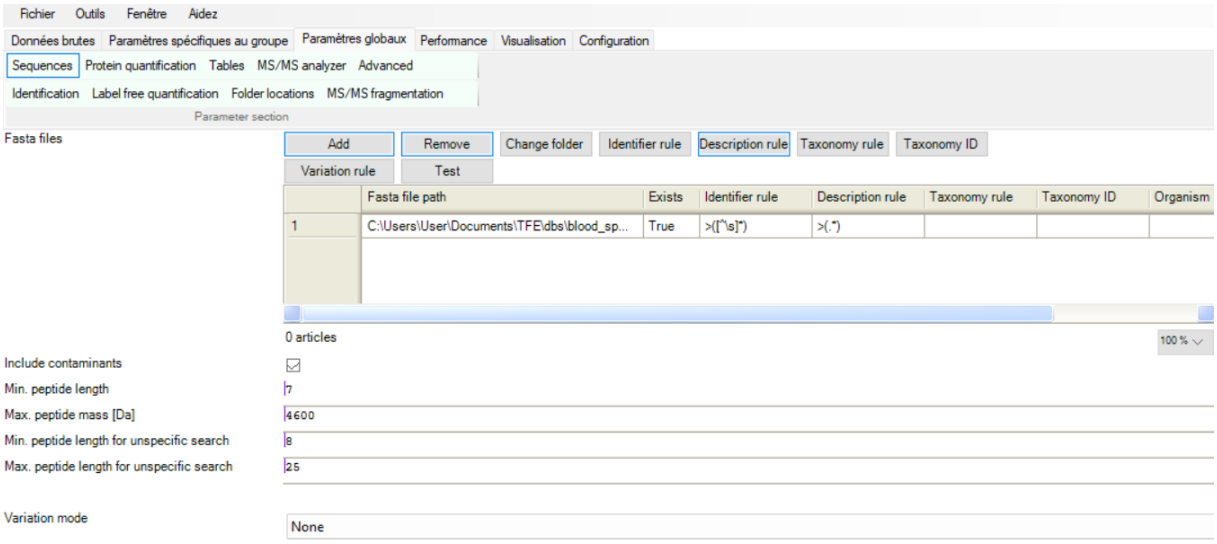


Figure 21 : Ajout d'une base de données au format fasta

La majorité des paramètres de Maxquant sont automatiquement déterminés via la reconnaissance du format des données de séquençage, ainsi que via l'utilisation de paramètres communs en analyse protéomique tels que la digestion à la trypsine ou les modifications.

En bas de la fenêtre, le nombre de processeurs alloués au programme peut être modifié. Enfin le bouton « Début » lancera l'analyse.

Lancement de l'analyse

Lorsque le bouton « Début » est pressé, plusieurs fichiers et dossiers sont créés. Parmis ceux-ci :

- Un dossier « combined » est ajouté à l'arborescence. Ce dernier contient les informations sur la recherche des pics, le temps nécessaire à chaque processus ainsi que les tables d'output générés à la fin de l'analyse.
- Un fichier mqpar.xml permettant de voir les paramètres de l'analyse et de les réutiliser lors d'une prochaine analyse.

Fichiers de sortie

Les fichiers de sorties sont des tables de données au format TSV trouvable dans le dossier combined/txt/. Chaque table a sa spécificité et son utilité. Dans le cadre de ce travail, le fichier le plus utilisé est le fichier peptide.txt, reprenant tous les peptides séquencés et les informations associées (identifiant de séquence, rapport m/z, charge, …).

C) Utilisation de scripts python

Les fichiers de sortie ont été filtés, triés et convertis au format Excel en utilisant des scripts en python (version 3.12.0) avec la librairie Pandas (version 2.2.0).

L'identifiant uniprot de chaque peptide a été envoyé à l'API d'Uniprot via la librairie requests (version 2.31.0) pour récupérer les informations sur chaque peptide et les inclure dans les tableaux de données.

L'utilisation de script a également permis l'élaboration d'un parseur permettant de récupérer les informations des fichiers générés par le programme pepdigest et de les traiter.

Enfin l'utilisation de la librairie matplotlib a permis la représentation graphique des données.

D) Traitement de données via Excel

Les tableaux Excel permettent de synthétiser les informations sur l'ensemble des données en utilisant des filtres, des tris, des tableaux croisés dynamiques et de graphiques.

3.5. Matériel informatique

Toutes les analyses informatiques ont été effectuées sur un PC portable Lenovo Ideapad 3, équipé d'un processeur AMD Ryzen 5 3500U avec Radeon Vega Mobile Gfx à 2,10 GHz, 8 cœurs logiques, 512 Go de stockage et 12 Go de RAM. Les analyses avec le logiciel MaxQuant ont été réalisées sous Windows 11 Famille (version 22H2), tandis que les programmes EMBOSS ont été exécutés sous Linux Gentoo (kernel 6.6.47-gentoo-dist, profile default/linux/amd64/23.0/split-usr/no-multilib).

4. Résultats

4.1. Prévisions théoriques

A) Construction de la banque de protéines sanguines

L'ordinateur sur lequel ce travail a été réalisé possédait au préalable une banque swissprot au format swiss indexée via la méthode EMBOSS avec les champs identifiants, code d'accès, organismes, mots clés, descriptions et la version de séquence, nommée « sp ». Cette banque servira de base à la création de la nouvelle banque.

Dans un premier temps, les entrées correspondant à la sous-unité alpha de l'hémoglobine ont été récupérés grâce à entret, puis le nombre de séquence a été compté via grep :

```
entret sp-id:HBA*
```

```
lucinux /data/tfe/hemoglobin # entret sp-id:HBA*
Retrieve sequence entries from flatfile databases and files
Full text output file [hba_tapge.entret]: hba.entret
lucinux /data/tfe/hemoglobin # less hba.entret
lucinux /data/tfe/hemoglobin # grep ^ID hba.entret | head
ID   HBA_TAPGE               Reviewed;         141 AA.
ID   HBA_LUTLU               Reviewed;         141 AA.
ID   HBAB_SERQU              Reviewed;         144 AA.
ID   HBAD_ACCGE              Reviewed;         141 AA.
ID   HBA_MANSP               Reviewed;         141 AA.
ID   HBA_AYTFU               Reviewed;         142 AA.
ID   HBA1_HAPGR              Reviewed;         140 AA.
ID   HBA_THUTH               Reviewed;         143 AA.
ID   HBA_RABIT               Reviewed;         142 AA.
ID   HBA2_NOTAN              Reviewed;         141 AA.
lucinux /data/tfe/hemoglobin # grep ^ID hba.entret |wc -l
359
```

Figure 22 : Récupération des séquences d'hémoglobine (sous-unité alpha)

On constate que le nombre d'hémoglobines sous-unité alpha présentes dans la banque swissprot est de 359 séquences. Ce procédé permet de récupérer toutes les séquences de protéines d'intérêt. Il a donc été répété pour tous les identifiants des protéines sanguines majeures, dont le nom et l'identifiant uniprot est répertorié ci-dessous.

Tableau 4 : Association des protéines sanguines majeures à leur identifiant uniprot

Nom des protéines	Identifiant uniprot
Hemoglobin subunit alpha	HBA
Hemoglobin subunit beta	HBB
Albumin	ALBU
Transferrin receptor	TFR1
Serotransferrin	TFR2
Lactotransferrin	TRFE
Fibrinogen alpha chain	FIBA
Fibrinogen beta chain	FIBB
Fibrinogen gamma chain	FIBG
Prothrombin	THRB
Tissue factor	TF
Coagulation factor VII	FA7
Coagulation factor VIII	FA8
Coagulation factor IX	FA9
Coagulation factor X	FA10
Coagulation factor XI	FA11
Coagulation factor XIII	FA12
IgG	IGHG
Immunoglobulin heavy constant alpha	IGHG
Alpha-2-macroglobulin	A2MG
Apolipoproteines	APO
Alpha-1-antitrypsin	A1AT

Le code permettant de récupérer toutes les séquences de chaque protéine et d'en compter le nombre se trouve en Annexe 1.

Le tableau 5 ci-dessous reprend le nombre de séquences récupérées pour chaque protéine sanguine. On constate que les protéines les plus abondantes dans la base de données sont les hémoglobines alpha et bêta. On remarque aussi la présence d'un grand nombre d'apolipoprotéines. En effet ces dernières n'ont pas été séparées en fonction du type d'apolipoprotéine, ce qui fait que toutes les protéines de cette famille y sont représentées ensemble.

Tableau 5 : Résultat du dénombrement des protéines sanguines majoritaires dans la base de données swissprot

Dénombrement	Identifiant de la protéine
1	IGHG
3	FA11
3	TFR2
4	FA8
6	FA12
7	FA10
7	FA7
7	FIBG
7	TF
7	THRB
10	TRF1
12	A2MG
12	FA9
20	TRFE
29	A1AT
30	ALBU
38	FIBB
53	FIBA
345	HBB
359	HBA
379	APO

Le script présenté en Annexe 1 a également créé un fichier pour chaque protéine contenant toutes les séquences trouvées dans swissprot. En combinant tous ces fichiers, on obtient un fichier au format swiss reprenant l'ensembre de nos protéines sanguines majoritaires. Ce fichier contient 1339 séquences. Il est à noter que sur ces séquences, 144 sont uniquement des fragments de la protéine, tandis que 497 sont des précurseurs, dont 340 uniquement pour les apolipoprotéines. Pour ces derniers, les résultats seront donc à nuancer.

```
cat *.dat >> blood.dat
grep -c ^ID blood.dat
1339
```

Cette banque peut maintenant être indexée via la méthode EMBOSS en utilisant le programme dbxflat, avec une ressource personnalisée pour la longueur des champs enregistrée dans le fichier emboss.default.

```
lucinux /data/tfe/blood_db # dbxflat
Index a flat file database using b+tree indices
Basename for index files: blood
Resource name: mares
      EMBL : EMBL
     SWISS : Swiss-Prot, SpTrEMBL, TrEMBLnew
        GB : Genbank, DDBJ
    REFSEQ : Refseq
     FASTQ : Fastq files
     USPTO : Iguspto files
Entry format [SWISS]:
Wildcard database filename [*.dat]: blood.dat
Database directory [.]:
        id : ID
       acc : Accession number
        sv : Sequence Version and GI
       des : Description
       key : Keywords
       org : Taxonomy
Index fields [id,acc]: id,acc,des,key,org
Compressed index files [Y]: N
General log output file [outfile.dbxflat]:
```

Figure 23 : Indexation de la sous-banque de protéines sanguines issue de swissprot.

```
lucinux /data/tfe/blood_db # cat outfile.dbxflat
Processing directory: ./
Processing file: blood.dat
entries: 1339 (1339) time: 0.4/0.0s (0.4/0.0s)
Total time: 0:00.0
Entry idlen 15 OK. Maximum ID length was 11 for 'A1ATR_HUMAN'.
Field acc acclen 15 OK. Maximum acc term length was 10 for 'A0A0G2JPK4'.
Field org orglen 125 OK. Maximum org term length was 54 for 'strain ATCC 43589 / DSM 3109 / JCM 10099 / NBR
C 100826'.
Field des deslen 125 OK. Maximum des term length was 27 for 'Alpha-1-antitrypsin-related'.
Field key keylen 125 OK. Maximum key term length was 35 for 'Congenital dyserythropoietic anemia'.
```

Figure 24 : Contrôle de la taille des champs et de l'indexation.

Ensuite, pour que la banque soit utilisable avec les programmes de EMBOSS, il faut la recenser dans le fichier emboss.default :

```
# Banque de protéines sanguines à partir de swissprot
DB blood [
      type: P
      dir: /data/tfe/blood_db/
      method: emboss
      format: swiss
      fields: "id acc org key des"
      comment: "Index des protéines sanguines dans swissprot (TFE)"
      ]
```

La commande ***showdb*** permet de vérifier que la banque est correctement reprise dans le fichier emboss.defaut tandis que la commande ***infoseq blood : | head*** permet de vérifier si les programmes emboss reconnaissent la banque (en montrant les premières lignes de résultat lorsqu'on interroge la banque).

Maintenant que la banque est indexée, on peut déterminer le nombre de protéines qu'elle contient pour chacun des organismes.

```
Infoseq blood : -only -org -noheading | sort | uniq -c | sort -nr >
      nb_per_org
```

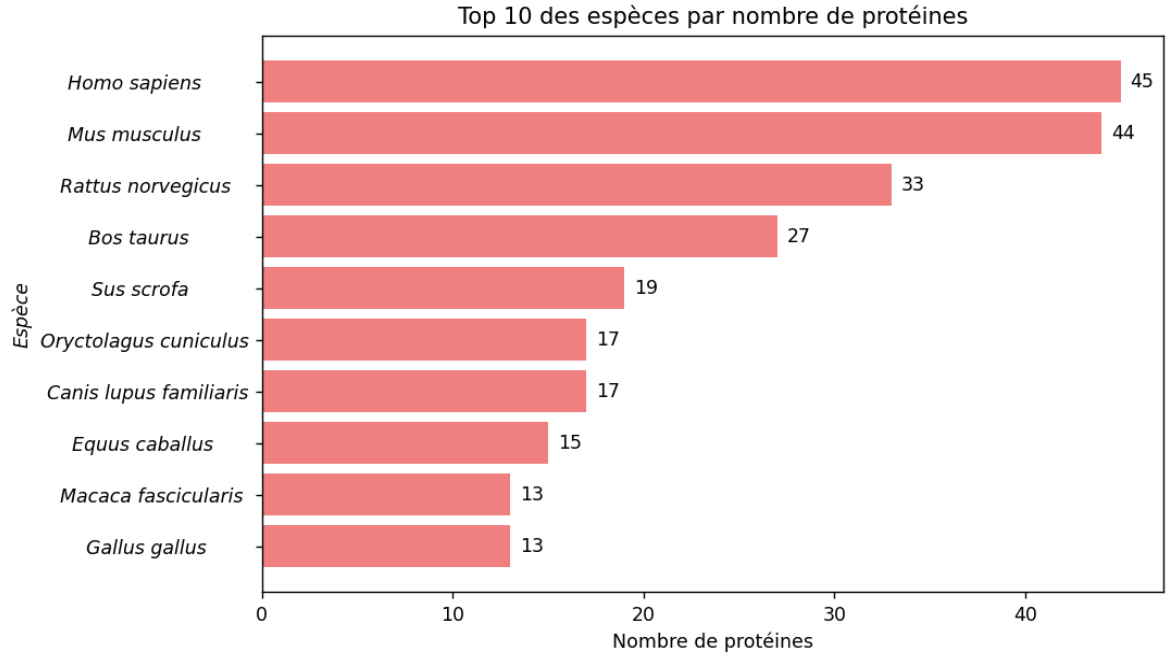


Figure 25 : Nombre de séquences de protéines présentes par organisme (10 organismes les plus représentés).

436 organismes différents sont présents dans la banque. La Figure 25 ci-dessus ne montre la répartition que pour les 10 organismes comprenant le plus de séquences dans la banque.

B) Digestion de la banque de données

```
infoseq blood: -only -usa -noheading | pepdigest @stdin -menu 1 -mono Y -
      overlap -out all_blood.pepdigest
```

La commande ci-dessus permet de récupérer l'identifiant de toutes les séquences de la base de données via la commande infoseq et de les passer directement à pepdigest pour digérer les séquences *in silico*. Les paramètres utilisés lors de la commande sont :

```
Infoseq
      blood: : banque de données indexée
      - only : ne montre que les champs demandés par la suite
      - usa : pour Uniform Sequence Adressing, récupérer les informations
      sur la banque et l'id d'où proviennent les séquences
      - noheading : ne pas montrer les titres des colonnes.

 Pepdigest
      @stdin : utilise l'identifiant récupéré depuis le standard input (à
      partir de ce qui est affiché à l'écran)
      -menu 1 : utilisation de la trypsine pour la digestion
      -mono Y : utilisation des masses monoisotopiques (masses calculées à
      partir de l'isotope le plus abondant)
      -overlap : autorise 1 misscleavage (le peptide n'est pas clivé après
      le premier résidu lysine ou arginine)
```

La structure des fichier générés par pepdigest est présentée en Figure 26. Chaque protéine de la base de données est digérée à la trypsine et les peptides tryptiques générés sont présenté dans l'ordre décroissant de leur masse moléculaire.

```
#========================================
#
# Sequence: A1AT_MESAU     from: 1   to: 413
# HitCount: 26
#
# Complete digestion with Trypsin yields 26 fragments
#
#========================================

  Start      End Mol_Weight Cterm  Nterm  Sequence
    114      154    4765.498 K      N      GFHNLLQTFNRPDNELQLTTGNGLFIHNNLKLVDKFLEEVK
    237      278    4702.335 K      M      VPMMSRLGMFDVHYVSTLSSWVLLMDYLGNATAIFILPDDGK
    363      396    3696.821 R      Q      GTEAAGATFMEIIPMSVPPEVNFNSPFIAIIYDR
      1       34    3605.818 .      Q      MKPSISWGILLLAGLCCLVPSFLAEDAQETDASK
     59       88    3223.659 R      G      ELVHQSNTTNIFFSPVSIATAFAMLSLGTK
    320      350    3168.692 K      A      TALDPLGITQVFSNGADLSGITEDVPLKLGK
     35       58    2810.290 K      E      QDQEHQACCKIAPNLADFSFNLYR
     89      113    2705.402 K      G      GVTHTQILEGLGFNLTEIAEAEVHK
    155      174    2317.013 K      V      NDYHSEAFSVNFTDSEEAKK
    213      231    2291.049 K      T      WKKPFDADNTEEADFHVDK
    302      319    1975.053 R      T      SANVHFPKLSISGTYNLK
    198      210    1541.849 K      G      DTVLALVNYIFFK
    279      288    1240.623 K      E      MQHLEQTLNK
    183      193    1156.682 K      D      GTQGKIVDLVK
    289      296     946.585 K      D      EIIGKFLK
    355      362     915.503 K      G      AVLTIDER
    175      182     904.502 K      G      VINGFVEK
    401      407     746.433 K      V      SPLFVGK
    408      413     685.376 K      .      VVDPTR
    232      236     548.317 K      V      TTTVK
    194      197     489.243 K      D      DLDK
    351      354     453.270 K      A      AVHK
    397      400     446.249 R      S      QTAK
    299      301     412.218 R      S      HTR
    297      298     289.139 K      H      DR
    211      212     203.127 K      W      GK

#----------------------------------------
#----------------------------------------
#========================================
#
# Sequence: A1AT_MESAU     from: 1   to: 413
# HitCount: 24
#
#
#
# Partial digest with Trypsin yields 24 extras.
# Only overlapping partials shown:
#
```

Peptides sans misscleavage

Peptides avec misscleavage

Figure 26 : Format des fichiers générés par pepdigest.

Chaque protéine est représentée 2 fois, une première fois pour les peptides sans misscleavage, et une seconde fois pour les peptides avec 1 misscleavage.

```
grep '^\s\+[1-9]' all_blood.pepdigest | wc -l
45774
```

Dénombrement des peptides

La commande ci-dessus permet de compter le nombre de peptides générés au total dans la banque. Cependant, certains peptides sont générés plusieurs fois parfois même au sein de la même protéine. Un script python a été créé pour compter le nombre de peptides différents générés et le nombre de protéines différentes dans lesquelles ces peptides sont retrouvées. Ce script récupère la séquence des peptides du fichier all_blood.pepdigest et regarde si elle a déjà été vue précédemment. Si non, une clé est créée dans un dictionnaire avec la séquence du peptide, et

l'identifiant de la protéine est ajouté à une liste vide correspondant à cette clé. Si oui, alors l'identifiant de la protéine est ajouté à la liste des identifiants déjà présents pour ce peptide. Les détails de ce script se trouvent en Annexe 2. Le script retourne une ligne par peptide avec le nombre de protéines qui contiennent ce peptide, la séquence du peptide et les identifiants des protéines.

```
python ../ids_per_pept.py all_blood.pepdigest | wc -l
28594
```

Au total, 28 594 peptides différents sont générés. Une petite modification du script permet de voir parmis ceux-ci combien de peptides sont spécifiques à une seule protéine de la banque (détaillé en Annexe 2).

```
python ../ids_per_pept.py all_blood.pepdigest | wc -l
23188
```

On peut voir que plus de 80% des peptides générés ne sont présent que pour une seule protéine de la banque. Cela signifie que l'utilisation de peptides unique permettra de distinguer facilement les protéines entre-elles au niveau des peptides séquencés.

Il est à noter que le programme Maxquant utilise une longueur minimale des peptides lors de l'identification. Cette valeur est fixée sur une longueur de 7 acides aminés. En ajoutant une condition que la longueur des peptides doit être au moins de 7 dans le script, nous obtenons le nombe total de peptides uniques pouvant être séquencés.

```
python ../ids_per_pept.py all_blood.pepdigest | wc -l → nombre total de
      peptides de 7 acides aminés ou plus
25874
python ../ids_per_pept.py all_blood.pepdigest | wc -l → nombre de peptides
      uniques de 7 acides aminés ou plus
21419
```

Dénombrement des peptides par espèce

Voyons maintenant ce que donne ce nombre rapporté aux protéines et aux organismes. Pour cela, nous allons à nouveau utiliser le script python, récupérer uniquement l'identifiant de l'organisme dont les peptides proviennent, les trier et les compter :

```
python ../ids_per_pept.py all_blood.pepdigest | sed -e "s/^.*_//" | sort |
      uniq -c | sort -n > nombre_pep_uniq_par_espece
```

Le nombre de peptides uniques par espèce est représenté en Figure 27. Par soucis de clareté, ce graphique ne montre que les espèces avec au moins 100 peptides uniques dans la banque, c'est-à-dire les 30 espèces avec le plus de peptides uniques générés lors de la digestion de la banque. On remarque une distribution semblable à celle présentée en Figure 25, ce qui est normal car plus il y a de protéines pour une espèce dans la banque, plus il y a de chance de trouver des peptides spécifiques à cette espèce.

N.B. : Le nom scientifique des organismes a été récupéré via une commande décrite dans l'Annexe 3.

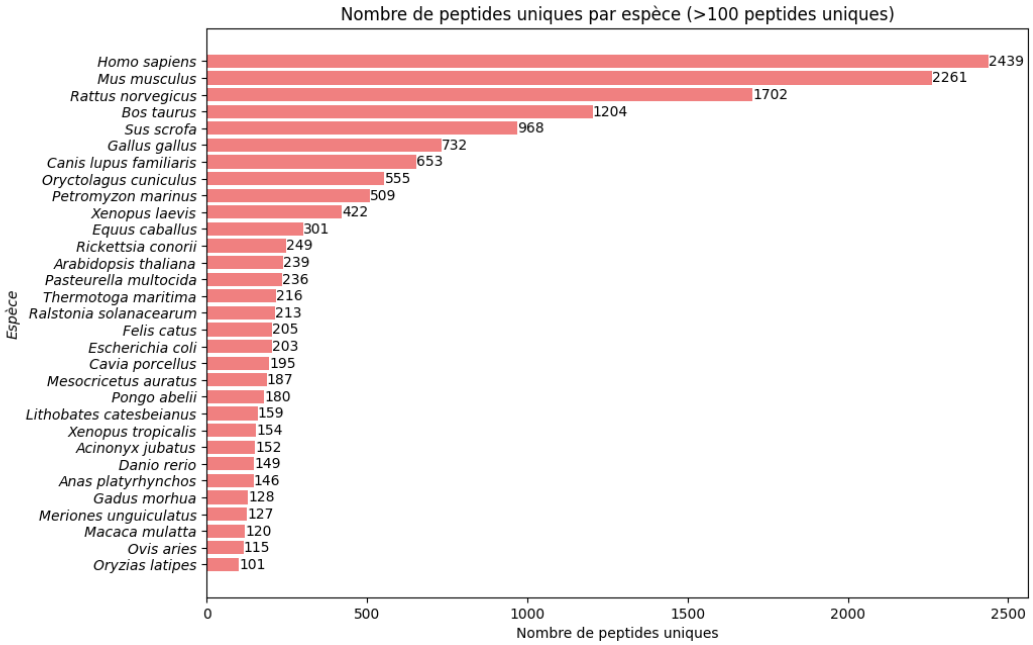


Figure 27 : Nombre de peptides uniques par espèce générés par la digestion de la banque (pour n>100)

Dénombrement des peptides par protéine

La même méthode que précédemment est appliquée afin de dénombrer les peptides uniques générés par chaque protéine.

```
python ../ids_per_pept.py all_blood.pepdigest | sed -e "s/^.* //" -e
      's/_.*$//' | sort | uniq -c | sort -n >
      nombre_pept_unique_par_protéine
```

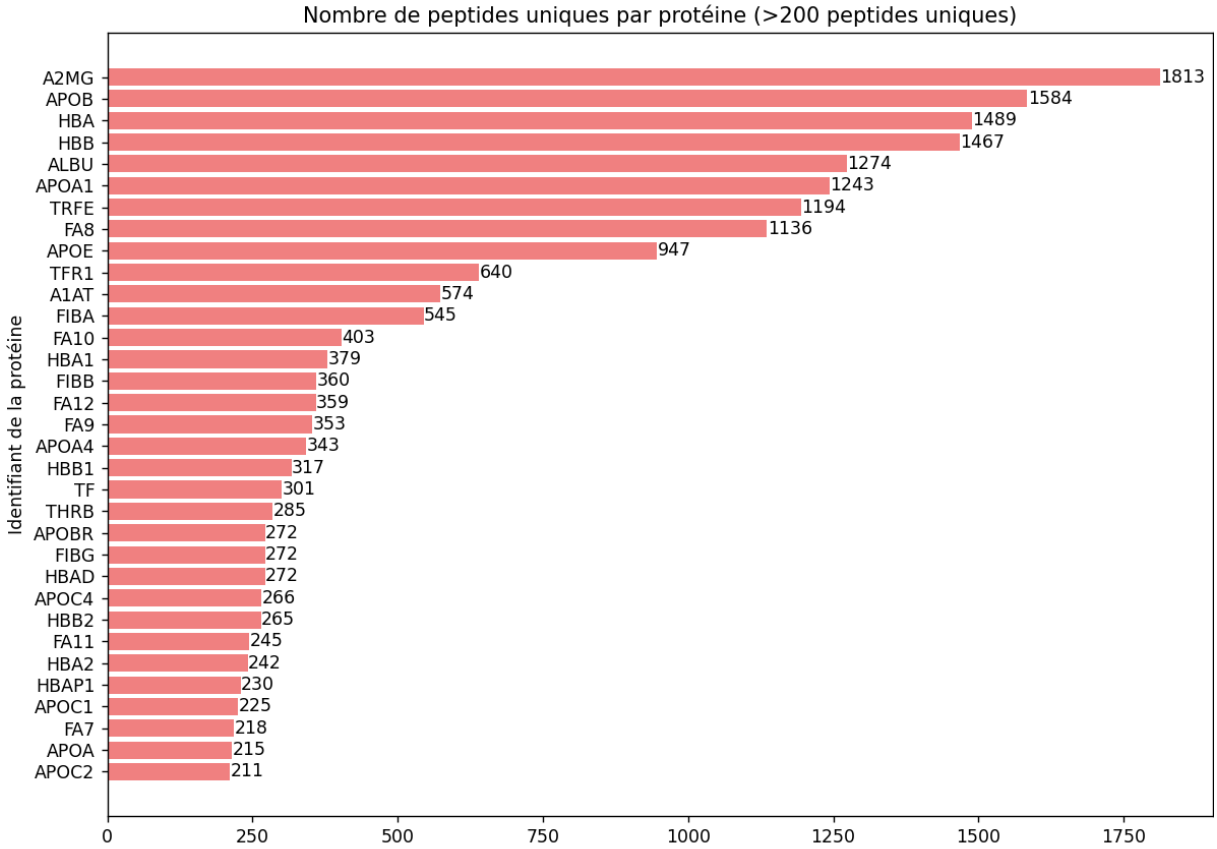


Figure 28 : Nombre de peptides uniques par protéine générés par la digestion de la banque (pour n>200)

Les résultats du dénombrement des peptides par protéine présenté en Figure 28 sont plus surprenants que pour le dénombrement par espèce. On s'attendait à voir les hémoglobines et les apolipoprotéines très haut du fait de leur nombre élevé dans la banque. Cependant, plusieurs protéines faiblement représentées dans la banque possèdent un très grand nombre de peptides uniques. C'est notamment le cas de l'albumine, qui n'a que 30 séquences dans la banque (10 fois moins que l'hémoglobine alpha) mais qui possède 1274 peptides uniques, presque autant que l'hémoglobine alpha. On peut noter également que malgré ses 12 séquences, l'alpha-2-macroglobuline possède le plus grand nombre de fragments uniques.

Compromis entre nombre de peptides uniques et nombre d'espèces représentées

Dans le but de trouver des peptides marqueurs pouvant à eux seuls déterminer la présence d'une espèce dans un échantillon de sang, il faut trouver des protéines qui possèdent à la fois un grand nombre de peptides spécifiques et une grande couverture des espèces dans la banque.

Pour analyser cela, il faut créer un tableau à double entrée type excel permettant de voir en ligne les identifiants des protéines et en colonnes les espèces.

La première étape est de récupérer l'identifiant de l'espèce et de la protéine pour chaque peptide unique trouvé :

```
python ../ids_per_pept.py all_blood.pepdigest id| awk '{print $2, $1}' |
      sed -e 's/_/ /' | sort -k2,2 -k1,1 | awk '{print $2, $1, $3}' >
      gen_pept_per_org_allblood
```

```
HUMAN TRFE NLREGTCPEAPTDECKPVK
HUMAN TRFE SDNCEDTPEAGYFAVAVVKK
HUMAN TRFE SDNCEDTPEAGYFAVAVVKKSASDLTWDNLK
HUMAN TRFE TAGWNIPMGLLYNKINHCRFDEFFSEGCAPGSKKDSSLCKLCMGSGLNLCEPNNK
HYDGR HBA NAELYGAETLTRLFAAHPTTK
HYDGR HBA NAELYGAETLTRLFAAHPTTKTYFPHFDLSPGSNDLK
HYDGR HBA TYFPHFDLSPGSNDLK
HYDGR HBA TYFPHFDLSPGSNDLKVHGKK
HYDGR HBA VAWVPVSK
HYDGR HBA VAWVPVSKNAELYGAETLTRLFAAHPTTK
HYDGR HBA VDPDNFQFLGLCLEVTIAAHSGGPLKPEVLLSVDKFLGQISK
```

Figure 29 : Structure du fichier gen_pept_per_org_allblood

Un nouveau script python permettra de combiner ce fichier avec le fichier généré pour le dénombrement des peptides par espèces (Annexe 3) et créer un tableau reprenant le nombre de peptides uniques par protéine et par organisme. Les détails du script peuvent être consultés en Annexe 4.

```
python ../table_create.py gen_pept_per_org_allblood
      nom_espèce_nombre_pep_uniq
```

Le tableau résultant est un tableau de 36 lignes (le nombre d'identifiants de protéines différents) et de 398 colonnes (le nombre d'organismes dans la banque). Une colonne « Total » a été ajoutée afin de dénombrer le nombre de peptides uniques générés par ligne, ainsi qu'une colonen « Nb-diff » représentant le nombre de colonnes (donc d'organismes) dans lesquelles au moins un peptide est est présent.

Sur base de ces différentes valeurs, une colonne supplémentaire est ajoutée calculant un coefficiant fictif appelé « Prot-org ». Ce dernier est arbitrairement défini par le produit entre le nombre de peptide générés par la protéine et le pourcentage d'organismes de la banque qu'elle couvre :

$$C_{po} = \sum peptides\ uniques\ * \frac{Nombre\ d'organismes\ associés}{Nombre\ d'organismes\ dans\ la\ banque} \quad (398)$$

Le Tableau 6 montre les protéines triées selon cet indice par ordre décroissant. On peut voir que les hémoglobines, certaines apolipoprotéines et l'albumine possèdent des scores permettant d'utiliser ces protéines pour identifier différentes espèces en fonction des peptides séquencés.

Tableau 6 : Tableau résumé de la comparaison entre le nombre de peptides d'une protéine et le nombre d'organismes associés.

Identifiant de la protéine	Nombre de peptides	Nombre d'organismes représentés	Coefficient Prot-org
HBB	1467	191	704,01
HBA	1489	186	695,86
APOA1	1243	53	165,53
APOE	947	59	140,38
ALBU	1274	21	67,22
A2MG	1813	12	54,66
TRFE	1194	16	48,00
HBA1	379	32	30,47
HBB1	317	31	24,69
A1AT	574	17	24,52
FIBA	545	17	23,28
APOC1	225	38	21,48
APOB	1584	5	19,90
HBAD	272	29	19,82
HBB2	265	29	19,31
APOC2	211	33	17,49
FIBB	360	19	17,19
TFR1	640	10	16,08
APOC4	266	24	16,04
HBA2	242	25	15,20
FA8	1136	4	11,42
FA9	353	12	10,64

APOA4	343	12	10,34

Les fichiers générés permettant l'identification des peptides peuvent être retrouvés via ce lien :

https://drive.google.com/drive/folders/1Xr7Du2h7cx9-58CycQU4jT3adU_TKsVi?usp=drive_link

4.2. Validations expérimentales

A) Mise en place de l'analyse

Plusieurs extractions ont été réalisées entre janvier et mai 2024 (voir Matériel et méthodes). Les données des protéines séquencées ont été reçue sous forme de dossiers « .d », caractéristiques des instruments Brucker.

Les outils nécessaires à l'analyse de données de spectrométrie de masse sont :

- Les données brutes (ions avec leur rapport m/z et leur intensité)
- Une base de données de référence.

La base de données utilisée sera la banque indexée de protéines sanguines, réalisée au point 4.1.A). Les séquences au format fasta de la banque peuvent être obtenues via le programme ***seqret*** d'EMBOSS

```
seqret blood: -out blood_sp.fasta
```

Le fichier fasta généré a une taille de 443ko.

B) Premiers résultats

Le premier séquençage a été réalisé sur les insectes FG10, TG02 et TG16.

Le fichier « peptide.txt » généré par l'analyse peut être ouvert directement via Excel. Cependant, il contient un certain nombre de colonnes inutiles pour l'analyse effectuée, telles que le décomte de chaque acide aminé dans la séquence du peptide, les positions de début et de fin du peptide dans la protéine, …

Une colonne « Potential contaminant » remplie de vide ou de « + » indique que la séquence peut provenir d'un contaminant. Maxquant a une base de données intégrée de protéines classées comme contaminants telles que la kératine (peau ou cheveux tombé(e)s lors des manipulations) ou de la trypsine (provenant principalement du porc).

Enfin, une colonne « Reverse » également remplie de lignes vides ou contenant un « + » permet de vérifier le taux de faux-positifs lors de l'analyse. En effet Maxquant compare les pics avec la base de données mais où les séqeuces sont lues à l'envers. Ces séquences inversées ne sont pas biologiquement réelles, mais servent à estimer le taux de fausses découvertes (FDR).

La première étape est de filtrer le tableau pour ne garder que les colonnes utiles par la suite, n'utiliser que les peptides uniques à une seule espèce et enlever toutes les protéines identifiées comme contaminantes ou provenenant de séquence reverse. Le scipt « filter_peptide.py » permettant la filtration du fichier « peptide.txt » se trouve en Annexe 5.

Ensuite, les identifiants présents dans la colonne « Proteins » sont récupés et envoyés à l'API d'Uniprot afin de récupérer les informations sur la protéine d'où provient le peptide. Le script permettant la récupération des informations sur les protéines se trouve en Annexe 6.

Enfin, le tableau reprenant les informations pour chaque protéine doit être fusionné au tableau peptide filtré afin d'avoir toutes les données sur un seul et même fichier, permettant le dénombrement des espèces et des protéines séquencées. Le script permettant la fusion des deux tableaux se trouve en Annexe 7.

Les tableaux croisés dynamiques ci-dessous permettent de synthétiser les données pour chaque individu.

Tableau 7 : Résultats du séquençage de l'individu FG10

Intensity FG10	(Plusieurs éléments)
Étiquettes de lignes	**Nombre de Sequences**
Homo sapiens	**11**
Alpha-1-antitrypsin	1
Apolipoprotein B-100	3
Coagulation factor VIII	1
Fibrinogen alpha chain	2
Fibrinogen beta chain	1
Fibrinogen gamma chain	1
Serotransferrin	2
Struthio camelus	**2**
Hemoglobin subunit alpha-D	1
Hemoglobin subunit beta	1
Ovis aries	**1**
Hemoglobin subunit beta	1
Dinomys branickii	**1**
Apolipoprotein E	1
Erethizon dorsatum	**1**
Apolipoprotein E	1
Erinaceus europaeus	**1**
Hemoglobin subunit alpha	1
Oryctolagus cuniculus	**1**
Hemoglobin subunit alpha-1/2	1
Equus caballus	**1**
Transferrin receptor protein 1	1
Petromyzon marinus	**1**
Fibrinogen gamma chain	1
Pongo abelii	**1**
Albumin	1
Sus scrofa	**1**
Apolipoprotein M	1
Leiostomus xanthurus	**1**
Hemoglobin subunit alpha	1
Torpedo marmorata	**1**
Hemoglobin subunit beta-2	1
Mus musculus	**1**
Hemoglobin subunit beta-1	1
Octodon degus	**1**
Apolipoprotein A-I	1
Total général	**26**

On peut voir que plusieurs protéines sanguines différentes sont présentes pour l'homme. Cela indique une très grande probabilité que ce phlébotome se soit nourri sur un homme. Toutes les autres espèces n'ont été identifiées qu'avec un seul peptide unique (deux pour *Struthio*

camelus), ce qui n'est pas suffisant pour pouvoir affirmer que ces organismes ont bien été piqués par le phlébotome.

Tableau 8 : Résultats du séquençage de l'individu TG02

Intensity TG02	(Plusieurs éléments)
Étiquettes de lignes	**Nombre de Sequences**
Gallus gallus	**81**
Albumin	46
Fibrinogen alpha chain	5
Fibrinogen beta chain	12
Hemoglobin subunit alpha-D	1
Ovotransferrin	17
Ovis aries	**2**
Albumin	1
Hemoglobin subunit beta	1
Petromyzon marinus	**1**
Fibrinogen gamma chain	1
Spalax ehrenbergi	**1**
Hemoglobin subunit alpha	1
Pongo abelii	**1**
Albumin	1
Macaca mulatta	**1**
Apolipoprotein(a)	1
Aldabrachelys gigantea	**1**
Hemoglobin A/D subunit beta	1
Mus musculus	**1**
Hemoglobin subunit beta-1	1
Phoenicopterus ruber	**1**
Hemoglobin subunit alpha-A	1
Oryctolagus cuniculus	**1**
Apolipoprotein E	1
Rhinoceros unicornis	**1**
Hemoglobin subunit beta	1
Struthio camelus	**1**
Hemoglobin subunit beta	1
Homo sapiens	**1**
Coagulation factor VIII	1
Sus scrofa	**1**
Apolipoprotein M	1
Torpedo marmorata	**1**
Hemoglobin subunit beta-2	1
Pan troglodytes	**1**
Hemoglobin subunit zeta	1
Peromyscus californicus	**1**
Hemoglobin subunit alpha	1
Total général	**98**

On peut voir que plusieurs protéines sanguines différentes sont présentes pour la poule. Cela indique une très grande probabilité que ce phlébotome se soit nourri sur une poule.

L'ovotransferrine est l'équivalent de la serrotransférine chez les mammifères. Toutes les autres espèces n'ont été identifiées qu'avec un seul peptide unique (deux pour le mouton), ce qui n'est pas suffisant pour pouvoir affirmer que ces organismes ont bien été piqués par le phlébotome.

Tableau 9 : Résultats du séquençage de l'individu TG16

Intensity TG16	(Plusieurs éléments)
Étiquettes de lignes	**Nombre de Sequences**
Bos taurus	**19**
Apolipoprotein D	1
Fibrinogen gamma-B chain	18
Bos javanicus	**5**
Hemoglobin subunit beta-A	5
Homo sapiens	**4**
Apolipoprotein B-100	1
Coagulation factor VIII	1
Serotransferrin	2
Bos gaurus frontalis	**2**
Hemoglobin subunit alpha	2
Oryctolagus cuniculus	**2**
Apolipoprotein E	1
Hemoglobin subunit beta-1/2	1
Rattus norvegicus	**2**
Apolipoprotein B-100	2
Spalax ehrenbergi	**2**
Hemoglobin subunit alpha	1
Hemoglobin subunit beta	1
Canis lupus familiaris	**2**
Albumin	2
Mus musculus	**2**
Apolipoprotein D	1
Hemoglobin subunit beta-1	1
Struthio camelus	**1**
Hemoglobin subunit beta	1
Procavia capensis habessinica	**1**
Hemoglobin subunit alpha	1
Phoenicopterus ruber	**1**
Hemoglobin subunit alpha-A	1
Bradypus tridactylus	**1**
Hemoglobin subunit alpha	1
Salmo salar	**1**
Apolipoprotein A-I	1
Megaderma lyra	**1**
Hemoglobin subunit beta	1
Gallus gallus	**1**
Albumin	1
Tachyglossus aculeatus aculeatus	**1**
Hemoglobin subunit beta	1
Pongo abelii	**1**
Albumin	1

Tapirus terrestris	**1**
Hemoglobin subunit alpha-1/2	1
Cercocebus atys	**1**
Hemoglobin subunit beta	1
Trichechus inunguis	**1**
Hemoglobin subunit alpha	1
Dasyurus viverrinus	**1**
Hemoglobin subunit beta	1
Balaenoptera acutorostrata	**1**
Hemoglobin subunit alpha	1
Sus scrofa	**1**
Apolipoprotein M	1
Capra hircus	**1**
Hemoglobin subunit beta-A	1
Peromyscus californicus	**1**
Hemoglobin subunit alpha	1
Theropithecus gelada	**1**
Hemoglobin subunit alpha	1
Octodon degus	**1**
Apolipoprotein A-I	1
Ondatra zibethicus	**1**
Hemoglobin subunit beta	1
Ornithorhynchus anatinus	**1**
Hemoglobin subunit alpha	1
Total général	**61**

On peut voir que plusieurs protéines sanguines différentes sont présentes pour la vache. Cela indique une très grande probabilité que ce phlébotome se soit nourri sur une vache. On voit également que l'homme a été séquencé avec trois protéines différentes, ce qui peut aussi indiquer sa présence dans le repas sanguin du phlébotome. Toutes les autres espèces n'ont été identifiées qu'avec un ou deux peptide(s) unique(s), ce qui n'est pas suffisant pour pouvoir affirmer que ces organismes ont bien été piqués par le phlébotome.

Cependant on voit que 5 peptides proviennent de l'hémoglobine bêta-A de *Bos gaurus*. Cela pourrait être un indice sur la présence de cette espèce dans le repas sanguin du phlébotome, mais c'est une espèce endémique de l'ile de Java, beaucoup trop éloignée de la zone de récolte des phlébotomes. De plus, les peptides ne sont présents que pour une seule protéine. Il s'agit probablement d'une séquence appartenant à la vache qui a mal été attribuée par Maxquant.

C) Résultats combinés des séquençages « single insect »

Le même mode opératoire a été utilisé pour traiter le résultat de séquençage des autres insectes. Le résultat combiné pour les insectes FG10, TG02, TG16, TG05, TG14, TG22, PA2A, PA2B, PA2C et PA2D est présenté dans le tableau 10

.

Tableau 10 : Résultat combiné des séquençages "single insect"

Étiquettes de lignes	Nombre de Sequences
Gallus gallus	**82**
Albumin	47
Fibrinogen alpha chain	5
Fibrinogen beta chain	12
Hemoglobin subunit alpha-D	1
Ovotransferrin	17
Mus musculus	**26**
Albumin	2
Apolipoprotein A-IV	1
Apolipoprotein B-100	5
Apolipoprotein D	1
Beta-2-glycoprotein 1	1
Coagulation factor X	1
Fibrinogen beta chain	2
Fibrinogen gamma chain	1
Hemoglobin subunit alpha	1
Hemoglobin subunit beta-1	4
Hemoglobin subunit beta-2	2
Serotransferrin	5
Homo sapiens	**19**
Alpha-1-antitrypsin	2
Apolipoprotein A-IV	2
Apolipoprotein B-100	3
Apolipoprotein L5	1
Coagulation factor VIII	1
Fibrinogen alpha chain	4
Fibrinogen beta chain	2
Fibrinogen gamma chain	1
Serotransferrin	3
Bos taurus	**19**
Apolipoprotein D	1
Fibrinogen gamma-B chain	18
Ovis aries	**14**
Albumin	11
Alpha-1-antiproteinase	2
Hemoglobin subunit beta	1
Canis lupus familiaris	**12**
Albumin	10
Coagulation factor VIII	1
Transferrin receptor protein 1	1
Oryctolagus cuniculus	**9**
Albumin	2
Apolipoprotein B	1
Apolipoprotein E	1
Coagulation factor X	1

Hemoglobin subunit alpha-1/2	1
Hemoglobin subunit beta-1/2	1
Serotransferrin	2
Rattus norvegicus	**8**
Apolipoprotein B-100	6
Hemoglobin subunit beta-2	1
Serotransferrin	1
Microtus pennsylvanicus	**6**
Hemoglobin subunit alpha	3
Hemoglobin subunit beta	3
Petromyzon marinus	**5**
Fibrinogen alpha-1 chain	1
Fibrinogen alpha-2 chain	1
Fibrinogen gamma chain	2
Serum albumin SDS-1	1
Sus scrofa	**4**
Albumin	2
Apolipoprotein M	1
Coagulation factor VIII	1
Sphenodon punctatus	**4**
Hemoglobin subunit alpha-A	2
Hemoglobin subunit alpha-D	1
Hemoglobin subunit beta-1	1

Les espèces comprenant moins de 4 peptides séquencés ou moins de 2 protéines différentes n'ont pas été repris dans ce tableau récapitulatif. Grâce à l'ensemble des données, on peut affirmer avec un haut degré de confiance que certaines espèces sont des hôtes pour les phlébotomes capturés :

- *Gallus gallus* : la poule
- *Mus musculus* : la souris
- *Homo sapiens* : l'homme
- *Oryctolagus cuniculus* : le lapin

Des doutes peuvent être émis sur d'autres espèces. En effet, la vache (*Bos taurus*) a été séquencé 19 fois, dont 18 fois pour la même protéine. De plus l'abscence de peptides provenant d'hémoglobine (protéine avec le meilleur coefficient peptide-organisme) renforce ce doute. Cependant, la digestion de la banque a montré qu'aucun peptide spécifique pour la vache ne provenait d'une hémoglobine. Il est donc attendu de ne pas en retrouver. De nouveaux séquençages pourraient montrer la présence d'autres protéine spécifique à la vache si elle est bien un hôte du phlébotome.

Le chien, le mouton, le rat et le cochon pourraient également être des hôtes car ils possèdent chacun au moins 3 protéines identifiées.

Certaines espèces cependant semblent être également des hôtes potentiels, mais ne peuvent physiquement pas être hôte des phlébotomes. *Sphenodon punctatus* est un reptile endémique de la Nouvelle-Zélande, *Microtus pennsylvanicus* vit en Amérique du Nord tandis que *Petromyzon marinus* est une espèce de lamproie marine.

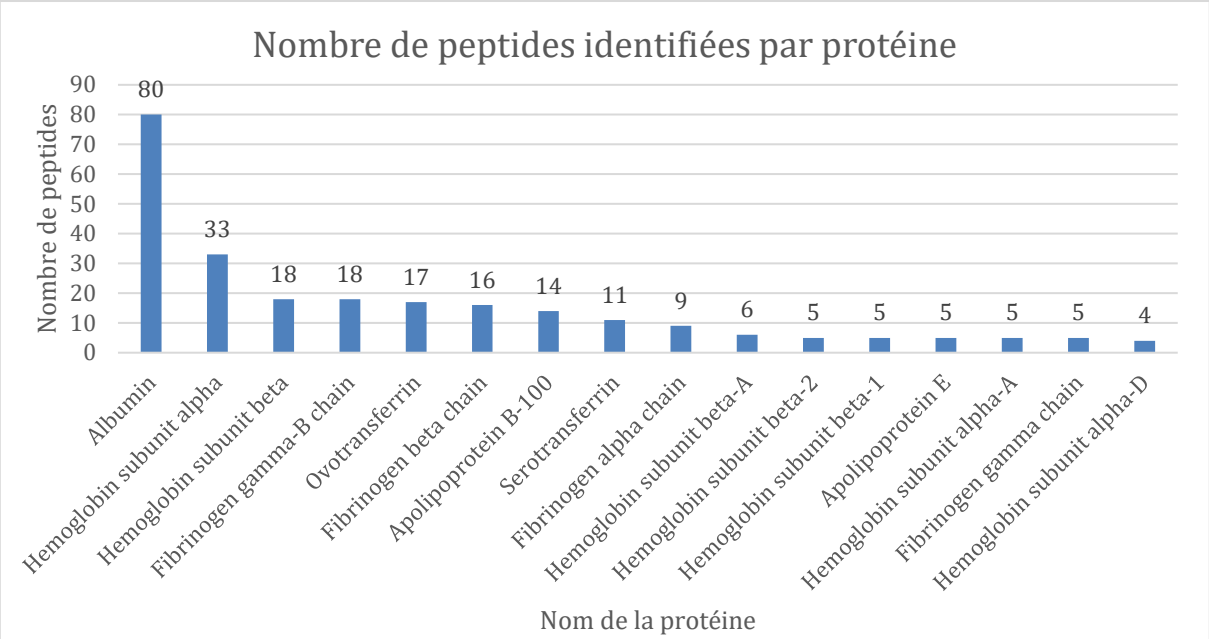


Figure 30 : Graphique du nombre de peptides séquencés par protéines sur l'ensemble des données.

La Figure 30 montre le nombre total de peptides séquencés pour chaque protéine. L'albumine arrive largement en tête, avec l'hémoglobine en second et troisième. Cela peut s'expliquer avec le nombre d'albumine séquencé uniquement pour la poule. En effet plus de la moitié des séquences provient de cette espèce. Ce graphique ci-dessus permet cependant de montrer les protéines les plus abondamment séquencées

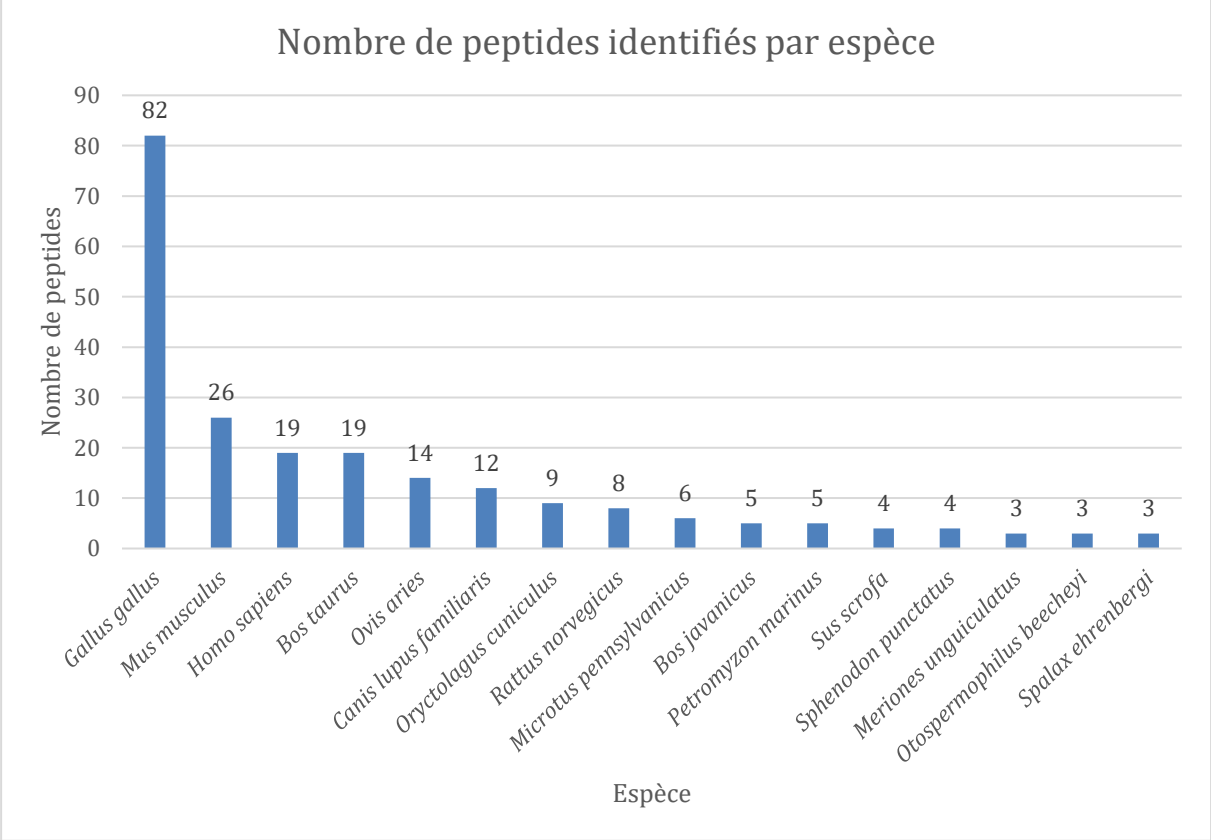


Figure 31 : Graphique du nombre de peptides séquencés par espèce sur l'ensemble des données.

La Figure 31 montre le nombre total de peptides séqencés pour chaque espèce. Cependant cette représentation est à prendre avec une vision plus critique car on a l'impression que la poule est un hôte majoritaire alors que la plupart des séquences provient de TG02, pour lequel beaucoup d'albumine et d'ovotransferrine de poule ont été séquencé.

D) Validation par alimentation spécifique

Tableau 11 : Résultats du séquençage de P. papatasi nourris spécifiquement sur l'homme

Intensity PH4A	(Plusieurs éléments)
Étiquettes de lignes	**Nombre de Sequence**
Homo sapiens	**84**
Alpha-1-antitrypsin	14
Alpha-2-macroglobulin	11
Apolipoprotein B-100	22
Apolipoprotein D	4
Coagulation factor VIII	2
Fibrinogen alpha chain	9
Fibrinogen beta chain	10
Fibrinogen gamma chain	9
Serotransferrin	3
Ovis aries	**8**
Albumin	7
Hemoglobin subunit beta	1
Sus scrofa	**5**
Apolipoprotein A-I	1
Apolipoprotein M	1
Hemoglobin subunit beta	2
Prothrombin	1
Oryctolagus cuniculus	**4**
Hemoglobin subunit beta-1/2	1
Ig gamma chain C region	1
Serotransferrin	2
Capra hircus	**4**
Hemoglobin subunit beta-C	4
Macaca nemestrina	**3**
Hemoglobin subunit alpha-1/2/3	3
Total général	**108**

Le tableau ci-dessus montre que la banque utilisée reconnait majoritairement l'espèce ciblée lors du séquençage. Cependant, la présence de différentes protéines sanguines pour le cochon (*Sus scrofa*) semble indiquer la présence de sang de cette espèce dans les échantillons. Cela pourrait s'expliquer par des instruments qui n'ont pas été nettoyé correctement, laissant une peptite contamination dûe à une précédente analyse.

Tableau 12 : Résultats du séquençage de P. papatasi nourris spécifiquement sur le mouton

Intensity PO1B	(Plusieurs éléments)
Étiquettes de lignes	**Nombre de Sequence**
Homo sapiens	**98**
Alpha-1-antitrypsin	7
Alpha-2-macroglobulin	12
Apolipoprotein A-IV	5
Apolipoprotein B receptor	1
Apolipoprotein B-100	28
Apolipoprotein D	2
Apolipoprotein L1	1
Coagulation factor VIII	1
Coagulation factor XII	1
Fibrinogen alpha chain	11
Fibrinogen beta chain	10
Fibrinogen gamma chain	11
Prothrombin	3
Serotransferrin	5
Ovis aries	**44**
Albumin	37
Alpha-1-antiproteinase	7
Gallus gallus	**21**
Albumin	14
Apolipoprotein A-I	1
Apolipoprotein B	1
Ovotransferrin	5
Sus scrofa	**4**
Albumin	1
Hemoglobin subunit beta	2
Serotransferrin	1
Capra hircus	**4**
Albumin	1
Hemoglobin subunit beta-C	3
Trichechus inunguis	**3**
Hemoglobin subunit alpha	1
Hemoglobin subunit beta	2
Oryctolagus cuniculus	**3**
Ig gamma chain C region	2
Serotransferrin	1
Total général	**177**

Le

Tableau 12 montre des résultats assez inattendus. En effet, les phlébotomes devraient être nourris avec du sang de mouton, mais ce dernier ne représente que 25% des protéines séquencées alors que l'humain réprésentait 80% des protéines séquencées dans le Tableau 11. De plus, la grande diversité de protéines identifiées pour l'humain et leur nombre ne laisse pas de place au doute, il y a clairement du sang humain dans les échantillons. Peut-être qu'à nouveau la colonne n'a pas été suffisamment nettoyée entre les deux injections (PH4A et PO3B).

Tableau 13 : Résultats du séquençage de P. sergenti nourris spécifiquement sur l'homme

Intensity SH4	(Plusieurs éléments)
Étiquettes de lignes	**Nombre de Sequence**
Homo sapiens	**109**
Alpha-1-antitrypsin	13
Alpha-2-macroglobulin	12
Apolipoprotein B-100	45
Apolipoprotein D	3
Apolipoprotein L1	2
Coagulation factor VIII	2
Coagulation factor X	1
Fibrinogen alpha chain	5
Fibrinogen beta chain	10
Fibrinogen gamma chain	12
Serotransferrin	4
Sus scrofa	**6**
Albumin	1
Apolipoprotein A-I	2
Apolipoprotein M	1
Hemoglobin subunit beta	1
Prothrombin	1
Ovis aries	**6**
Albumin	6
Tadarida brasiliensis	**5**
Hemoglobin subunit alpha-1	2
Hemoglobin subunit beta	3
Capra hircus	**3**
Hemoglobin subunit beta-C	2
Hemoglobin subunit zeta	1
Cercocebus atys	**3**
Hemoglobin subunit alpha	2
Hemoglobin subunit beta	1
Otolemur crassicaudatus	**3**
Hemoglobin subunit alpha-A	1
Hemoglobin subunit beta-1/2	2
Macaca nemestrina	**3**
Hemoglobin subunit alpha-1/2/3	3
Microtus pennsylvanicus	**3**
Hemoglobin subunit beta	3
Total général	**141**

Les phlébotomes de l'espèce *P. sergenti* nourris spécifiquement sûr l'homme montrent à nouveau de bons résultats au niveau du séquençage, avec plus de 75% des protéines identifiées appartenant à l'humain. Le cochon est à nouveau identifié avec plusieurs protéines spécifiques.

Tableau 14 : Résultats du séquençage de P. sergenti nourris spécifiquement sur le mouton

Intensity SO3A	(Plusieurs éléments)
Étiquettes de lignes	**Nombre de Sequence**
Gallus gallus	**65**
Albumin	39
Apolipoprotein A-I	3
Apolipoprotein B	1
Ovotransferrin	22
Ovis aries	**46**
Albumin	39
Alpha-1-antiproteinase	7
Homo sapiens	**8**
Apolipoprotein B-100	4
Apolipoprotein D	1
Coagulation factor VIII	2
Serotransferrin	1
Capra hircus	**5**
Albumin	1
Hemoglobin subunit beta-C	4
Sus scrofa	**4**
Albumin	1
Apolipoprotein M	1
Prothrombin	1
Serotransferrin	1
Mus musculus	**4**
Hemoglobin subunit beta-1	4
Trichechus inunguis	**3**
Hemoglobin subunit alpha	1
Hemoglobin subunit beta	2
Total général	**135**

À nouveau les résultats des phlébotomes nourris spécifiquement au sang de mouton n'ont pas révélé avec certitude la présence majoritaire de mouton.

E) Identification avec la banque de données Swissprot

Jusqu'à présent, la base de données utilisée pour l'identification était la base de données blood_sp.fasta créée au point 4.1.A). Le temps d'analyse moyen des données était entre 6h et 8h en utilisant 4 processeurs.

La banque de données Swissprot au format fasta a une taille de 289Mo, soit 500 fois plus volumineuse que la banque de protéines sanguines créée. On peut donc s'attendre à ce que l'analyse des données en utilisant la banque swissprot prenne beaucoup plus de temps.

Le programme Maxquant a mis 5 jours et 6h pour compléter l'analyse avec swissprot et toutes les données de séquençage. Les données ont été traitées avec les mêmes méthodes que précédemment. Cependant, les résultats ne prenaient pas en compte uniquement les protéines sanguines. Les résultats ont donc été filtrés avec l'outil filtre d'Excel afin de n'afficher que les protéines sanguines identifiées de manière unique. Au total, 98 peptides issus de protéines sanguines ont pu être séquencés, mais le Tableau 15 omet 7 espèces identifiées avec uniquement

une seule protéine. Ces espèces proviennent majoritairement d'Amérique du Nord, ce qui est incompatible avec le lieu de capture des phlébotomes.

Tableau 15 : Résultats du séquençage de tous les insectes comparés à la banque Swissprot

Étiquettes de lignes	Nombre de Sequences
Homo sapiens	**46**
Alpha-1-acid glycoprotein 1	1
Alpha-1-acid glycoprotein 2	1
Alpha-1-antitrypsin	5
Alpha-2-macroglobulin	3
Apolipoprotein A-IV	2
Apolipoprotein B-100	2
Apolipoprotein D	2
Fibrinogen alpha chain	3
Fibrinogen beta chain	3
Fibrinogen gamma chain	3
Hemoglobin subunit delta	2
Immunoglobulin heavy constant gamma 2	1
Immunoglobulin heavy constant gamma 3	2
Immunoglobulin heavy constant gamma 4	1
Immunoglobulin heavy constant mu	1
Immunoglobulin kappa variable 4-1	1
Lactotransferrin	11
Serotransferrin	2
Gallus gallus	**19**
Albumin	14
Apolipoprotein A-I	1
Ovalbumin	2
Ovotransferrin	2
Ovis aries	**15**
Albumin	15
Bos javanicus	**4**
Hemoglobin subunit beta-A	4
Mus musculus	**3**
Albumin	1
Hemoglobin subunit alpha	1
Hemoglobin subunit beta-1	1
Canis lupus familiaris	**2**
Albumin	2
Bos taurus	**2**
Apolipoprotein D	1
Fibrinogen gamma-B chain	1
Total général	**91**

On peut très clairement constater la présence de l'homme dans les résultats de séquençage. Ceci s'explique avec le mélange de toutes les données, y compris les phlébotomes d'élevage ayant montré de bons résultats d'identification de l'homme. Ces deux expériences influencent surement la représentation de l'espèce. On constate toujours la présence de *Bos javanicus*, certainement venant d'hémoglobine de vache. En effet, la séquence HBB_BOVIN est strictement identique à la

séquence HBBA_BOSJA dans la base de données à l'exception du 18è acide aminé, ou la lysine de la vache est remplacée par une histidine chez *Bos javanicus*, ce qui explique la présence de cette espèce, confondue avec une vache.

La Figure 32 ci-dessous montre le nombre de peptides séquencés par espèce, tandis que la Figure 33 montre le nombre de peptides séquencés par protéine. Les données sont semblables à celles décrites précédemment avec la sous-banque de protéines sanguines.

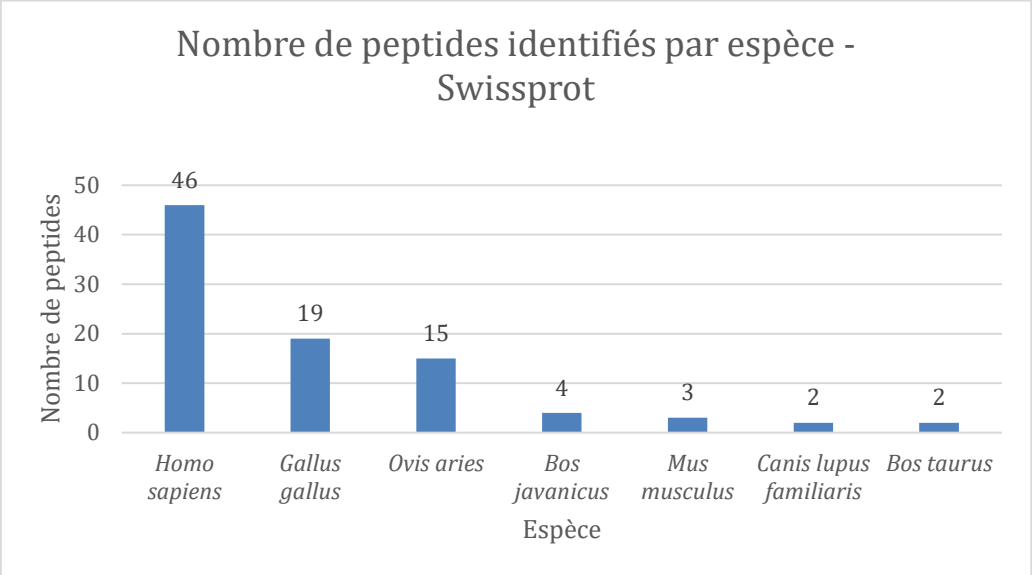


Figure 32 : Nombre de peptides identifiés par espèce avec la banque de données Swissprot

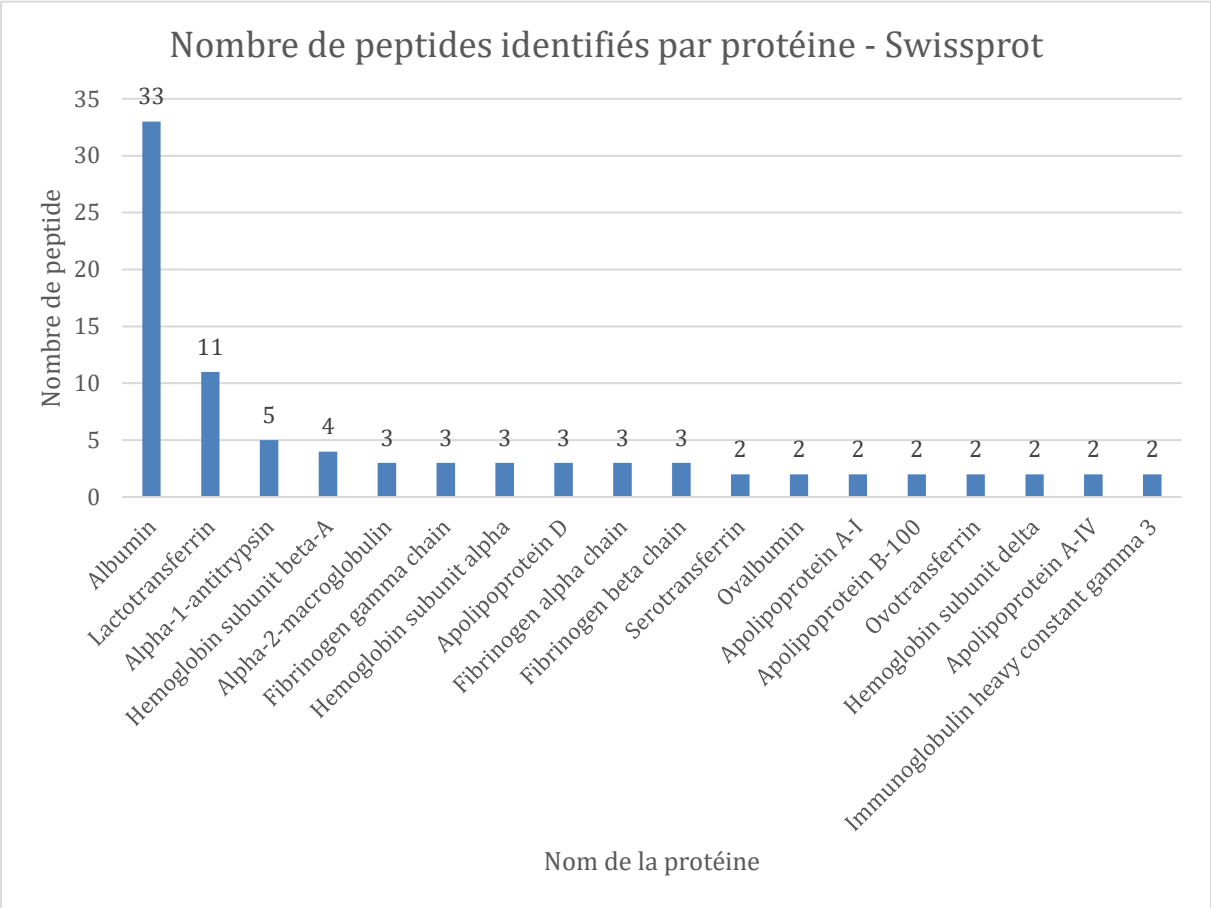


Figure 33 : Nombre de peptides identifiés par protéine avec la banque de données Swissprot

5. Discussion

5.1. Comparaison des banques de données

L'utilisation de la banque blood_sp permet une identification précise et rapide des protéines sanguines par spectrométrie de masse. Cependant, l'utilisation de la banque swissprot demande plus de temps, mais elle est moins sensible et permet de réduire le bruit. Les deux banques ont leur avantage car la banque blood_sp permet d'identifier des protéines qui seraient présentes en moins grande quantité.

5.2. Comparaison des peptides expérimentaux avec les peptides théoriques

Bien que l'hémoglobine semble être une cible de premier choix pour l'identification des espèces, elle ne garanti pas l'identification sûre de toutes les espèces étant donné que certaines ne génèrent aucun peptide unique lors de la digestion de leur hémoglobine. C'est le cas notamment du mouton et de la vache.

Les données expérimentales révèlent aussi que malgré le grand nombre de peptides générés par l'hémoglobine et le grand nombre d'espèces pour lesquelles l'hémoglobine possède au moins 1 peptide unique, ce n'est pas la protéine la plus séquencée. Cela peut être dû aux prorpiétés physico-chimiques de l'hémoglobine. En effet, le passage par la chomatographie lique en phase inverse avant la spectrométrie de masse peut entrainer une perte de certains peptides hydrophiles qui ne sont pas du tout retenus par la colonne. Cela peut également être dû à la répartition des charges du peptide qui va influencer son passage dans l'analyseur du spectromètre de masse.

5.3. Validation de la base de données via alimentation spécifique

Le séquençage des phlébotomes alimentés spécifiquement au sang humain a montré de bons résultats, avec entre 75% et 80% des peptides séquencés identifiés à l'humain. Cependant, la validation sur le mouton n'a pas montré de bons résultats. La cause la plus probable est une contamination par des produits utilisés en laboratoire, bien que les traces de contamination auraient alors dû être présentes également dans les échantillons humains.

5.4. Identification des meilleurs candidats pour l'identification

L'albumine a montré un très grand nombre d'identification par rapport aux autres protéines. Cela peut à nouveau être dû aux propriétés physico-chimiques de la protéine, lui permettan de facilement atteindre le détecteur. Cependant, bien qu'elle génère un grand nombre de peptides uniques , elle n'est présente actuellement disponible que pour 21 espèces dans la banque swissprot. La sélection de deux ou trois protéines supplémentaires comme l'hémoglobine ou les apolipoprotéines pourra permettre de déterminer avec certitude la présence d'une espèce dans un échantillon sanguin.

5.5. Conclusions et perspectives

Ce travail de fin d'études a permis de développer une base de données de protéines sanguines destinée à l'identification des organismes hôtes des phlébotomes, vecteurs de la leishmaniose. Grâce à une approche protéomique combinée à la spectrométrie de masse, nous avons pu identifier avec précision les protéines présentes dans l'intestin des phlébotomes après un repas sanguin. L'utilisation d'une base de données compacte, basée sur SwissProt et digérée in silico, a non seulement réduit le temps d'analyse, mais a également augmenté la sensibilité de l'identification des peptides, y compris ceux ayant un spectre de faible intensité.

Les résultats obtenus démontrent que cette méthodologie est efficace pour identifier les hôtes des phlébotomes, ce qui est crucial pour comprendre les dynamiques de transmission de la leishmaniose. De plus, les analyses comparatives entre les peptides théoriques et expérimentaux ont permis de cibler les protéines les plus pertinentes pour de futures recherches.

En conclusion, ce projet pose les bases pour des applications futures dans le contrôle des populations de phlébotomes et la prévention de la leishmaniose. La création d'une base de données spécifique aux protéines sanguines des hôtes offre un outil précieux pour des analyses rapides et précises, et pourrait être étendue à d'autres vecteurs de maladies. Ces résultats ouvrent la voie à de nouvelles stratégies de lutte contre la leishmaniose, contribuant ainsi à réduire l'impact de cette maladie dans les régions endémiques.

6. Bibliographie

[1] Huebner, Erwin. Fichier:Rhodnius prolixus70-300.jpg — Wikipédia. 2 septembre 2009, https://commons.wikimedia.org/wiki/File:Rhodnius_prolixus70-300.jpg.

[2] *Principaux repères sur la leishmaniose.* 2022. World Health Organization. https://www.who.int/fr/news-room/fact-sheets/detail/leishmaniasis. Consulté le 2 juin 2024.

[3] Torres-Guerrero, Edoardo, et al. *Leishmaniasis: A Review*. 6:750, F1000Research, 26 mai 2017. *f1000research.com*, https://doi.org/10.12688/f1000research.11120.1.

[4] *Control of the Leishmaniases WHO TRS N° 949*. https://www.who.int/publications-detail-redirect/WHO-TRS-949. Consulté le 2 juin 2024.

[5] *Global Leishmaniasis Surveillance: 2019–2020, a Baseline for the 2030 Roadmap*. https://www.who.int/publications-detail-redirect/who-wer9635-401-419. Consulté le 3 juin 2024.

[6] Bhunia, Gouri Sankar, et Pravat Kumar Shit. « Introduction of Visceral Leishmaniasis (Kala-Azar) ». *Spatial Mapping and Modelling for Kala-Azar Disease*, Springer, Cham, 2020, p. 1-18. *link.springer.com*, https://doi.org/10.1007/978-3-030-41227-2_1.

[7] Smokefoot. *English: antimony gluconic acid complex, improved structure*. 24 janvier 2013. Travail personnel, *Wikimedia Commons*, https://commons.wikimedia.org/wiki/File:ImprovedSbgluconicAcid.png?uselang=fr.

[8] Clar, Bob Saint. *English: Structure of meglumine antimoniate.* mai , 09:18:20 2014. Travail personnel, *Wikimedia Commons*, https://commons.wikimedia.org/wiki/File:Meglumine_antimoniate.png?uselang=fr

[9] PubChem. Amphotericin b Deoxycholate. https://pubchem.ncbi.nlm.nih.gov/compound/23668620. Consulté le 6 mars 2024.

[10] *Sodium Stibogluconate*. https://go.drugbank.com/drugs/DB05630. Consulté le 4 juin 2024.

[11] Feng, Mei, et al. « Sterol Profiling of Leishmania Parasites Using a New HPLC-Tandem Mass Spectrometry-Based Method and Antifungal Azoles as Chemical Probes Reveals a Key Intermediate Sterol That Supports a Branched Ergosterol Biosynthetic Pathway ». *International Journal for Parasitology: Drugs and Drug Resistance*, vol. 20, décembre 2022, p. 27-42. *DOI.org (Crossref)*, https://doi.org/10.1016/j.ijpddr.2022.07.003.

[12] *Amphotericin B*. https://go.drugbank.com/drugs/DB00681. Consulté le 4 juin 2024.

[13] Akhoundi, Mohammad, et al. « A Historical Overview of the Classification, Evolution, and Dispersion of Leishmania Parasites and Sandflies ». *PLoS Neglected Tropical Diseases*, vol. 10, nº 3, mars 2016, p. e0004349. *PubMed Central*, https://doi.org/10.1371/journal.pntd.0004349.

[14] « Trypanosomatidae ». *Wikipédia*, 16 septembre 2023. *Wikipedia*, https://fr.wikipedia.org/w/index.php?title=Trypanosomatidae&oldid=207887513.

[15] Alemayehu, Bereket, et Mihiretu Alemayehu. « Leishmaniasis: A Review on Parasite, Vector and Reservoir Host ». *Health Science Journal*, vol. 11, nº 4, 2017. *DOI.org (Crossref)*, https://doi.org/10.21767/1791-809X.1000519.

[16] Mann, Sarah, et al. « A Review of Leishmaniasis: Current Knowledge and Future Directions ». *Current Tropical Medicine Reports*, vol. 8, nº 2, 2021, p. 121-32. *PubMed Central*, https://doi.org/10.1007/s40475-021-00232-7.

[17] Chisha, Paulos. (2019). REVIEW ON LEISHMANIASIS. Asian Journal of Animal and Veterinary Advances

[18] Mishra, Jyotsna, et al. « Chemotherapy of Leishmaniasis: Past, Present and Future ». *Frontiers in Medicinal Chemistry*, édité par Atta-ur-Rahman et al., BENTHAM SCIENCE PUBLISHERS, 2012, p. 97-130. *DOI.org (Crossref)*, https://doi.org/10.2174/978160805464011306000 7.

[19] « Leishmaniosis in Dogs - Generalized Conditions ». *MSD Veterinary Manual*, https://www.msdvetmanual.com/generalized-conditions/leishmaniosis/leishmaniosis-in-dogs.

[20] Lewis, D. J. « Phlebotomid Sandflies ». *Bulletin of the World Health Organization*, vol. 44, nº 4, 1971, p. 535-51.

[21] Maroli, M., et al. « Phlebotomine Sandflies and the Spreading of Leishmaniases and Other Diseases of Public Health Concern ». *Medical and Veterinary Entomology*, vol. 27, nº 2, juin 2013, p. 123-47. *DOI.org (Crossref)*, https://doi.org/10.1111/j.1365-2915.2012.01034.x.

[22] *Phlebotomine Sand Flies - Factsheet for Experts*. 19 juin 2017, https://www.ecdc.europa.eu/en/disease-vectors/facts/phlebotomine-sand-flies.

[23] Bongiorno, G., et al. « Host Preferences of Phlebotomine Sand Flies at a Hypoendemic Focus of Canine Leishmaniasis in Central Italy ». *Acta Tropica*, vol. 88, nº 2, octobre 2003, p. 109-16. *PubMed*, https://doi.org/10.1016/s0001-706x(03)00190-6.

[24] Photographies réalisées par Sabrina. Bousbata et Sofia El Kacem le 29 mai 2024.

[25] Hlavackova, Kristyna, et al. « A novel MALDI-TOF MS-based method for blood meal identification in insect vectors: A proof of concept study on phlebotomine sand flies ». *PLoS Neglected Tropical Diseases*, vol. 13, nº 9, septembre 2019, p. e0007669. *PubMed Central*, https://doi.org/10.1371/journal.pntd.0007669.

[26] « PASEF: Redefining New Standards for Proteomics Research ». *News-Medical*, 15 mars 2019, https://www.news-medical.net/whitepaper/20190315/PASEF-Redefining-New-Standards-for-Proteomics-Research.aspx.

[27] Cox, Jürgen, et al. « Andromeda: A Peptide Search Engine Integrated into the MaxQuant Environment ». *Journal of Proteome Research*, vol. 10, nº 4, avril 2011, p. 1794-805. *DOI.org (Crossref)*, https://doi.org/10.1021/pr101065j.

[28] Ostasiewicz, Paweł, et al. « Proteome, Phosphoproteome, and N-Glycoproteome Are Quantitatively Preserved in Formalin-Fixed Paraffin-Embedded Tissue and Analyzable by High-Resolution Mass Spectrometry ». *Journal of Proteome Research*, vol. 9, nº 7, juillet 2010, p. 3688-700. *DOI.org (Crossref)*, https://doi.org/10.1021/pr100234w.

[29] Paul, Joseph, et Timothy D. Veenstra. « Separation of Serum and Plasma Proteins for In-Depth Proteomic Analysis ». *Separations*, vol. 9, nº 4, avril 2022, p. 89. *www.mdpi.com*, https://doi.org/10.3390/separations9040089.

7. Annexes

Annexe 1 : Récupération des séquences de protéines sanguines et décompte de leur nombre

lucinux /data/tfe/blood_db # cat ../all_entry/blood_prot

```
sp-id:HBA*
sp-id:HBB*
sp-id:ALBU*
sp-id:TFR1*
sp-id:TFR2*
sp-id:TRFE*
sp-id:FIBA*
sp-id:FIBB*
sp-id:FIBG*
sp-id:THRB*
sp-id:TF_*
sp-id:FA7_*
sp-id:FA8_*
sp-id:FA9_*
sp-id:FA10_*
sp-id:FA11_*
sp-id:FA12_*
sp-id:IGHG_*
sp-id:APO*
sp-id:A2MG_*
sp-id:A1AT*
```

```
for id in $(cat /data/tfe/all_entry/blood_prot)
do
        name=$(echo $id | sed -e 's/^sp-id://' | sed -e 's/\*$//')
        echo $id
```

```
        echo $name
        entret $id $name.dat
        grep -c ^ID $name.dat > $name.count
done
```

Annexe 2 : Parseur pour les fichiers « .pepdigest »

```
#!/usr/bin/env python
# -*- coding: utf-8 -*-

import sys

with open(sys.argv[1], "r") as pep_file:
   text = pep_file.readlines()

peptides = {}

for line in text:
   if line.startswith("# Sequence:"):
      seq_id = line.split()[2]
      continue
   if line.startswith("#"): continue
   if line == "\n": continue
   if line.startswith("   Start"): continue
   peptide = line.split()[5]
   if peptide not in peptides.keys():
      peptides[peptide] = []
   peptides[peptide].append(seq_id)

# Pour tous les peptides (peu importe la taille)
# for pept in peptides.keys(): #Décommenter la ligne adéquate
   # Récupère uniquement les peptides uniques à une seule protéine
   # if len(peptides[pept]) == 1: print(pept, ".join(peptides[pept]))

   # Récupère le nombre d'identifiants par peptide
   # print(len(peptides[pept]), pept, peptides[pept])

# Pour les peptides de minimum 7 a.a
for pept in peptides.keys():
        if len(peptides[pept]) == 1 and len(pept) >= 7:print(pept, ".join(peptides[pept]))

   # if len(pept) >= 7:
      # print(len(peptides[pept]), pept, peptides[pept])
```

Annexe 3 : Récupération du nom scientifique des organismes dans la banque de données

```
for org in $(sed -e 's/^.*\s//g' nombre_pep_uniq_par_espece);
do
        entret blood:*_$org stdout 2>/dev/null | grep "^OS" | sort | uniq | sed -e 's/^OS\s\+//' | sed -e 's/
        (.*$//' >> nouveau_org_names
done
```

Le fichier doit être manuellement modifé car certains noms sont trop longs et utilisent deux lignes. Il faut enlever toutes les lignes qui ne commencent pas par une majuscule.

paste nombre_pep_uniq_par_espece nouveau_org_names > nom_espèce_nombre_pep_uniq

Annexe 4 : Code permettant de créer le tableau à double entrées du dénombrement des peptides par protéine et par espèce

```
#!/usr/bin/env python
# -*- coding: utf-8 -*-

import pandas as pd
import sys
import subprocess

# Fonction pour lire le fichier et créer un DataFrame
def read_file(file_path):
    data = {}

    with open(file_path, 'r') as f:
        for line in f:
            parts = line.split()
            if len(parts) < 3:
                continue
            animal_id = parts[0]
            protein_id = parts[1]

            # Initialiser les sous-dictionnaires si nécessaire
            if protein_id not in data:
                data[protein_id] = {}
            if animal_id not in data[protein_id]:
                data[protein_id][animal_id] = 0

            # Incrémenter le compteur de peptides pour la paire (animal_id, protein_id)
            data[protein_id][animal_id] += 1

    return data

# Fonction pour obtenir les noms des organismes
def get_org_names(file_path):
    org_names = {}
    with open(file_path, 'r') as f:
        for line in f:
            parts = line.split(maxsplit=2)
            if len(parts) < 3:
                continue
            org_id = parts[1]
            name = parts[2].strip()
            org_names[org_id] = name

    return org_names
```

```
# Créer le DataFrame à partir des données
def create_dataframe(data, org_names):
    df = pd.DataFrame(data).fillna(0).astype(int)

    # Transposer le DataFrame pour que les identifiants des animaux soient en colonnes
    df = df.T

    # Remplacer les noms des colonnes par les noms des organismes
    new_columns = [org_names[x] for x in df.columns]
    df.columns = new_columns

    # Ajouter les totaux pour chaque ligne et colonne
    df['Total'] = df.sum(axis=1)

    return df

# Chemin vers le fichier de données
if len(sys.argv) != 3:
    print("Usage: python script.py <gen_pept_file> <nb_uniq_pep_par_org_file>")
    sys.exit(1)

gen_pep_file = sys.argv[1]
count_per_org_file = sys.argv[2]

# Lire le fichier et créer le DataFrame
data = read_file(gen_pep_file)
org_names = get_org_names(count_per_org_file)
df = create_dataframe(data, org_names)

# Afficher le DataFrame
print(df)

# Sauvegarder le DataFrame dans un fichier CSV
df.to_csv(f'{gen_pep_file.split("_")[-1]}.csv')
```

Annexe 5 : Script filter_peptide.py

```
import pandas as pd

# Chemin vers le fichier peptides.txt
input_file = 'txt/peptides.txt'

print(f"Récupération et filtrage des données : {input_file}")

# Charger les données
peptides_df = pd.read_csv(input_file, sep='\t')

# Filtrer les données dans peptides.txt
peptides_filtered = peptides_df[(peptides_df['Reverse'] != '+') & (peptides_df['Potential
contaminant'] != '+') & (peptides_df['Unique (Proteins)'] == 'yes')]

intensity_columns = [col for col in peptides_df.columns if col.startswith('Intensity ')]
experiment_columns = [col for col in peptides_df.columns if col.startswith('Experiment ')]
```

```
# Sélectionner les colonnes pertinentes
selected_columns = intensity_columns + ['Sequence', 'Length', 'Mass', 'Proteins', 'Charges', 'PEP',
'Score', 'MS/MS Count', 'id'] + experiment_columns

# Créer un DataFrame filtré avec les colonnes sélectionnées
peptides_filtered = peptides_filtered[selected_columns]

# Chemin pour enregistrer la matrice filtrée
output_file = 'data-filtered/filtered_peptides'

# Exporter la matrice filtrée
peptides_filtered.to_csv(f"{output_file}.txt", sep='\t', index=False)
peptides_filtered.to_excel(f"{output_file}.xlsx", index=False)

print("Matrice peptides.txt filtrée et exportée avec succès.")
```

Annexe 6 : Fichier fetch_upids.py permettant de récupérer les informations des protéines séquencées via l'API d'Uniprot

```
import requests
import pandas as pd
import sys
import time
import json

def get_excel_file():
    try:
        return sys.argv[1]
    except IndexError:
        print("Please enter an Excel file")
        sys.exit()

def read_excel_file(file):
    try:
        return pd.read_excel(file)
    except Exception as e:
        print(f"An error occurred: {e}")
        sys.exit()

def submit_uniprot_job(ids):
    url = "https://rest.uniprot.org/idmapping/run"
    query = {
        'from': 'UniProtKB_AC-ID',
        'to': 'UniProtKB',
        'ids': ','.join(ids)
    }
    response = requests.post(url, data=query)
    response.raise_for_status()
    return response.json()

def check_job_status(job_id, max_attempts=10, wait_time=5):
    status_url = f"https://rest.uniprot.org/idmapping/status/{job_id}"
```

```
    attempt = 0
    while attempt < max_attempts:
        status_response = requests.get(status_url)
        status_response.raise_for_status()
        status_json = status_response.json()
        if 'results' in status_json:
            return status_json
        elif 'jobStatus' in status_json and status_json['jobStatus'] == 'RUNNING':
            print('Job status: ', status_json['jobStatus'])
            print('Nouvel essai dans 5 secondes...')
            time.sleep(wait_time)
        elif 'jobStatus' in status_json and status_json['jobStatus'] == 'FINISHED':
            return status_json
        elif 'jobStatus' in status_json and status_json['jobStatus'] == 'FAILED':
            raise Exception("ID mapping job failed.")
        else:
            print(
                f"Attempt {attempt + 1} failed. Job status: {status_json.get('jobStatus', 'Unknown')}.
Retrying in {wait_time} seconds...")
            attempt += 1
            time.sleep(wait_time)
    raise Exception("Maximum attempts reached, job not completed.")

def fetch_job_results(job_id):
    result_url = f'https://rest.uniprot.org/idmapping/uniprotkb/results/{job_id}'
    result_response = requests.get(result_url)
    result_response.raise_for_status()
    return result_response.json()

def extract_protein_data(result_json):
    results = []
    for item in result_json['results']:
        to = item['to']
        organism = to['organism']
        protein_description = to['proteinDescription']
        if 'recommendedName' in protein_description:
            protein_name = protein_description['recommendedName']['fullName']['value']
        elif 'submissionNames' in protein_description and protein_description['submissionNames']:
            protein_name = protein_description['submissionNames'][0]['fullName']['value']
        else:
            protein_name = 'Unknown'

        results.append({
            'UniProt ID': to['uniProtkbId'],
            'Primary Accession': to['primaryAccession'],
            'Organism Common Name': organism.get('commonName', 'N/A'),
            'Organism Scientific Name': organism.get('scientificName', 'N/A'),
            'Molecular Weight': to['sequence'].get('molWeight', 'N/A'),
            'Protein Length': to['sequence'].get('length', 'N/A'),
            'Protein Name': protein_name,
            'Protein Sequence': to['sequence'].get('value', 'N/A')
        })
    return results

def save_results_to_excel(results, output_file):
```

```
    df = pd.DataFrame(results)
    df.to_excel(output_file, index=False)
    print(f'Data has been saved to {output_file}')

def main():
    excel_file = get_excel_file()
    print(f"Début de la récupération des identifiants du fichier {excel_file}")

    data = read_excel_file(excel_file)
    ids = data['Proteins'].unique()
    # print(f"Liste des identifiants pour le fichier {excel_file} : {ids}")
    print(f"Nombre d'identifiants demandés : {len(ids)}")

    # Split IDs into chunks if necessary
    chunk_size = 20
    id_chunks = [ids[i:i + chunk_size] for i in range(0, len(ids), chunk_size)]

    all_results = []
    for chunk in id_chunks:
        # print(f"Processing chunk: {chunk}")
        try:
            response_json = submit_uniprot_job(chunk)
            job_id = response_json['jobId']

            status_json = check_job_status(job_id)
            with open(f"{excel_file.split('.')[-2] + '_api_status.json'}", "w") as file:
                json.dump(status_json, file, indent=4)

            result_json = fetch_job_results(job_id)
            with open(f"{excel_file.split('.')[-2] + '_api_result.json'}", "w") as file:
                json.dump(result_json, file, indent=4)

            results = extract_protein_data(result_json)
            all_results.extend(results)
        except Exception as e:
            print(f"An error occurred with chunk {chunk}: {e}")
            continue

    print(f"Nombre d'identifiants récupérés : {len(all_results)}")
    output_file = f"data-filtered/idmapping_{excel_file.split('/')[-1]}"
    save_results_to_excel(all_results, output_file)

if __name__ == "__main__":
    main()
```

Annexe 7 : Fichier merge_upids.py permettant la fusion du fichier peptide.txt avec les informations récupérées via l'API d'Uniprot

```
import pandas as pd
import sys

def read_excel_file(file_path):
    try:
```

```
        return pd.read_excel(file_path)
    except Exception as e:
        print(f"An error occurred while reading {file_path}: {e}")
        sys.exit()

def check_column_values(data_df, idmapping_df):
    missing_ids = set(data_df['Proteins']) - set(idmapping_df['Primary Accession'])
    if missing_ids:
        print(f"Identifiants manquants dans le fichier de mapping : {missing_ids}")

def merge_data(data_file, idmapping_file, output_file):
    data_df = read_excel_file(data_file)
    idmapping_df = read_excel_file(idmapping_file)

    # print("Columns in data_df:", data_df.columns)
    # print("Columns in idmapping_df:", idmapping_df.columns)

    # print(f"{data_file} DataFrame:")
    # print(data_df.head())
    # print(f"\n{idmapping_file} DataFrame:")
    # print(idmapping_df.head())

    # Check for missing identifiers before merging
    check_column_values(data_df, idmapping_df)

    # Convert the identifiers to string to ensure the merge works correctly
    data_df['Proteins'] = data_df['Proteins'].astype(str)
    idmapping_df['Primary Accession'] = idmapping_df['Primary Accession'].astype(str)

    # Merge dataframes on the protein identifiers
    merged_df = data_df.merge(idmapping_df, left_on='Proteins', right_on='Primary Accession',
how='left')
    merged_df.drop(columns=['Proteins'], inplace=True)

    merged_df.to_excel(output_file, index=False)
    print(f"Fichier fusionné et exporté avec succès vers {output_file}")

def main():
    try:
        file_type = sys.argv[1]
    except IndexError:
        print("Usage: python merge_upids.py <file_type (evidence or peptides)>")
        sys.exit()

    data_file = f'data-filtered/filtered_{file_type}.xlsx'
    idmapping_file = f'data-filtered/idmapping_filtered_{file_type}.xlsx'
    output_file = f'data-filtered/merged_{file_type}.xlsx'

    print(f'Fusion des tableaux {data_file} et {idmapping_file} :')
    merge_data(data_file, idmapping_file, output_file)

if __name__ == "__main__":
    main()
```