Université de Montréal

Identification des peptides du complexe majeur d'histocompatibilité de classe I par spectrométrie de masse

par
Alexandre Bramoullé

Département de biochimie
Faculté des arts et des sciences

Mémoire présenté à la Faculté des études supérieures
en vue de l'obtention du grade de Maître ès sciences (M.Sc.)
en bioinformatique

Décembre, 2010

Université de Montréal
Faculté des études supérieures

Ce mémoire intitulé:

Identification des peptides du complexe majeur d'histocompatibilité de classe I par spectrométrie de masse

présenté par:

Alexandre Bramoullé

a été évalué par un jury composé des personnes suivantes:

Dr Nicolas Lartillot,	président-rapporteur
Dr Pierre Thibault,	directeur de recherche
Dr Sébastien Lemieux,	codirecteur
Dr Naglaa Shoukry,	membre du jury

Mémoire accepté le: . . . . . .

RÉSUMÉ

L'immunité adaptive et la discrimination entre le soi et le non-soi chez les vertébrés à mâchoire reposent sur la présentation de peptides par les récepteurs d'histocompatibilité majeur de classe I. Les peptides antigéniques, présentés par les molécules du complexe d'histocompatibilité (CMH), sont scrutés par les lymphocytes T CD8 pour une réponse immunitaire appropriée. Le répertoire des peptides du CMH de classe I, aussi appelé immunopeptidome, est généré par la dégradation protéosomale des protéines endogènes, et a un rôle essentiel dans la régulation de l'immunité cellulaire. La composition de l'immunopeptidome dépend du type de cellule et peut présenter des caractéristiques liées à des maladies comme le cancer. Les peptides antigéniques peuvent être utilisés à des fins immunothérapeutiques notamment dans le traitement voire la prévention de certains cancers. La spectrométrie de masse est un outil de choix pour l'identification, le séquençage et la caractérisation de ces peptides. Cependant, la composition en acides aminés, la faible abondance et la diversité de ces peptides compliquent leur détection et leur séquençage. Nous avons développé un programme appelé StatPeaks qui permet de calculer un certains nombres de statistiques relatives à la fragmentation des peptides. À l'aide de ce programme, nous montrons sans équivoque que les peptides du CMH classe I, en mode de fragmentation par dissociation induite par collision (CID), fragmentent très différemment des peptides trypsiques communément utilisés en protéomique. Néanmoins, la fragmentation par décomposition induite par collision à plus haute énergie (HCD) proposée par le spectromètre LTQ-Orbitrap Velos améliore la fragmentation et fournit une haute résolution qui permet d'obtenir une meilleure confiance dans l'identification des peptides du CMH de classe I. Cet avantage permet d'effectuer le séquençage *de novo* pour identifier les variants polymorphes qui ne sont normalement pas identifiés par les recherches utilisant des bases de données. La comparaison des programmes de séquençage Lutefisk, pepNovo, pNovo, Vonode et Peaks met en évidence que le dernier permet d'identifier un plus grand nombre de peptides du CMH de classe I. Ce programme est intégré dans une chaîne de traitement de recherche d'antigènes mineurs d'histocompatibilité. Enfin, une base de données contenant les informations spectrales de plusieurs centaines de peptides du CMH de classe I accessible par Internet a été développée.

Mots clés : antigènes, CMH de classe I, immunopeptidome, spectrométrie de masse, séquençage *de novo*, polymorphisme mononucléotidique.

ABSTRACT

Adaptive immunity and discrimination between self and nonself in jawed vertebrates relies on the presentation of peptides by the major histocompatibility (MHC) class I receptors. Foreign or self peptide antigens presented by the MHC molecules are probed by CD8 T-cell lymphocyte for proper immune response. The repertoire of MHC I peptides collectively referred to as the immunopeptidome is generated through the proteasomal degradation of endogenous proteins and plays an important role in the regulation of cellular immunity. The composition of the immunopeptidome is cell specific and can harbor important hallmark of human diseases including cancer. Antigenic peptides can also be used in immunotherapy to mount an appropriate immune response against cancer cells displaying these peptides. Mass spectrometry is a tool of choice for the identification, sequencing and characterization of these peptides. However, the amino acid composition, the low abundance and diversity of these peptides make their detection and sequencing more challenging. We developed a software, called StatPeaks, that calculates statistics relative to the fragmentation of peptides. Using this software, we demonstrate that under collision induced dissociation (CID) MHC class I peptides fragment in a very different fashion than tryptic peptides, commonly used in proteomics. However, the higher-energy collisional dissociation (HCD) mode available on the LTQ-Orbitrap Velos enhances peptide fragmentation and provides high resolution fragment information that significantly improves the confidence in MHC class I peptide identification. This inherent advantage confers the ability to perform de novo sequencing to identify polymorphic variants that would normally elude conventional database searches. The comparison of de novo peptide sequencing software Lutefisk, pepNovo, pNovo, Vonode and Peaks indicated that the later software enabled higher rates of correct identification for MHC class I peptides. This software was integrated into a data analysis pipeline for the identification minor histocompatibility antigens (MiHAs). A web-based library that stores spectral information of hundreds of synthetic MHC class I peptides was developed in support to the needs of the immunopeptidome discovery program.

Keywords: antigen, MHC Class I, immunopeptidome, mass spectrometry, *de novo* sequencing, single nucleotide polymorphism.

TABLE DES MATIÈRES

LISTE DES TABLEAUX

LISTE DES FIGURES

LISTE DES ANNEXES

LISTE DES SIGLES

AgMH	Antigènes Mineurs d'Histocompatibilité
BLAST	Basic Local Alignment Search Tool
CMH-I	Complexe majeur d'histocompatibilité de classe I
C-trap	C-shaped storage trap
CD8	Cluster of Differentiation 8
CD8+	Cytotoxic T cells with CD8 surface protein
CID	Collision-Induced Dissociation
CMH-I	Complexe Majeur d'Histocompatibilité de Classe I
CTL	Cytotoxic T Cell
ESI	Electrospray Ionization
FAB	Fast Atom Bombardment
GVH	Graft Versus Host
GVL	Graft Versus Leukemia
HCD	Higher-energy C-trap Dissociation
HLA	Human Leukocyte Antigen
CLHP	Chromatographie en phase liquide à haute performance
IC50	Median Inhibition Concentration
IED	Immune Epitope Database

IPI	International Protein Index
LSIMS	Liquid Secondary Ion Mass Spectrometry
LTQ	Linear Trap Quadrupole
MALDI	Matrix-assisted laser desorption/ionization
MGF	Mascot Generic Format
MHCDB	Major Histocompatibility Complex DataBase
MS/MS	Spectrométrie de masse en tandem
MVC	Modèle Vue Contrôleur
MySQL	(My) Structured Query Language
nanoLC-MS	nano Liquid Chromatography Mass Spectrometry
m/z	Masse divisée par la charge
ORF	Open Reading Frame
ORM	Object-relational mapping
PFF	Peptide Fragment Fingerprint
RE	Réticulum Endoplasmique
ROC	Receiver Operating Characteristic
SARS	Severe Acute Respiratory Syndrome
SNP	Single-Nucleotide Polymorphism
TAP	Transporter Associated with Antigen Processing
TCR	T Cell Receptor

CHAPITRE 1

INTRODUCTION

1.1 Les objectifs

Les peptides antigéniques du complexe d'histocompatibilité majeur de classe I ont une importance cruciale dans le système immunitaire et sont impliqués dans plusieurs maladies. De nombreuses études et approches thérapeutiques reposent sur l'identification et la caractérisation de ces peptides. La spectrométrie de masse est devenue l'outil de choix pour leur analyse. Cependant, la composition en acides aminés, la faible abondance et la diversité de ces peptides compliquent leur détection et leur séquençage. En premier lieu, un programme permettant de calculer un certain nombre de statistiques relatives à la fragmentation des peptides a été développé. Ensuite, une description des caractéristiques de fragmentation des peptides du CMH de classe I a été faite à l'aide d'une librairie de plusieurs centaines de spectres de peptides synthétiques. Nous avons comparé les spectres issus de deux modes de fragmentation différents, CID et HCD. Après avoir démontré le gain d'information apporté par le second, nous avons évalué les performances de 5 programmes de séquençage *de novo* avec les peptides du CMH-I. Le programme montrant les meilleures performances a été retenu pour intégrer une chaîne de traitement de recherche d'antigènes mineurs d'histocompatibilité. Pour terminer, on décrit brièvement la base de données MHCDB qui contient un ensemble de peptides du CMH de classe I associés à leurs spectres de masse.

1.2 L'immunité acquise et le complexe d'histocompatibilité majeur de classe I

Il est fort à parier que dès l'apparition de la vie sur Terre, celle-ci a été en émulation avec elle-même. Il n'y a pas de forme de vie qui ne soit pas en compétition avec une autre. Très tôt les organismes ont dû mettre en oeuvre des stratégies pour se protéger de leurs compétiteurs, y compris les plus petits, et assurer leur pérennité. La survie de l'espèce humaine et de bien d'autres espèces repose en grande partie sur leur système de défense immunitaire qui est le fruit de millions d'années d'évolution. Il faudra sans conteste de nombreuses années avant de découvrir tous les arcanes de ce système extrêmement développé. Parmi la plupart des espèces animales et notamment les mammifères, le système immunitaire repose sur deux composantes. L'immunité innée, la plus ancienne, est basée sur la reconnaissance d'éléments caractéristiques des pathogènes. Elle déclenche une réaction inflammatoire. L'immunité acquise, plus récente dans l'évolution, est capable d'apprendre en détail la composition moléculaire de l'organisme tout entier (le soi), et reconnaît de façon très spécifique tout élément étranger (non-soi). Ce dernier peut être un virus ou encore à une protéine mutée pouvant conduire à un cancer. Dans le cas de l'immunité acquise à laquelle on s'intéresse dans notre étude, la réaction immunitaire est initiée par la présence d'une cellule cible infectée ou néoplasique avec un lymphocyte T cytotoxique (CTL) qui exprime le marqueur CD8+. Son rôle est d'identifier, via les récepteurs TCR (T Cell Receptor), les cellules présentant à leur surface des fragments peptidiques provenant de pathogènes ou de protéines découlant de transformations malignes (figure 1.1). Ces fragments, aussi appelés peptides antigéniques, sont fixés et présentés par les molécules du *complexe majeur d'histocompatibilité de classe I* (CMH-I) [73]. Ces peptides antigéniques sont issus de protéines sources dégradées par le protéasome. Les peptides résultants sont internalisés par le canal TAP (Transporter

Associated with Antigen Processing) du réticulum endoplasmique puis clivés par des aminopeptidases pour générer des peptides de 8 à 11 acides aminés qui s'associeront par la suite aux récepteurs du CMH-I.

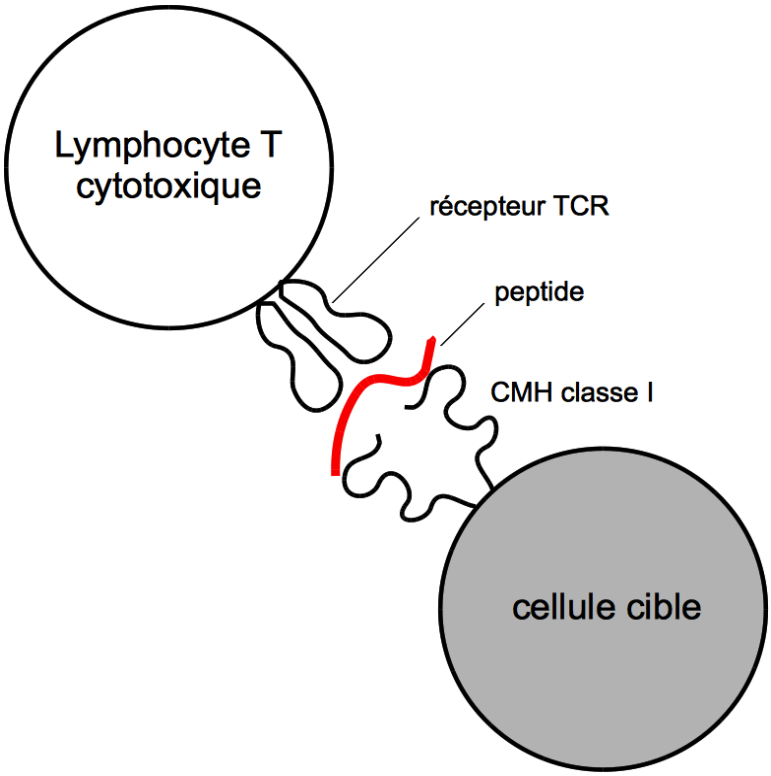


Figure 1.1 – Schéma de présentation des antigènes peptidiques par les molécules du CMH de classe I au récepteur TCR. Le lymphocyte T cytotoxique interagit avec la cellule cible à travers la liaison de son récepteur TCR avec le complexe d'histocompatibilité majeur de classe I sur lequel est lié le peptide.

Dans le cas de cellules saines, exemptes d'infection virale par exemple, ce sont des fragments de protéines du soi qui sont présentés et aucune réponse immunitaire n'est normalement déclenchée. En revanche si un peptide n'est pas reconnu comme appartenant au soi, l'activation des CTL conduit à la lyse de la cellule afin que celle-ci ne se multiplie pas ou que l'infection se ne propage pas à ses voisines.

1.2.1 Le complexe d'histocompatibilité majeur de classe I

Chez l'homme, le complexe d'histocompatibilité majeur (CMH) est aussi appelé HLA (de l'anglais Human Leucocyte Antigen). Le CMH est composé d'une famille de protéines comprenant trois classes de gènes :

1. Les gènes de classe I codent la chaîne α des molécules du CMH-I. Ces chaînes lourdes, associées à la β2-microglobuline, sont reconnues par les lymphocytes T CD8+. Les produits des loci HLA-A, HLA-B, HLA-C pour l'humain et H-2k et H-2b pour la souris sont quasiment ubiquitaires et ont un grand polymorphisme allélique. Chez l'humain, ces gènes sont situés sur le chromosome 6.

2. Les gènes de classe II codent les chaîne α et β des molécules de CMH-II. Ces gènes sont exprimés dans un nombre restreint de cellules : les monocytes et les lymphocytes B. Ces cellules se chargent avant tout de la présentation des antigènes exogènes.

3. Les gènes de classe III contrairement aux gènes précédents ne sont pas exprimés à la surface de la cellule. Ils constituent un ensemble hétérogène de gènes dont certains ont néanmoins un lien avec le système immunitaire comme par exemple ceux codant les sous-unités TAP1 et TAP2, et d'autres codant des sous-unités du protéasome.

Ceux auxquels on s'intéresse sont les gènes de classe I. Les protéines du CMH-I se lient à des fragments de protéines et sont présentées à la surface des cellules nucléées. L'ensemble du répertoire des peptides présentés par les récepteurs du CMH-I décrit l'immunopeptidome d'une cellule à un moment donné. Le système de présentation des peptides du CMH-I :

1. produit des peptides d'une longueur comprise entre 8 et 11 acides aminés qui peuvent se loger dans le sillon des molécules de classe I ;
2. traitent des peptides de séquences variées qui représentent la diversité des motifs de séquence des molécules de classe I ;
3. et, est rapide pour permettre une réponse immunitaire aussi prompte que possible afin, par exemple, d'éviter la prolifération des pathogènes dans l'organisme.

1.2.1.1 La composition du complexe d'histocompatibilité majeur de classe I

Le complexe du CMH de classe I est l'association de différents partenaires :

1. la chaîne lourde α,
2. la β2-microglobuline,
3. et, un peptide de 8 à 11 acides aminés.

Tous les partenaires sont indispensables pour la stabilité de la molécule. En l'absence d'un seul d'entre eux, la molécule ne migre pas à la surface de la cellule. La molécule se compose d'un pied et d'une tête. Le pied, formé du domaine α3 de la chaîne lourde, assure l'ancrage dans la membrane plasmique qu'il traverse de part en part. La tête est formée des domaines α1 et α2 de la chaîne lourde (figure 1.2) [40].

Les domaines α1 et α2, constitués de 8 plis β anti-parallèles, forment un plateau. Deux hélices α constituent un sillon d'une longueur de 25 Å et d'une largeur de 10 Å. Les peptides présentés aux lymphocytes T viennent se loger dans ce sillon. Nombre de résidus polymorphes, différenciant les allèles les uns de autres, se situent sur la face interne des hélices α ou sur le plancher du sillon. Les peptides fixés par un allèle diffèrent des peptides fixés par un autre allèle.

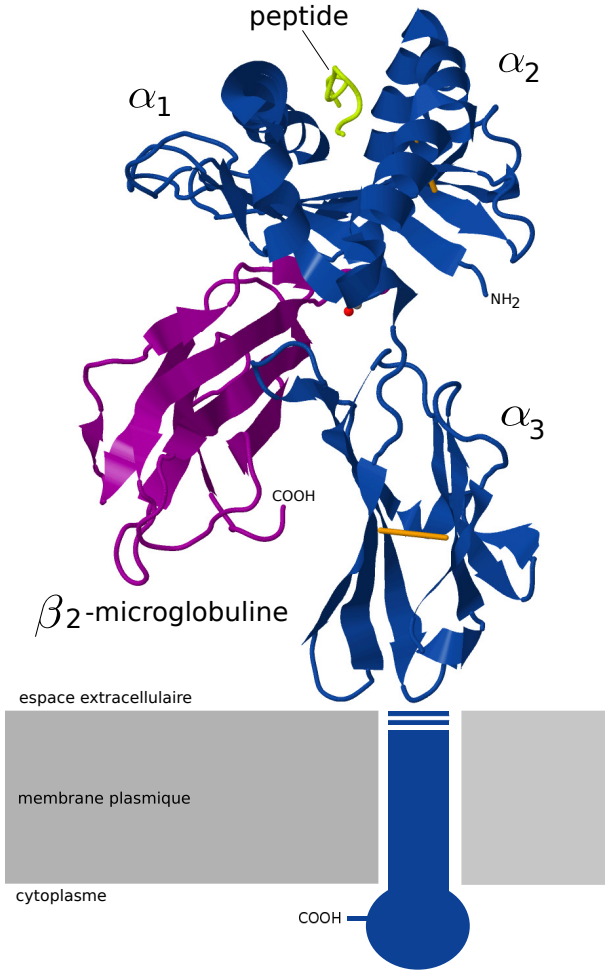


Figure 1.2 – La structure 3D de la molécule du CMH-I HLA-A*2 avec le peptide antigénique WT1 (126-134 R1Y). La chaîne α de la molécule du CMH-I a trois domaines extracellulaires α1, α2 et α3. Elle est liée de façon no-covalente à une chaîne polypeptide, la β2-microglobuline. Alors que cette dernière est invariable, la chaîne α est extrêmement polymorphique, surtout dans les domaines α1, α2. Ces domaines forment un sillon, ou autrement appelé une cavité, dans lequel vient se loger le peptide présenté aux lymphocytes T cytotoxiques. La molécule du CMH-I traverse la membrane plasmique. Les ponts disulfures sont représentés en orange. *(Figure dessinée à l'aide de Inkscape et Jmol à partir du fichier PDB : DOI :10.2210/pdb3myj/pdb).*

1.2.1.2 Les motifs du complexe d'histocompatibilité majeur de classe I

Les peptides antigéniques qui se lient à une isoforme de CMH-I particulière présentent à certaines positions des acides aminés identiques ou très similaires. Ceux-ci

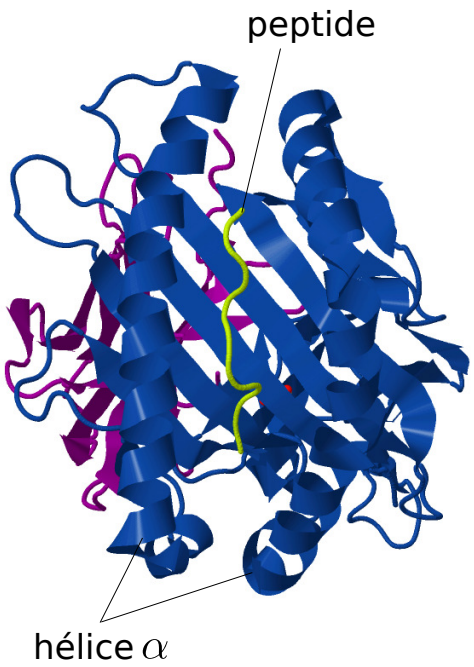


Figure 1.3 – La structure 3D de la molécule du CMH-I HLA-A*2 avec le peptide antigénique logé dans le sillon vu du dessus. Le peptide présenté aux lymphocytes T cytotoxiques vient se loger dans le sillon formé par les deux chaînes α. Ces peptides font dans la plupart des cas une longueur de 8 à 11 résidus. *(Figure dessinée à l'aide de Inkscape et Jmol à partir du fichier PDB : DOI :10.2210/pdb3myj/pdb)*

correspondent à ce que l'on appelle des résidus d'ancrage. L'ensemble de ces résidus constitue le motif de liaison du CMH-I. Leur identification nous permet notamment de prédire l'affinité de liaison entre des peptides et une isoforme du CMH-I [68] [46] [17]. La non correspondance d'un seul acide aminé peut compromettre la liaison du peptide avec la molécule du CMH-I. Le tableau 1.I liste plusieurs peptides connus pour se lier à la molécule HLA-A*0201.

À partir de ce sous-ensemble de peptides connus, on peut établir des règles définissant quels peptides sont susceptibles de se lier à cette molécule HLA. Par exemple à la

Tableau 1.I – Un sous-ensemble de séquences de peptides connus pour se lier à l'isoforme HLA-A*0201

P1	P2	P3	P4	P5	P6	P7	P8	P9
A	**L**	A	K	A	A	A	A	M
A	**L**	A	K	A	A	A	A	N
A	**L**	A	K	A	A	A	A	V
A	**L**	A	K	A	A	A	A	T
A	**L**	A	K	A	A	A	A	V
G	**M**	N	E	R	P	I	L	T
G	**I**	L	G	F	V	F	T	M
T	**L**	N	A	W	V	K	V	V
K	**L**	N	E	P	V	L	L	L
A	**V**	V	P	F	I	V	S	V

position 2, on peut dire que les acides aminés Leu (L), Met (M), Ile (I) et Val (V) sont permis. L'acide aminé Leu semble néanmoins être plus fréquemment observée. Même à partir de petits ensembles de données, des approches statistiques permettent de fournir une description des motifs de liaison et de prédire l'affinité de liaison d'un peptide donné pour telle ou telle isoforme du CMH-I [43] [17]. Des équipes ont développé des méthodes de prédiction mettant à profit des ensembles de plus de 200 peptides pour une même isoforme [68]. Plusieurs algorithmes de prédiction sont disponibles au premier desquels SYFPEITHI [65] et NetMHC-3.0 [44]. Ces outils bioinformatiques seront présentés dans la sous-section 1.4.

1.2.1.3 Le polymorphisme : un avantage et un inconvénient

Chacune des espèces de vertébrés a un grand nombre d'allèles différents de chacun des gènes du CMH-I. Les cellules humaines expriment concurremment les protéines des sous-classes A, B et C du CMH-I. Chacune de ces sous-classes comportent plusieurs allèles. Les molécules du CMH-I font partie de celles présentant le plus de polymorphisme. À l'exception des jumeaux monozygotes, la probabilité de trouver deux indi-

vidus présentant un CMH-I identique est très faible. Le polymorphisme, du point vue de l'espèce, est un avantage indubitable car il permet d'augmenter la probabilité qu'une sous-population marginale survive à une épidémie. En effet, les modèles de co-évolution hôte-parasite montrent une adaptation du parasite au génotype hôte le plus fréquent [70]. Par conséquent, un hôte qui possède un allèle CMH rare est avantagé dans la population. Ce phénomène crée une sélection fréquence-dépendante négative où les allèles rares sont favorisés jusqu'à ce que le parasite s'adapte. Cependant, le polymorphisme représente une difficulté, voire un obstacle, pour les transplantations. La compréhension de la genèse de l'immunopeptidome ainsi que de sa composition est essentielle pour mettre en oeuvre des procédés minimisant le risque de rejet.

1.2.2 La présentation des antigènes du CMH de classe I

Le peptide antigénique lié à la molécule du CMH-I arrivant à la surface de la cellule a dû franchir plusieurs étapes avant d'en arriver là. Plusieurs organelles et molécules interviennent dans le processus qui conduit à la présentation du peptide.

1.2.2.1 Le protéasome et les autres protéases impliquées

L'apprêtement des peptides à la surface cellulaire repose à la fois sur des protéines non spécialisées et des protéines exclusivement réservées à la présentation antigénique. Parmi les premières d'entre elles, on trouve les protéases cytosoliques (le protéasome notamment) et des protéines chaperonnes du réticulum endoplasmique (RE). Parmi les protéines spécialisées, on compte un transporteur de peptides situé à la membrane du RE, deux aminopeptidases et une protéine chaperonne. Dans le cytosol, les protéines devenues inutiles ou inutilisables sont dégradées par le protéasome [6]. En plus des protéines cytosoliques, celles non repliées ou mal assemblées sont pris en charge par cette

protéase [51]. Cette dernière fait nulle distinction entre les protéines endogènes et celles provenant de pathogènes. Elles sont prises en charge exactement de la même manière et subissent le même sort. En l'absence de pathogènes, seuls des peptides issus de la dégradation de protéines endogènes sont présentés. Mais ces derniers ne déclenchent, normalement, aucune réponse immunitaire car ils sont connus comme appartenant au soi. Le système d'apprêtement est lié à la dégradation des protéines. Cependant, plusieurs études ont montré qu'une partie des peptides présentés à la surface cellulaire provient de protéines ayant une courte demi-vie. Certaines n'ont d'ailleurs jamais été fonctionnelles [88]. L'apprêtement des peptides pourrait donc être autant lié au taux de traduction des protéines qu'à la dégradation de celles-ci. Cette hypothèse, bien que non unanime [80], est séduisante car elle répond à la nécessité d'induire une réponse immunitaire rapide lors d'une infection virale.

Le protéasome, représentant un ensemble de protéases multicatalytiques essentielles dans le métabolisme des protéines cellulaires [6], se situe au tout début de la chaîne de traitement menant à la présentation des peptides antigéniques. Il est composé d'une structure cylindrique, et composé de 28 protéines organisées en 4 anneaux appelés complexe 20S. Les complexes 26S régulent l'accès des substrats dans la cavité cylindre. Le rôle du protéasome est de dégrader les substrats marqués pour la dégradation par une poly-ubiquitinylation [13]. Le protéasome est équipé de trois spécificités catalytiques [39] qui sont similaires à la trypsine (clivage après les acides aminés Arg et Lys), à la chymotrypsine (clivage après les acides aminés Tyr, Phe, Leu et Ile) et à la caspase (clivage après Glu et Asp). Celles-ci lui confèrent la possibilié de dégrader quasiment toutes les protéines en peptides d'une longueur comprise entre 4 et 15 résidus. D'après certaines études, moins de 15% des peptides issus de la dégradation par le protéasome auraient une longueur supérieure à celle des peptides du CMH-I [38]. D'autres pré-

tendent que les expériences *in vitro* ne sont pas fidèles à ce qui se passe réellement dans la cellule et que la plupart des peptides produits par le protéasome ont des longueurs supérieures à 15 acides aminés [66]. Cette dernière hypothèse sous-entend donc l'implication d'autres protéases pour produire des peptides de taille canonique, à savoir entre 8 et 11 acides aminés. Sous l'influence de l'interféron gamma, les trois sous-unités catalytiques appelé protéasome constitutif sont remplacées pour donner lieu à la formation de l'immunoprotéasome [39]. Celui-ci est dépourvu de l'activité de type caspase qui est sans utilité pour la présentation des peptides du CMH-I. En effet, les molécules de classe I ne fixent pas de peptides avec un résidu acide en C-terminal. L'immunoprotéasome se veut donc être plus efficace pour la présentation des peptides du CMH-I.

Seule une toute petite partie des peptides produits par le protéasome finit à la surface de la cellule [39]. Plusieurs protéases sont impliquées dans la destruction des peptides [69] dont la demi-vie n'est pas supérieure à quelques secondes [66]. La tripeptidylpeptidase est une protéase qui affectionne les peptides de plus de 15 résidus. Celle-ci pourrait jouer un rôle dans la présentation des peptides à la surface cellulaire [66] bien que celui-ci semble être limité [35]. Les peptides de 8 à 13 résidus sont dégradés par la thimet oligopeptidase [74] alors que ceux étant plus courts sont pris en charge par d'autres aminopeptidases situées dans le cytsol et le RE [69]. Le point commun de toutes ces protéases est leur activité aminopeptidase. Aucune protéase connue, autre que le protéasome, ne se charge de l'autre extrémité. L'ensemble des protéases impliquées dans la constitution de la population de peptides antigéniques est responsable de la composition et de la distribution des acides aminés le long de la chaîne peptidique.

1.2.2.2 Le transport des peptides du cytosol au réticulum endoplasmique

La translocation des peptides du cytosol au site d'assemblage avec les molécules du CMH-I, le RE est assurée par le transporteur TAP [81] . Les cellules ayant un déficit en TAP expriment très peu de molécules du CMH-I à leur surface [82]. Parmi nombre d'espèces, notamment l'homme et la souris, le transporteur TAP semble avoir plus d'affinité pour les peptides de 8 à 16 résidus. Ceci est en adéquation avec la longueur des peptides pouvant être associés aux molécules du CMH-I. En revanche, le motif des séquences préférées pour la souris diffèrent de celles de l'homme [52] [8]. Chez l'homme, les acides aminés volumineux (Tyr et Phe) et basiques (Arg) sont préférés alors que chez la souris ils ne le sont pas. Des études ont mis en évidence par exemple la contribution importante des trois positions aminoterminales pour le transport d'un peptide par TAP [8]. Alors que les peptides matures ne passeront par aucune autre étape protéolytique une fois rendus dans le RE, les peptides précurseurs tout juste transloqués peuvent subirent une dernière protéolyse du côte aminoterminal [67]. Deux aminopeptidases ont été identifiées : ERAP1 et ERAP2 (qui n'est pas exprimée chez la souris). La première préfère les résidus hydrophobes alors que ERAP2 affectionne davantage les résidus basiques [72]. Il a été montré que ces deux peptidases jouent un rôle majeur dans la constitution de l'immunopeptidome [28].

1.2.2.3 L'assemblage des peptides avec la molécule du CMH-I

Quatre protéines chaperones sont impliquées dans la formation du complexe peptide-molécule du CMH-I (figure 1.4) [16]. La molécule du CMH-I demeure instable sans la présence d'un peptide. En attendant, qu'un peptide vienne se fixer à elle avec une bonne affinité, elle reste associée aux protéines chaperones. La calnexine s'associe aux

chaînes lourdes tout juste synthétisées comme illustrée par la figure 1.4. La calréticuline remplaçant la calnexine ainsi que ERp57 aide la β2-microglobuline à s'associer à ces dernières. Il y a formation de ponts disulfures entre les chaînes lourdes. Enfin intervient la tapasine en liant le complexe à un transporteur TAP pourvu d'un peptide. Une fois la molécule complètement assemblée, elle se libère de TAP pour migrer vers la surface cellulaire.

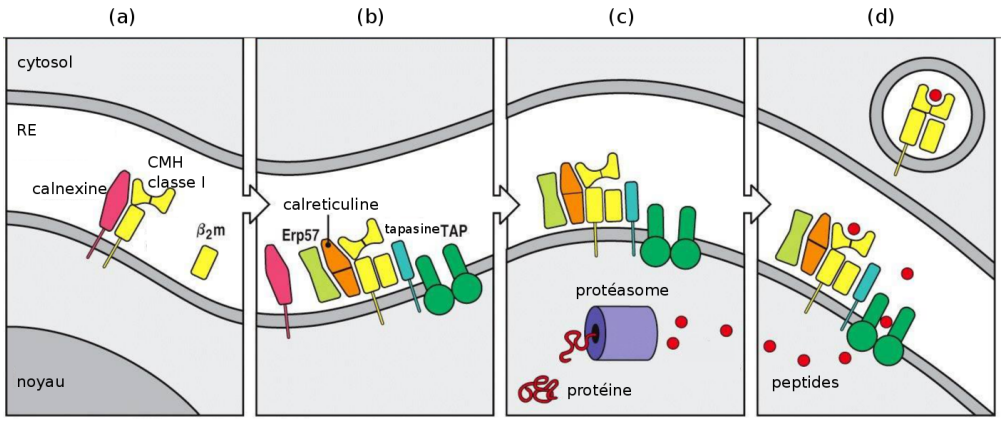


Figure 1.4 – Apprêtement des antigènes et chargement des molécules du CMH-I. (a) La chaîne α partiellement repliée se lie à la calnexine jusqu'à ce que la β2-microglobuline se lie à elle. (b) Le complexe se libère de la calnexine, se lie au complexe de protéines (calreticuline, Erp57) et se lie à TAP via la tapasine. (c) Les protéines cytosoliques sont dégradées en fragments par le protéasome notamment. (d) TAP libère le peptide qui se lie à la molécule du CMH-I. La molécule du CMH-I complètement assemblée se libère de TAP et se dirige vers la surface cellulaire.

1.2.2.4 La présentation croisée des antigènes exogènes par les molécules du CMH de classe I

Les cellules dendritiques sont capables de présenter des antigènes internalisés. Ces dernières mettent en oeuvre ce qu'on appelle la présentation croisée [86]. Elle constitue une voie de présentation qui permet aux antigènes extracellulaires internalisés de subir le même processus que celui décrit plus haut [86]. La présentation croisée permet aux

cellules dendritiques de déclencher des réponses immunitaires contre des pathogènes en utilisant les récepteurs du CMH-I et en présentant ceux-ci aux cellules T cytotoxiques CD8.

1.2.3 L'importance de l'immunopeptidome

Outre l'intérêt pour la connaissance que représente la compréhension de la genèse de l'immunopeptidome, l'utilisation thérapeutique des peptides du CMH-I connaît un regain d'intérêt depuis quelques années [64]. Il convient de faire un tour d'horizon des maladies concernées par l'immunopeptidome pour montrer combien son étude relève d'une grande importance.

1.2.3.1 Les maladies autoimmnunes

Le système immunitaire a pour rôle de protéger l'intégrité de l'organisme. Néanmoins, il arrive que celui-ci se trompe de cible et s'attaque à des tissus sains. On parle alors de réactions autoimmunes. Il existe nombre de maladies pour lesquelles ces réactions en sont la cause : le lupus, diabète de type 1, la polyarthrite rhumatoïde, la spondylarthrite ankylosante, le syndrome de Goujerot-Sjögren, la maladie de Crohn, la myasthénie, etc. Lorsque le système immunitaire s'attaque à une protéine de l'organisme, il ne parvient jamais à l'éliminer car celle-ci est en permanence synthétisée. C'est pourquoi ces réactions sont à l'origine de maladies chroniques. Le rôle du CMH-I dans les maladies auto-immunes a été établis par plusieurs études comme dans le cas du lupus par exemple [84]. Certaines recherches s'intéressent à la possibilité d'administrer des antigènes qui pourraient stimuler des méchanismes tolérogéniques [63] pour retarder ou amoindrir l'expression phénotypique de la maladie.

1.2.3.2 Les allergies

Il est avéré que la prévalence des allergies, notamment dans les pays industrialisés, ne cesse d'augmenter. Celles-ci peuvent causer des symptômes bénins mais peuvent dans certains cas s'avérer être plus grave ou nuire à la qualité de vie des personnes qui en souffrent. Elles correspondent à des réactions inadaptées ou exagérées du système immunitaire. Certaines recherches prometteuses se sont intéressées à l'utilisation de vaccins basés sur des peptides du CMH-I comme immunothérapies contre les allergies [85].

1.2.3.3 Le cancer

Le cancer est une des principales causes de mortalité. Le cancer résulte de la croissance de cellules ayant perdu toute régulation. Pour détruire le cancer, il faut donc détruire les cellules déréglées. Une des voies possibles consiste à générer une réponse immunitaire contre les cellules néoplastiques. Une équipe menée par le Dr Perreault a montré que l'injection de lymphocytes T CD8 pré-activés contre l'antigène H7a peut guérir des mélanomes malins chez la souris [50]. D'autres équipes se sont intéressées au développement de méthodes et d'outils bioinformatiques pour la découvertes d'antigènes tumoraux [75]. Leur objectif est de développer des immunothérapies basées sur plusieurs épitopes associés à des cancers afin d'assurer une réponse immunitaire spécifique et efficace contre les cellules cancéreuses.

1.2.3.4 Les vaccins

L'humanité n'est jamais à l'abri d'une pandémie dévastatrice. L'utilisation de la vaccination, dont le principe a été expliqué par Louis Pasteur, a permis d'éradiquer des maladies ou d'en diminuer de façon importante leur prévalence. L'identification de pep-

tides du CMH de classe I qui induisent une réponse immunitaire constitue une voie menant au développement de vaccin. Seule une petite fraction des peptides du protéome du pathogène est capable d'induire une réponse immunitaire. Ceci est principalement dû à la sélectivité de l'apprêtement des peptides du CMH de classe I. Pour chaque allèle de classe I, seulement 1 sur 2000 peptides sera immunodominant [87]. La détermination de peptides candidats, induisant une réponse immunitaire, pour le développement de vaccin constitue un véritable défit.

1.2.3.5 La greffe allogénique

L'allogreffe se distingue de l'autogreffe par le fait que le donneur et le receveur sont deux personnes différentes. Le patrimoine génétique est forcément différent entre les deux personnes à moins qu'il ne s'agisse de vrais jumeaux (isogreffe). Intervient alors la notion d'histocompatibilité. Le donneur et le receveur doivent être HLA compatibles. Au sein d'une fratrie, la probabilité d'obtenir un donneur et un recevoir histocompatibles est plus grande puisqu'il partage une partie de leur patrimoine génétique. La greffe myéloablative, indiquée dans les cas de leucémie par exemple, consiste à détruire les cellules hématopoïétiques du receveur avant de transplanter les cellules souches saines d'un donneur. Dans cette stratégie thérapeutique, les cellules tumorales, le système sanguin du patient responsable de la maladie ainsi que son système immunitaire sont détruits. Les cellules néoplastiques sont la cible du système immunitaire du donneur qui remplace celui du receveur incapable d'identifier ces dernières. C'est ce que l'on appelle la réaction du greffon contre la tumeur (GVL) que l'on cherche à provoquer. Les lymphocytes T provenant du greffon sont capables de reconnaître et de détruire les cellules tumorales chez le receveur. Pourvu d'un nouveau système immunitaire, le patient sera armé pour lutter contre la maladie et minimiser les probabilités de rechute. Néanmoins cette greffe

comporte plusieurs risques. Parmi lesquels figure la réaction du greffon contre l'hôte (GVH) où les lymphocytes T provenant du greffon s'attaquent aux tissus du receveur. Même lorsqu'il s'agit de greffes allogéniques au sein d'une fratrie, il y a un taux important de réaction de ce type. Le taux de GVH chronique varie entre 40 et 70% [49]. Celle-ci peut être induite par les antigènes mineurs d'histocompatibilité auxquels on s'intéressera dans le chapitre 7. La spectrométrie de masse constitue une solution de choix pour l'identification de ces antigènes mineurs.

1.3 La spectrométrie de masse

La spectrométrie de masse est devenue un outil très précieux pour l'étude du protéome. Elle l'est tout autant pour l'étude des peptides du CMH-I. Les progrès techniques permettent aujourd'hui d'espérer des découvertes prometteuses découlant de son utilisation dans le domaine de l'immunologie. Dans le cadre de notre étude, elle est utilisée pour déterminer la séquence des peptides du CMH-I.

1.3.1 Une vue d'ensemble

La spectrométrie de masse est une technique de détection très sensible qui permet de déterminer des structures moléculaires. Le spectromètre de masse est souvent couplée avec un système de chromatographie, d'une méthode de séparation ainsi que d'une méthode d'identification. Un composé organique introduit dans le spectromètre de masse est ionisé, et ce, par différentes méthodes d'ionisation possibles. Certaines d'entre elles sont très énergétiques et conduisent à une fragmentation importante. D'autres sont plus douces et ne produisent pas de fragmentation. L'obtention de l'ion moléculaire permet de déterminer la masse molaire du composé. Des ruptures de liaison chimique aboutis-

sant à la formation d'ions fragments peuvent avoir lieu. Celles-ci sont caractéristiques car elles suivent un processus déterministe. Les ions fragments sont séparés en fonction du rapport masse/charge (notée *m/z*). Ils sont collectés par un détecteur qui convertit le courant ionique en courant électrique. C'est l'ensemble des ions fragments produits qui constitue le spectre de masse permettant *in fine* la détermination de la structure moléculaire du composé. Les fragments non chargés ne peuvent être détectés. La figure 1.5 montre la structure d'un spectromètre de masse.

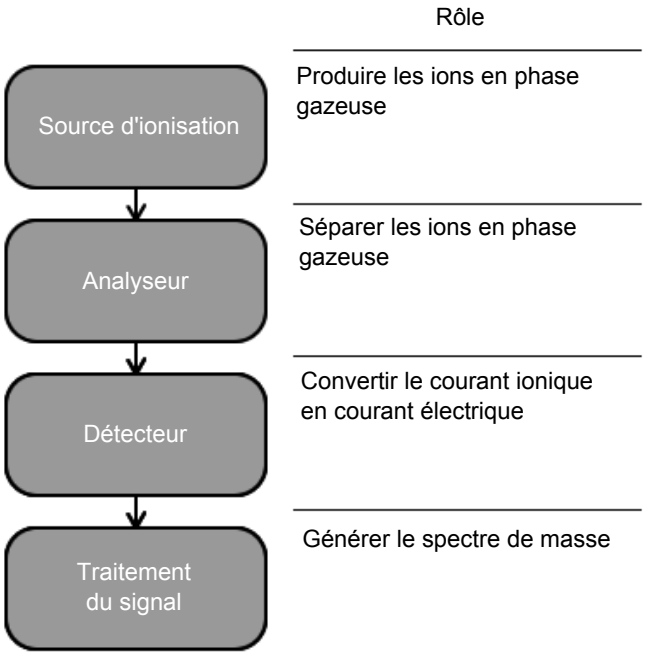


Figure 1.5 – Structure d'un spectromètre de masse. Le spectromètre de masse se compose principalement de 4 éléments : la source d'ionisation, l'analyseur, le détecteur et la partie qui traite le signal pour générer le spectre de masse.

1.3.2 La source d'ionisation

La source d'ionisation, comme son nom l'indique, permet l'ionisation des substances à analyser. L'une des caractéristiques importantes est la quantité d'énergie interne transférée. Plus celle-ci est élevée, plus elle conduit à la fragmentation de la molécule. La source d'ionisation employée dépend de la nature physicochimique de la molécule à analyser. Pour l'étude des protéines et des peptides, les méthodes d'ionisation les plus utilisées sont le bombardement par atomes ou ions rapides (FAB ou LSIMS), l'électronébulisation (ESI) et la désorption-ionisation laser assistée par matrice (MALDI). Ces techniques peu énergétiques conduisent à la formation d'ions stables et sans fragments. Les spectromètres que nous avons employés pour notre étude, à savoir le LTQ Orbitrap XL et le LTQ Orbitrap Velos, utilisent l'électronébulisation. Cette technique, utilisant un processus électrochimique, a été inventée par Fenn dans les années 1980 et s'est répandue depuis les 1990. En effet, elle présente de nombreux avantages aux premiers desquels la sensibilité très élevée et la possibilité de couplages avec la chromatographie liquide ou l'électrophorèse capillaire [53] [10]. La technique consiste à appliquer à pression atmosphérique un fort champ électrique sur un liquide traversant un tube capillaire avec un faible débit (moins de 100 μl/min). Entre le capillaire et la contre-électrode est appliquée une différence de potentiel de 3 à 6kV générant un champ électrique. Lorsque la pression des charges accumulées à la surface du liquide est suffisante, la force exercée surpasse la tension superficielle et des gouttelettes se détachent formant un jet nanométrique. L'évaporisation du solvant conduit au rétrécissement des gouttelettes et donc à l'augmentation de la densité de charge. Les forces coulombiennes répulsives vont contrebalancer les forces de cohésions et conduire à l'explosion des gouttelettes. Un nuage de fines gouttelettes se forme. Ce processus se répète jusqu'à ce que la densité de charge

devienne suffisante pour conduire à la désorption des ions. Ce procédé permet une ionisation douce. Ce sont donc en majorité des ions moléculaires qui sont formés.

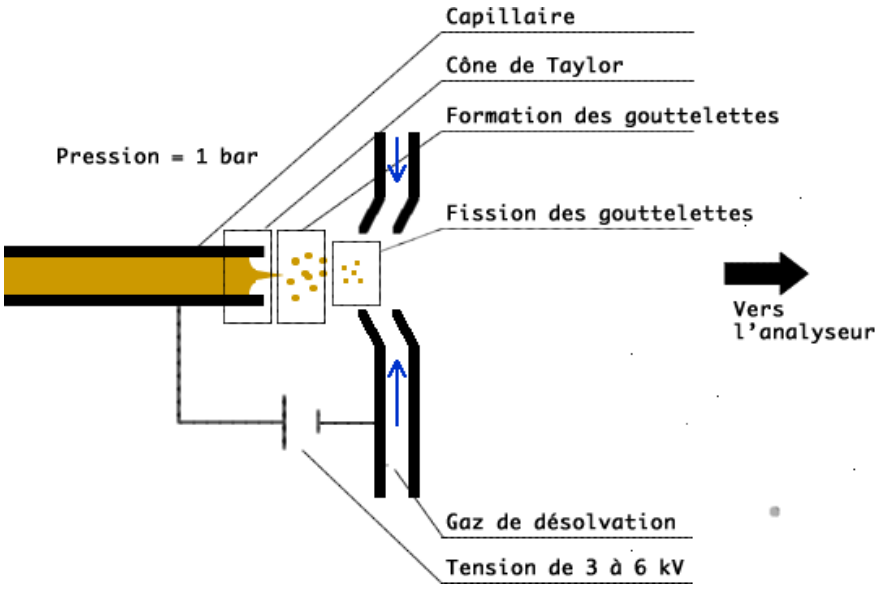


Figure 1.6 – Source d'ionisation par électronébulisation. Le nuage de gouttelettes se forment à l'extrémité du capillaire jusqu'à la désorption des ions qui sont dirigés vers l'analyseur.

1.3.3 Les analyseurs

Les analyseurs réceptionnent les ions produits et se chargent de les séparer suivant leur m/z. Les spectromètres utilisés dans la présente étude utilisent deux analyseurs différents : la trappe ionique linéaire et la trappe électrostatique.

1.3.3.1 La trappe ionique linéaire

La trappe ionique linéaire (LTQ) est une trappe ionique ayant la spécifié d'être de géométrie linéaire. Comme pour les autres trappes ioniques, son fonctionnement est basé sur l'action d'un champ électrique radiofréquence sur les ions. Le mouvement de ceux-ci dépend de leur masse m ainsi que de leur charge z. Il y a des zones de stabilité pour

lesquelles les ions d'une certaine masse m peuvent avoir un mouvement stable et donc rester piégés dans la trappe ionique. Les ions piégés dans la trappe sont éjectés sélectivement en fonction de la valeur m/z. Le LTQ permet d'augmenter de façon considérable la rapidité d'acquisition et par conséquent les performances de l'instrument. Elle a une excellente sensibilité du balayage et est tout adapté à la spectrométrie en tandem (dont le principe est expliqué dans la sous-section 1.3.4).

1.3.3.2 La trappe électrostatique

La trappe électrostatique (aussi appelé Orbitrap) a été inventée par Alexander Alekseevitch Makarov [48] [33] et a fait l'objet d'un premier brevet en 1999. Elle se compose d'une électrode en forme de tonneau à l'intérieur de laquelle se trouve une électrode en forme de fuseau. Une tension continue est appliquée entre les deux électrodes. Les ions sont injectés tangentiellement et avec une énergie cinétique de quelques keV. Ils oscillent alors suivant l'axe z en formant des spirales autour de l'électrode interne. La forme des deux électrodes conduit à la formation d'un champ électrostatique quadrologarithmique. Les oscillations des ions suivant l'axe z ont une fréquence (1.1) :

$$\omega = \sqrt{(\frac{z}{m})k} \tag{1.1}$$

Seul le rapport m/z affecte cette fréquence. Le méthode de la transformée de Fourier est utilisée pour convertir la fréquence en m/z. La précision en masse est très bonne (1-2 ppm) et la résolution (100 000).

1.3.4 La spectrométrie de masse en tandem

La spectrométrie de masse en tandem (communément appelée MS/MS) est une méthode qui implique au minimum deux étapes d'analyse de masse (figure 1.7). La première étape consiste à sélectionner et isoler un ion, appelé ion précurseur. Ensuite celui-ci est fragmenté dans un cellule à collision rempli d'un gaz inerte. La deuxième étape consiste à analyser les ions fragments générés par la première étape. Il existe deux types de spectromètre de masse en tandem. Le premier, basé sur une séparation spatiale, met en oeuvre le couplage de deux analyseurs. Le deuxième, basé sur une séparation temporelle, utilise un dispositif de stockage d'ions. Dans notre étude, nous employons deux spectromètres de masse en tandem qui se basent sur une séparation spatiale utilisant une cellule de collision. En protéomique, la spectrométrie de masse en tandem est notamment utilisée pour élucider la séquence des peptides. En effet, à cause de la stabilité des liaisons, le peptides se clivent principalement aux niveaux des liaisons peptidiques. Ceci est mis à profit pour déterminer leur séquence. Nous parlerons dans la sous-section 1.3.7 de la fragmentation des peptides et l'interprétation des spectres résultants.

1.3.5 La dissociation induite par collision

La spectrométrie de masse en tandem nécessite la fragmentation des ions précurseurs. La collision de ceux-ci avec un gaz cible entraine leur activation ainsi qu'un gain d'énergie interne suffisant pour induire la dissociation des liens peptidiques. La méthode la plus utilisée se base sur les collisions (CID) et se déroule en deux étapes. La première correspond à la collision de l'ion durant laquelle une portion de l'énergie translationnelle est convertie en énergie interne. Cette étape conduit l'ion dans un état excité. Dans une deuxième étape, l'ion se décompose en fragments. Il existe différentes méthodes pour

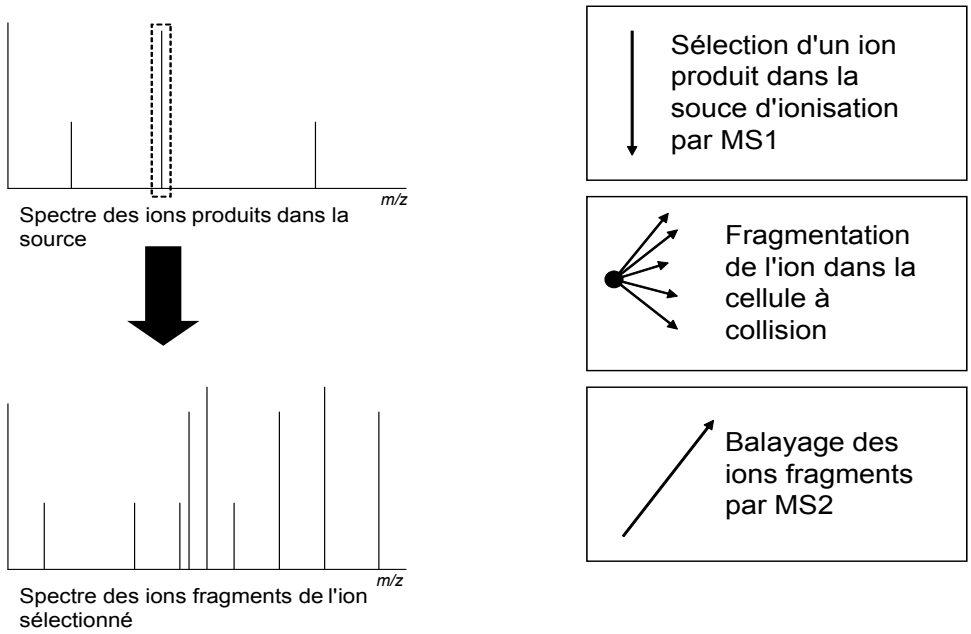


Figure 1.7 – Principe de la spectrométrie en tandem. un ion est sélectionné par le premier analyseur MS1. Celui est ensuite fragmenté dans la cellule à collision. Finalement, les ions fragments résultants sont analysés par le second analyseur MS2.

l'activation d'ions par collision. Dans les spectromètres qui ont été utilisés se trouve une cellule à collision située entre les deux analyseurs qui renferme un gaz inerte à une pression suffisante. L'ajout d'énergie conduit à une augmentation de la population des formes ayant des énergies plus élevées. Le proton est délocalisé à différents endroits sur chaîne peptidique. La présence de ce proton initie la fragmentation au niveau des liaisons peptidiques pour former des ions fragments [20]. Les spectromètres de masse utilisent dans le cadre de cette étude un régime collisionnel à basse énergie (entre 1 et 200 eV) et les fragmentations à haute énergie (quelques keV). Les deux modes de fragmentation CID et HCD utilisés dans notre étude se rapporte à la fragmentation par collision à basse énergie. Le mode HCD permet notamment une fragmentation à plus haute énergie que le mode CID et sans perte de fragments de basses masses. Elle offre une précision en masse et une sensibilité accrues.

1.3.6 Les spectromètres utilisés

1.3.6.1 Le spectromètre de masse LTQ-Orbitrap XL

Le spectromètre de masse LTQ-Orbitrap XL est un appareil hybride constitué d'une trappe ionique linéaire couplée à une trappe électrostatique via une C-trap. Il combine les avantages des deux analyseurs. Les avantages de la trappe ionique linéaire sont la haute sensibilité, la gamme dynamique élevée, le temps de cycle rapide. La fragmentation des précurseurs a lieu dans cette première trappe. Quant à la trappe électrostatique, elle fournit une haute précision en masse et une haute résolution. Elle se charge de l'acquisition des spectres de masse. Les caractéristiques de ce spectromètre en fait l'outil adapté pour l'analyse des protéines et des peptides.

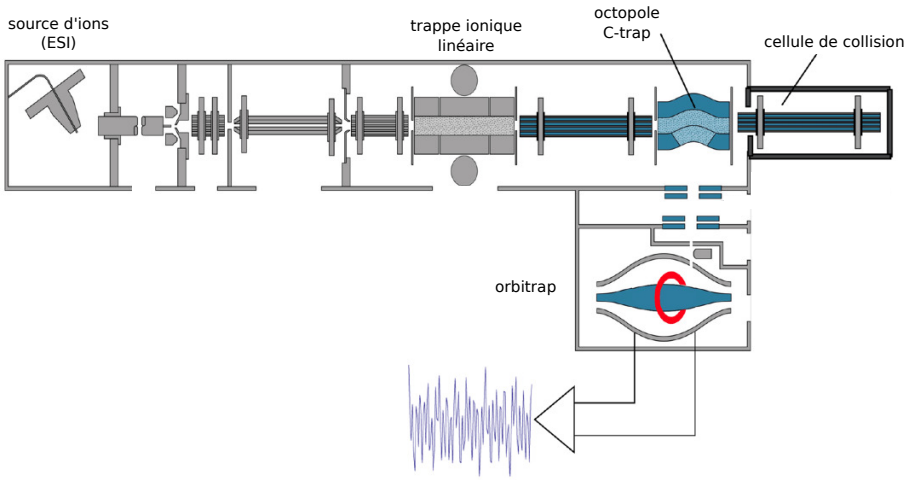


Figure 1.8 – Schéma général du LTQ-Orbitrap-XL. LTQ-Orbitrap XL est un appareil hybride constitué d'une trappe ionique linéaire couplée à une trappe électrostatique (orbitrap) via une C-trap qui assure le transfert des ions de la trappe linéaire à la trappe électrostatique. *(souce : `www.thermo.com`)*

L'appareil est équipé d'une source d'ionisation électronébulisation. Elle utilise une sonde dynamique pour le couplage à un système chromatographique à nano débit (nanoLC). Le premier analyseur de masse est une trappe ionique linéaire, permettant de

travailler sur une gamme de masse allant de 15 à 4000 m/z (15 – 200 m/z, 50-2000 m/z, et 100-4000 m/z). Le transfert des ions est assuré par un octopole, qui amène les ions dans un piège intermédiaire appelé C-Trap. Les ions sont ensuite pulsés vers le trappe électrostatique. Le vide très poussé et l'injection rapide des ions dans la trappe électrostique permettent aux ions d'être stables pendant plusieurs secondes. Ceci permet d'obtenir une haute résolution et une haute précision de masse. La gamme de masse de la trappe électrostatique est la même que celle du premier analyseur. Dans notre étude, nous utilisons ce spectromètre pour les analyses de la fragmentation en mode CID.

1.3.6.2 Le spectromètre de masse LTQ-Orbitrap Velos

Cet appareil est le dernier de la série des LTQ-Orbitrap et bénéficie des dernières améliorations. Les trois nouveaux composants qui confèrent à cet appareil des performances inégalées sont :

1. un système de transfert des ions plus performant (S-lens),
2. une trappe ionique à double pression,
3. et une cellule HCD plus performante.

À la place de la trappe ionique linéaire utilisé dans le spectromètre brièvement décrit plus haut, l'assemblage de deux trappes ionique à pression différentes séparées par une lentille est utilisée. La première trappe est remplie d'hélium à haute pression alors que la deuxième a une plus faible pression. Grâce à ses composants et à une sensibilité 10 fois supérieure que celle du LTQ-Orbitrap XL, ce spectromètre permet d'obtenir une haute résolution et une haute précision en masse de façon routinière [57]. Ce spectromètre permet de détecter des ions fragments à de faible masse et donc d'obtenir des spectres

plus informatifs. Dans notre étude, nous utilisons ce spectromètre pour les analyses de la fragmentation en mode HCD.

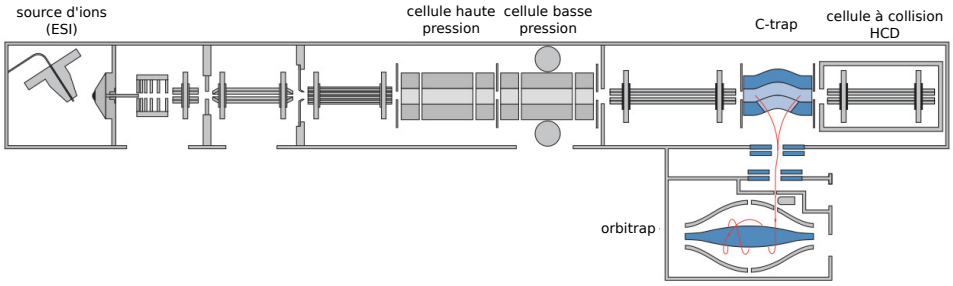


Figure 1.9 – Schéma général du LTQ-Orbitrap-Velos. Le conception à trappe ionique à double pression permet une capture efficace et une activation à haute pression (à gauche) et une détection dans la cellule à basse pression (à droite). Une lentille (S-lens) permet d'augmenter le flux d'ions de la source d'ionisation (ESI) dans l'instrument par un facteur de 5 à 10. L'association du C-trap et de la cellule à collision HCD fournissent une extraction des ions fragments améliorés et un meilleur piégeage.

1.3.7 La fragmentation des peptides et leur séquencage

Les peptides sont des chaînes d'acides aminés reliées par des liaisons peptidiques. On dénombre 20 acides aminés. Les peptides fragmentent principalement le long de la chaîne principale avec souvent le transfert d'un atome d'hydrogène (ou 2) conférant à l'ion une stabilité. On peut distinguer deux catégories de fragments. Les premiers proviennent du clivage d'une ou de deux liaisons de la chaîne peptidique. Les seconds subissent en plus un clivage de la chaîne latérale d'un acide aminé. Le clivage au niveau de la chaîne peptidique peut se faire aux niveaux de trois liaisons : $C\alpha - C$, $C - N$, $N - C\alpha$ comme le montre la figure 1.10(a). Le clivage au niveau de la première liaison peut donner des fragments a_n si la charge est du côté N-terminal et x_n si elle est du côte C-terminal. De façon analogue, le clivage au niveau de la deuxième liaison donne des fragments b_n et y_n. Et enfin au niveau de la troisième liaison, on peut observer des frag-

ments c_n et z_n. Alors que l'activation à basse énergie génère principalement les types d'ion b_n et y_n, à haute énergie on peut observer l'ensemble des types de fragments. Cependant, comme il sera discuté plus loin, la fragmentation dépend aussi grandement de la composition en acides aminés du peptide. C'est la différence de masse entre les ions fragments consécutifs qui permet d'identifier les acides aminés et reconstruire la séquence du peptide. La figure 1.10 (b) décrit la nomenclature pour les différents ions fragments. La figure 1.10 (c) montre comment reconstruire la séquence peptidique à partir du spectre. Par exemple, la masse séparant les pics correspondants aux fragments b_1 et b_2 est égale à celle de l'acide aminé A_2, la même chose pour l'acide aminé A_3. La connaissance de la masse de l'ion précurseur permet de compléter la séquence. Un peptide de n résidus génère au maximum $(n-1)$ ions fragments a_n, b_n ,c_n, x_n, y_n et z_n. De plus, les fragments b_n et y_n notamment peuvent subir des pertes de molécules neutres comme H_2O ou NH_3 à partir des chaînes latérales des acides aminés. On observe aussi des fragments internes qui résultent de la fragmentation à la fois du côté N-terminal et C-terminal. Les ions immoniums peuvent également apparaître dans les spectres. Ils correspondent aux clivages multiples de la chaîne peptidique.

La position N-terminale des peptides est un site favorable à la charge. Les atomes d'azote des liaisons amides peuvent aussi être porteur de charge. Dans les peptides contenant des acides aminés basiques (Arg, Lys, His et Pro), la charge est préférentiellement localisée sur ces résidus. Les peptides dépourvus de ces acides aminés produisent indifféremment des fragments b_n et y_n même si les premiers sont favorisés avec l'augmentation de la taille du peptide. En revanche, les peptides contenant des résidus basiques vont produire davantage de fragments y_n s'ils sont localisés sur le côté C-terminal. Cependant, si ces résidus sont du côté N-terminal, on observera en majorité des fragments b_n. Ceci illustre, en partie seulement, l'influence de la composition des peptides sur le profil de

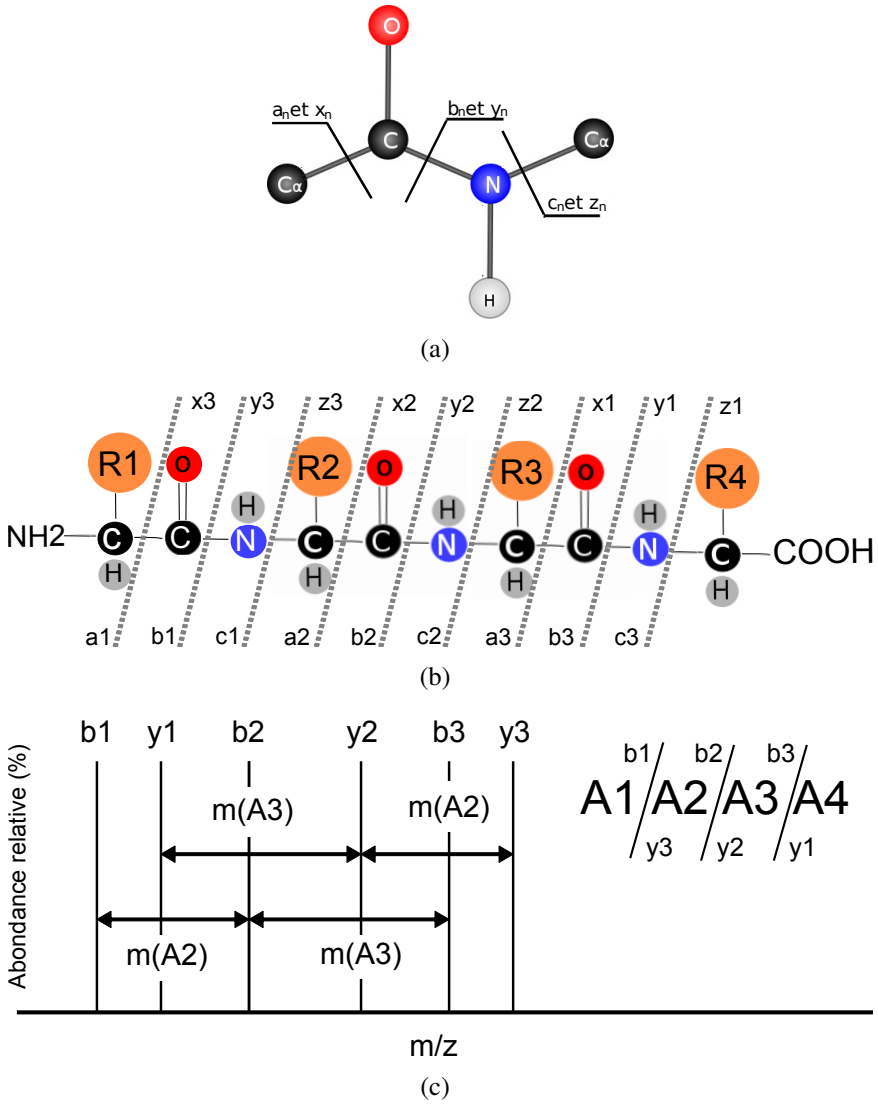


Figure 1.10 – Fragmentation de la chaîne peptidique et nomenclature. (a) Les différentes liaisons brisées par l'activation par collision . (b) La nomenclature pour des différents types d'ions fragments. (c) Le spectre théorique pour le peptide $A_1A_2A_3A_4$. On observe 6 pics correspondant aux ions fragments y_n et b_n. Les ions fragments ont tous la même abondance relative mais ce n'est jamais le cas en pratique. Les ions b_1 et b_2 sont séparés par un m/z correspondant à la masse de l'acide aminé A_2 et ainsi de suite.

fragmentation.

1.3.8 L'identification des peptides

À partir des spectres de masse obtenus, deux principales méthodes bioinformatiques s'offrent à nous pour identifier les peptides et éventuellement les protéines sources. La première se base sur l'utilisation d'une base de donnée de séquences de protéines ou génomiques. Alors que la deuxième vise à interpréter le spectre sans aucune autre information. L'une comme l'autre ont leurs avantages et leurs inconvénients.

1.3.8.1 Les moteurs de recherche

Pour une séquence peptidique donnée, il est possible de déterminer un spectre de masse théorique. Nous pourrions imaginer générer les spectres théoriques pour tous les peptides possibles et calculer la similarité de ceux-ci avec le spectre expérimental dont on cherche à connaître la séquence. Cependant, on se heurte à un problème de taille. En effet, le nombre de peptides possibles pour uniquement ceux qui comptent 10 résidus est égal à 20^{10}. C'est pourquoi il a été proposé de comparer les spectres expérimentaux uniquement avec ceux calculés à partir des protéines répertoriées dans les bases de données [76]. Plusieurs moteurs de recherche utilisent ce principe de recherche de similarité spectrale : SEQUEST [22], X !tandem [15], MASCOT [60], etc. La recherche est limitée aux séquences peptidiques ayant des masses correspondant à celle du peptide précurseur. Il existe deux catégories d'algorithmes : les algorithmes heuristiques et les algorithmes probabilistes [34]. Les premiers se basent sur le comptage du nombre de pics partagés entre les spectres théoriques et expérimentaux (SEQUEST, X !tandem). Les deuxièmes, dont fait partie MASCOT, se base sur un calcul de score probabiliste. L'approche consiste à calculer la probabilité que la correspondance observée entre les données expérimentales et chaque entrée de la base de données est un événement fortuit.

La meilleure correspondance est celle qui présente la probabilité la plus faible. Le seuil de signification le plus utilisé est que la probabilité de l'événement observé se produisant par hasard soit inférieure à un sur vingt (p <0,05). En fait, MASCOT transforme la probabilité P par la formule $-10Log_{10}(P)$. Cela signifie que le meilleur candidat est celui qui a le score le plus élevé, et un bon candidat a typiquement un score supérieur à 50.

La comparaison du spectre expérimental avec ceux générés à partir d'une banque de séquences est couramment utilisée car elle est considérée plus fiable que l'approche exposée dans la sous-section 1.3.8.2. Néanmoins, elle présente l'inconvénient que seuls les peptides pour lesquels la protéine source est répertoriée peuvent être identifiés. L'identification de peptides porteurs de mutation, par exemple, est impossible.

Dans les études protéomiques, on utilise habituellement une enzyme, appelée trypsine, qui clive les protéines en position C-terminale des acides aminés Arg ou Lys. Pour une même protéine, on peut donc avoir plusieurs peptides pour lesquels on obtient les spectres de masse. Plus le nombre de peptides observés pour une protéine est important, plus son identification est probable. Si on observe 4 peptides d'une même protéine source, on peut être quasiment certain que cette protéine est effectivement présente dans notre échantillon. Ceci est moins vrai dans le cas où l'on n'a qu'un peptide. D'autant qu'une séquence peptidique n'est pas nécessairement associée à une seule protéine source [36]. Dans le cas qui nous intéresse, il y a rarement deux peptides du CMH-I de séquences peptidiques différentes pour une même protéine source. Ceci illustre l'une des difficultés présentées par les peptides du CMH-I.

1.3.8.2 Le séquençage *de novo*

Le séquençage *de novo* consiste à déterminer la séquence du peptide à partir du spectre MS/MS sans connaissance *a priori*. Il peut être formulé comme suit : pour une spectre expérimental S, un peptide de masse m, et un ensemble de types de fragments Δ , trouver une séquence (ou un ensemble de séquences) avec la masse m qui donne l'appariement maximal pour le spectre S. Choisir la meilleure correspondance implique l'implémentation d'un calcul de score. L'approche qui est utilisée par quasiment tous les algorithmes de séquençage de novo repose sur 2 étapes :

1. déterminer un ensemble de séquences candidates,

2. et déterminer un score pour chacun des candidats en fonction du spectre R.

La majorité des algorithmes de séquençage *de novo* utilisent des approches fondées sur les graphes spectraux. À chacun des pics d'un spectre est associé un sommet dans un graphe spectral. Deux sommets sont reliés par une arc si leur masse diffère de la masse d'un acide aminé. Chacun des pics du spectre expérimental est transformé en plusieurs sommets dans un graphe spectral où chacun d'entre eux représente une association plausible entre le pic et un type de fragment (y, b, etc.). La détermination de la séquence *de novo* revient donc à trouver le plus long chemin dans le graphe acyclique orienté. D'autres méthodes basées sur une recherche exhaustive impliquant la production de toutes les séquences peptidiques possibles ont été développées. Elles consistent à trouver le spectre théorique présentant le meilleur appariement avec le spectre expérimental. Vu le nombre important de séquences possibles, cette stratégie oblige l'utilisation de techniques d'élagage qui écarte parfois la bonne solution. Nous verrons au chapitre 4 des algorithmes qui utilisent des variantes de ces stratégies. D'autres approches intéressantes et prometteuses qui sont basées sur les chaînes de Markov cachées sont utilisées

(novoHMM [23]). Les variables aléatoires observables sont les pics de masse expérimentaux et les variables cachées correspondent à la séquence peptidique. Néanmoins, comme expliqué au chapitre 4, le programme novoHMM n'a pas été retenu. Le principal avantage du séquençage *de novo* par rapport à l'identification basée sur les banques de données est sa tolérance aux erreurs de séquence, aux mutations et aux délétions d'acides aminés. De plus, cette approche peut être couplée avec des algorithmes de recherche de similarités de séquences dans les banques de données. En fait, pour un processus de fragmentation idéal conduisant au clivage de toutes les liaisons peptidiques et un spectromètre de masse idéal (exempt d'erreur), le problème du séquençage *de novo* est relativement simple. Mais dans la pratique, c'est loin d'être idéal et par conséquent la séquençage n'est pas une tâche facile. Les spectres expérimentaux contiennent rarement tous les ions fragments et sont toujours bruités. De plus, l'imprécision sur la masse augmente le nombre de séquences peptidiques possibles de façon exponentielle.

1.4 La prédiction de peptides du CMH de classe I

La bioinformatique nous fournit des outils très utiles pour les investigations relatives aux peptides du CMH-I. Comme mentionnée dans la section 1.2.2, seule une petite partie des peptides issus de la dégradation des protéines dans le cytosol va être effectivement présentée à la surface de la cellule. On estime qu'en moyenne seul 0,5% de tous les peptides possibles d'une longueur de 9 résidues (soit $0,005 \times 20^9$) peuvent se lier aux molécules du CMH-I avec une grande affinité alors que 99% seront totalement ignorés [87]. Un peptide se retrouve finalement à la surface de la cellule parce qu'il correspond à un motif particulier. Chez l'humain, les molécules du CMH-I sont encodés par 3 différents loci appelé A, B et C. Chacun des ces gènes sont très polymorphiques.

Pour chaque locus, des centaines d'allèles existent. Et chacune des ces dernières se lient à un sous-ensemble spécifique de peptides. Les allèles ont été regroupés en 9 supertypes partageant grossièrement la même spécificité [43]. L'intérêt de ces supertypes, du point vue bioinformatique, est de pouvoir limiter la recherche de peptides du CMH-I potentiels à une sous-population d'allèles du CMH-I. Pour la plupart des molécules du HLA, le deuxième et le dernier résidu du peptide constituent des positions d'ancrage [65]. Les premiers algorithmes développés, notamment SYFPEITHI [65], partent de l'hypothèse que chacun des résidus le long de la séquence contribue indépendamment à une énergie de liaison [59]. L'énergie cumulée reflète l'affinité de liaison du peptide avec la molécule du CMH-I. Cependant, ces approches ne tiennent pas compte du fait que l'affinité d'un acide aminé peut être influencée par ses voisins. Des algorithmes basés sur les réseaux de neurones artificiels ont été développés et montrent les meilleures performances [46]. D'autres approches algorithmiques ont été mises en oeuvre telles que les chaînes de Markov cachées [89] ou les machines à vecteurs de support [17]. Nous verrons au chapitre 5 une utilisation de ces outils bioinformatiques pour évaluer la plausibilité qu'un peptide d'une séquence donnée se lie effectivement à la molécule du CMH-I.

CHAPITRE 2

LE PROGRAMME STATPEAKS

La spectrométrie de masse est d'une grande aide pour l'identification des peptides du CMH-I. Cependant, leur composition en acides aminés complique leur détection et leur séquençage. En effet, on observe que les spectres de ces peptides sont moins informatifs et ont des profils différents de ceux des peptides trypsiques. L'amélioration de l'interprétation de leurs spectres repose en premier lieu par la caractérisation de ceux-ci. À ce jour, aucune étude sur les profils de fragmentation des peptides du CMH-I n'a été publiée. Pour mener cette étude, un outil bioinformatique, appelé StatPeaks, a été développé par moi-même. Il calcule un ensemble exhaustif de statistiques relatives à la fragmentation des peptides à partir d'une librairie pouvant contenir des milliers de spectres expérimentaux pour lesquels on connaît les séquences peptidiques correspondantes. Bien que StatPeaks ait été développé initialement pour l'étude de la fragmentation des peptides du CMH-I, aucune restriction n'est imposée sur la nature ou l'origine des peptides. Par conséquent, il peut être utilisé dans le cadre d'autres études.

2.1 La détermination des spectres théoriques

La première tâche de StatPeaks est d'annoter les fragments observés dans les spectres expérimentaux. Pour ce faire, le programme calcule le spectre théorique correspondant à la séquence peptique associée au spectre expérimental. Le programme calcule la masse de chacun des ions pouvant résulter de la fragmentation d'un peptide donné sachant sa séquence S (a_1, a_2, ... a_n) de longueur n. Le calcul du spectre utilise les valeurs de masse les plus précises disponibles (`http://pubchem.ncbi.nlm.nih.gov`). L'utilisa-

tion du type double précision (double) est généralisée et la propagation d'erreurs d'arrondi est négligeable. On utilise un compilateur GNU Compiler Collection (GCC) permettant l'utilisation du codage double précision étendue.

Pour chaque sous-séquence possible de la séquence S de taille n contenant le résidu N-terminal (ions fragments a_k, b_k, ou c_k), finissant à la position k ($k \in [1;n-1]$) et simplement chargé, StatPeaks calcule les masses suivant les équations suivantes :

$$M_{a_k}(a_1,a_2,...,a_k) = \sum_{i=1}^{k} m(a_i) + m(H) - m(CO) \tag{2.1}$$

$$M_{b_k}(a_1,a_2,...,a_k) = \sum_{i=1}^{k} m(a_i) \tag{2.2}$$

$$M_{c_k}(a_1,a_2,...,a_k) = \sum_{i=1}^{k} m(a_i) + m(H) + m(NH_2) \tag{2.3}$$

où $m(a_i)$ est la masse du résidu à la position i, $m(X)$ est la masse de la molécule de formule chimique X. Pour chaque sous-séquence possible de la séquence S de taille n contenant le résidu C-terminal (fragments x_k, y_k, ou z_k) et commencant à la position k ($k \in [2;n]$), il calcule la masse suivant l'équation suivante :

$$M_{x_k}(a_k,...,a_n) = \sum_{i=n-k+1}^{n} m(a_i) + m(H) + m(O) + m(CO) \tag{2.4}$$

$$M_{y_k}(a_k,...,a_n) = \sum_{i=n-k+1}^{n} m(a_i) + m(H) + m(O) + m(NH_2) \tag{2.5}$$

$$M_{z_k}(a_k,...,a_n) = \sum_{i=n-k+1}^{n} m(a_i) + m(H) + m(O) - m(NH) \tag{2.6}$$

La masse du fragment interne commençant à la position j et se terminant à la position k est calculée suivant la formule suivante :

$$M_{j,k}(a_j,...,a_k) = \sum_{i=j}^{k} m(a_i) + m(H) \tag{2.7}$$

Pour chaque type d'ions énumérés plus haut, on calcule les ions avec perte de neutre en soustrayant la masse de l'ammoniaque (NH_3) et de l'eau (H_2O) aux masses calculées par les formules ci-dessus. StatPeaks calcule également la masse pour ces mêmes ions fragments doublement chargés.

La masse de l'ion immonium correspondant à l'acide aminé a et de chaîne latérale R_a se calcule par la formule suivante :

$$M_{Immonium}(a) = m(R_a) + 3 \times m(H) + m(N) + m(C) \tag{2.8}$$

où $m(R_a)$ est la masse de la chaîne latérale de l'acide aminé a.

La fragmentation à haute énergie, qui n'est pas utilisée dans le cadre de notre étude, peut produire d'autres types de fragments correspondant au clivage de la chaîne peptidique et de la chaîne latérale de l'acide aminé. Deux types de fragments découlent du clivage de la liaison entre les carbones β et γ de la chaîne latérale (R) des acides aminés. Si la charge positive est du côté N-terminal, on désigne le fragment d_n, si elle est du côté C-terminal, on le désigne w_n. Leurs masses respectives se calculent par les formules suivantes :

$$M_{d_k} = M_{a_k} - m(R_k) + m(H) \tag{2.9}$$

$$M_{w_k} = M_{z_k} - m(R_k) + m(H) \tag{2.10}$$

où $m(R_k)$ est la masse de la chaîne latérale de l'acide aminé a_k. Ces fragments permettent notamment de distinguer entre les acides aminés Leu et Ile. Les acides aminés qui ont un groupe aromatique rattaché au carbone β (His, Phe, Tyr, Trp) ne présentent que peu de fragments de ce type [32]. Un autre type de fragment résulte du clivage complet de la chaîne latérale. Il est désigné par v_n. Ce type de fragment est abondant lorsque le fragment w_n ne l'est pas [32]. Sa masse théorique se calcule par la formule suivante :

$$M_{v_k} = M_{y_k} - m(R_k) - m(H) \tag{2.11}$$

Notons que les acides aminés Thr et Ile contiennent deux chaînes latérales. Ces acides aminés peuvent donc donner deux fragments w_n de masse différente. Néanmoins, il faut mentionner qu'en basse énergie, on n'observe pas en principe les fragments d_k, w_k et v_k. Ils ne sont donc pas pris en considération dans nos analyses. Mais, il en est fait mention ici car StatPeaks calcule les pics théoriques correspondants et ils peuvent être utiles dans le cadre d'autres études.

2.2 La normalisation des spectres

La comparaison de spectres et l'étude des caractéristiques de fragmentation sur un ensemble de spectres nécessitent de normaliser ces derniers. L'intensité des pics pour un même peptide peut être différente d'une acquisition à une autre pour des raisons chimiques ou physiques et conduire à des courants ioniques totaux différents. La plus simple et la plus intuitive des approches consiste à normaliser le pic i d'intensité I_i par

rapport à l'intensité du pic ayant la plus grande intensité I_{max} suivant la formule suivante :

$$I_i norm = 100 \frac{I_i}{I_{max}} \tag{2.12}$$

Quand on s'intéresse à la distribution des intensités relatives des pics suite à cette normalisation, on observe un artéfact qui correspond à la sur-représentation des pics ayant une intensité relative de 100%. En effet, cette normalisation a pour conséquence que chacun des spectres contient un pic de 100%. Les spectres normalisés par cette approche contenant plus d'un pic d'intensité relative de 100% sont en pratique rarissimes. Une autre solution consiste à normaliser par le courant ionique total (somme de toutes les intensités mesurées) de telle sorte que la somme de toutes les intensités normalisées soit égale à 1 :

$$\sum I_i norm = 1 \tag{2.13}$$

Il a été montré que cette dernière solution montre de meilleurs résultats [4]. Une autre possibilité consiste à normaliser de façon à ce que la somme des carrés des intensités soit égale à 1 :

$$\sum I_i norm^2 = 1 \tag{2.14}$$

Cette dernière n'apporte pas de meilleurs résultats que la deuxième. De plus, elle demande plus de calcul. Statpeaks utilise les deux premières normalisations.

2.3 Les intervalles de tolérance de masse et l'assignation des pics

Les intervalles de tolérance se rapportent à la précision de masse du spectromètre. La précision de masse peut se définir comme la capacité de mesurer ou de calibrer la réponse de l'instrument pour une entité connue. La précision, exprimée en parties par million (ppm), indique la déviation de la réponse de l'instrument pour un masse monoisotopique calculée. Tous les spectromètres n'ont pas les mêmes précisions de masse, tant pour le précurseur que pour les fragments. La précision avec laquelle un spectromètre de masse peut mesurer le m/z d'un ion se situe entre quelques ppm dans le cas de la haute résolution (ex. : LTQ-Orbitrap) à plus de 500 ppm, dans le cas de basse résolution [55]. StatPeaks permet de préciser l'intervalle de tolérance et donc de réaliser des analyses avec n'importe quelle précision de masse. Un pic observé sera associé à un pic théorique à condition que son m/z se situe dans l'intervalle de tolérance spécifié.

La masse théorique calculée *in silico* de chaque ion est comparée à l'ensemble des masses du spectre MS/MS expérimental. Si une masse correspond à l'intervalle de tolérance près, StatPeaks assigne au pic le type d'ion correspondant. S'il arrive qu'un pic observé puisse correspondre à plusieurs pics théoriques équiprobables, on lui associe les différents types de fragments correspondants. Par exemple pour les peptides du CMH-I, on considère que les observations des fragments y et b sont équiprobables, et plus probable que celle des fragments a.

2.4 Le seuil du bruit

On distingue deux sources de bruit : chimique et électrique. Le bruit de fond d'origine chimique est dû à la présence des espèces moléculaires ionisées ne correspondant pas à celles qui sont recherchées. Différentes approches ont été proposées pour débruiter

les spectres avant l'identification des pics. Certains définissent le bruit comme la valeur médiane des intensités dans les spectres et fixent le seuil de sorte de ne garder que les pics ayant une intensité supérieure [14]. D'autres utilisent les parties négatives des données des spectres normalisés pour estimer la variance du bruit utilisé pour estimer le seuil du bruit [71]. Statpeaks laisse la possibilité à l'utilisateur de fixer le seuil qu'il souhaite. La valeur de seuil correspond au pourcentage de l'intensité du pic ayant la plus grande intensité ou du courant ionique total.

2.5 Le ré-échantillonnage

L'objectif de l'analyse de la fragmentation des peptides du CMH-I est d'établir des règles ou tout au moins des tendances générales à partir des quelques spectres dont nous disposons. Il est inconcevable et impossible de procéder à l'analyse de la fragmentation de la totalité des peptides possibles du CMH-I. Par exemple, le nombre de tous les peptides du CMH-I possibles de 8 résidus est égal à $0,005 \times 20^8 = 1,28 \times 10^8$ [87]. Ce nombre est trop important pour que l'on puisse tous les synthétiser et obtenir les spectres correspondants. Notre analyse se base donc sur un échantillon de quelques centaines de peptides du CMH-I. À partir de celui-ci, il est possible d'obtenir une moyenne estimée d'une mesure. En revanche, on ne peut connaître la dispersion de la mesure obtenue, ni l'intervalle de confiance, à moins de faire appel à des méthodes d'inférence statistique. On ne fait aucune hypothèse quant à la distribution fondamentale de la population des données, les méthodes de ré-échantillonnage sont donc indiquées dans le cas présent. Le bootstrap non paramétrique présente l'avantage de ne requérir aucune autre donnée que celle de l'échantillon. Elle consiste à construire B échantillons (sous-ensembles de notre échantillon initial que l'on appelle échantillons bootstrap) afin de procéder à des

inférences. Chaque échantillon bootstrap est construit en tirant un nombre m d'éléments provenant de l'échantillon initial et ce avec remise (c'est-à-dire qu'une fois l'élément tiré, il est remis et peut être retiré de nouveau). Chacun des éléments peut donc être tirés plusieurs fois ou pas une seule fois pour chaque échantillon. Le programme StatPeaks permet de spécifier le nombre B d'échantillons bootstrap. Le nombre de d'échantillons nécessaire peut être déterminé de façon empirique. En théorie, ce nombre devrait être infini. Mais en réalité, l'utilité du bootstrap est qu'il converge relativement rapidement et donc un nombre fini d'échantillons est suffisant. Une façon de vérifier que le nombre d'échantillons est suffisant est de s'assurer de la reproductibilité des résultats obtenus avec un même nombre d'échantillons. Il faut noter qu'un trop grand nombre d'échantillons bootstrap aura un coût informatique. On conseille un nombre minimal de simulations de 1000. L'échantillon aléatoire initial de données de taille n est noté par X = $(X_1, ..., X_n)$. Il constitue un ensemble des variables aléatoires $X_1, ..., X_n$ indépendantes. L'échantillon est utilisé pour inférer (estimer) une caractéristique de la population, notée θ (par exemple, l'intensité des ions fragments y_1) dont la valeur estimée pour l'échantillon est notée $\hat{\theta}_n$. Pour déterminer $\hat{\theta}_n$, on calcule un estimateur pour chacun des B échantillons, notés $\hat{\theta}_n^b$. Les B valeurs $\hat{\theta}_n^b$ constituent une approximation de la distribution fréquentiste de l'estimateur. On s'intéresse alors à calculer pour $\hat{\theta}_n$: l'écart-type (la dispersion) et les intervalles de confiance déterminés par les quantiles de la distribution. Le but de l'intervalle de confiance est d'indiquer la confiance que l'on peut avoir dans la valeur de t. StatPeaks est configuré pour utiliser un intervalle de confiance à un seuil de 5%. Notre intervalle de confiance est donc délimité par les quantiles 2,5% et 97,5% de la distribution bootstrap. Notons que la taille de l'échantillon initial ne doit pas être trop petite (n > 2 est conseillé). Aussi, la non-symétrie de la distribution bootstrap doit éveiller des soupçons qu'en aux paramètres utilisés ou la taille des échantillons.

2.6 Le filtrage sur la séquence

Il est possible de procéder aux analyses statistiques sur un sous-ensemble de séquences partageant un motif. Par exemple, on peut lancer une analyse sur les seuls peptides de 8 résidus se terminant par un acide aminé Tyr en utilisant le motif suivant : `.......Y`. Statpeaks filtre les fichiers en fonction de ce motif. Cette option permet notamment de mettre en évidence l'influence d'un acide aminé particulier à une position donnée. Le tableau 6.I liste des exemples de recherche par motif.

Tableau 2.I – Exemples de filtrage par motif

Motif	Résultat de la recherche
$*$	tous les peptides sans distinction
$\ldots\ldots\ldots$	les peptides de 8 résidus
$\ldots\ldots\ldots L$	peptides de 8 résidus se terminant par L
$.*L$	tous les peptides se terminant par L
$L.*$	tous les peptides se commençant par L
$.*K\|.*R$	les peptides se terminant par K ou R
$K.*\|F.*$	les peptides commençant par K ou F
$\ldots R\ldots$	les peptides de 7 résidus ayant un R en 4ème position

2.7 Les différentes statistiques calculées

Statpeaks calcule différentes statistiques afin de comparer les spectres d'un ensemble de peptides avec un autre ensemble (par ex. : peptides trypsiques *versus* peptides du CMH-I). Il fournit aussi des statistiques pour comprendre l'influence de chacun des acides aminés sur la fragmentation des peptides.

2.7.1 La complétude et l' incomplétude de la fragmentation

On définit ici le concept de complétude et son opposé : l'incomplétude. On dit qu'un spectre est complet pour indiquer qu'aucune information additionnelle n'est utile à son

interprétation. Par exemple, si pour un peptide de n résidus le spectre expérimental contient tous les fragments y_k avec $k \in [1, n-1]$ alors il s'agit d'un spectre complet. S'il manque un clivage, le spectre à un incomplétude égale à $\frac{1}{n-1}$. Pour un peptide de n résidus, l'incomplétude $\mathfrak{I}$ est définie par la formule suivante :

$$\mathfrak{I} = \frac{\sum_{k=1}^{n-1} min(e(y_k), e(b_k))}{n-1} \tag{2.15}$$

où

$$e(y_k) = \begin{cases} 0 \; si \; y_k \; existe \\ 1 \; sinon \end{cases} \tag{2.16}$$

et

$$e(b_k) = \begin{cases} 0 \; si \; b_k \; existe \\ 1 \; sinon \end{cases} \tag{2.17}$$

On formule donc la complétude C par $C = 1 - I$.

2.7.1.1 Les rapports de chacun des types de fragments

Par rapport au nombre total de pics StatPeaks calcule pour chacun des types de fragments leur proportion par rapport au nombre total de pics. Il indique également le nombre de pics non assignés, c'est-à-dire les pics pour lesquels il ne fournit pas d'explications sur leur origine.

Par rapport à la longueur du peptide Statpeaks détermine le rapport du nombre d'ions de chaque type par la longueur du peptide moins 1. Pour un type de fragment x, f_x est désigné pour représenter ce rapport. Pour les ions de type a, b, c, x, y et z,

une valeur de 100% signifie que tous les ions du type donné sont observés. $N(x)$ est le nombre de pics correspond à l'ion de type x. Pour un peptide de n résidus, ce nombre est au maximum égal à $n-1$. f_x se calcule par la formule suivante :

$$f_x = 100 * \frac{N(x)}{n-1} \tag{2.18}$$

Pour un échantillon de m peptides et pour un type d'ions x, StatPeaks calcule la moyenne du $\bar{f}_x$ suivant la formule :

$$\bar{f}_x = \frac{\sum_1^m f_x}{m} \tag{2.19}$$

Le nombre de fragments internes pour un peptide de n résidus est compris entre 0 et $\sum_{k=2}^{n}(n-k)$. Ce nombre peut donc excéder la valeur $n-1$. Par conséquent, on peut donc obtenir une valeur supérieure à 100% pour ce type de fragments. L'intervalle de confiance est calculé par la technique du bootstrap.

2.7.1.2 Le profil de fragmentation

Le profil de fragmentation renseigne sur l'intensité relative de chacun des ions fragments d'un type particulier. Pour un échantillon de m peptides d'exactement n résidus, Statpeaks calcule pour $i \in [1;n-1]$, les intensités moyennes $\bar{I}_i(x)$ des fragments de type x ($x \in \{a,b,c,d,w,x,y,z\}$). Pour chaque $\bar{I}_i(x)$, un intervalle de confiance est calculé par la technique du bootstrap.

2.7.2 L'influence des résidus adjacents au site de clivage

Statpeaks calcule l'effet que chaque acide aminé a sur le site de clivage et l'ion associé lorsqu'il est adjacent à celui-ci, du côté C-terminal ou N-terminal. Pour calculer la fréquence de clivage $\Phi_T(a)$ pour chaque acide aminé a, au site correspondant au

terminus $T \in \{N, C\}$, il compte le nombre de fois qu'un ion est observé correspondant au clivage pour cet acide aminé $O_T(a)$, et divise par le nombre d'occurrences du résidu dans l'ensemble des séquences $O(a)$. La technique du bootstrap est utilisée pour déterminer l'intervalle de confiance.

$$\Phi_T(a) = \frac{\sum O_T(a)}{\sum O(a)} \tag{2.20}$$

2.7.2.1 L'influence des résidus non-adjacents au site de clivage

Statpeaks permet d'évaluer l'influence des résidus non-adjacents au site de clivage sur la production des fragments. Il calcule le nombre de fois que chaque acide aminé a est présent à l'intérieur des fragments identifiés à l'exception de la position adjacente au site de clivage et divise par le nombre de fois que l'acide aminé est présent à l'intérieur de tous les fragments possibles. L'intervalle de confiance est déterminé par la technique du bootstrap.

2.7.2.2 Le N-biais par acide aminé

Statpeaks calcule l'effet de chacun des acides aminés sur la directionnalité de la fragmentation. Pour ce faire, on utilise la méthode utilisée par Tabb et al. [77]. Pour chacun des résidus dans le peptide, le programme cherche dans le spectre les pics des ions produits par la fragmentation à chacun des côtés C-terminal et N-terminal. Pour calculer le N-biais ($\mathscr{N}$), le programme soustrait l'intensité du pic C-terminal au pic N-terminal pour un type d'ion donné (ex. : b ou y), et divise par la somme des deux intensités. Par exemple, pour calculer le N-biais $\mathscr{N}(Pro)_b$ de l'acide aminé Pro pour les ions b on utilise la formule suivante :

$$\mathscr{N}(Pro)_b = \frac{I_b(i) - I_b(i+1)}{I_b(i) + I_b(i+1)} \tag{2.21}$$

où $I_b(i)$ est l'intensité de l'ion b produit par le clivage au côté N-terminal de l'acide aminé Pro, et $I_b(i+1)$ est l'intensité de l'ion b produit par le clivage au côté C-terminal. Si le pic pour l'ion b produit du côté C-terminal (N-terminal) n'est pas trouvé, le N-biais est égal à 1 (-1). On calcule ensuite la moyenne sur l'ensemble des spectres et on applique la méthode du bootstrap pour déterminer l'intervalle de confiance.

2.7.3 La carte de fragmentation par paires

Statpeaks calcule la fréquence de clivage entre les résidus des 400 couples d'acides aminés possibles ainsi que leur abondance (autrement dit, l'intensité moyenne des ions fragments résultants). Les résultats sont représentés sous forme de matrice. Les lignes représentent les résidus en position N-terminal du site du clivage alors que les colonnes représentent les résidus en position C-terminal. Comme montré sur la figure 2.1, chaque case de la matrice est remplie proportionnellement au nombre d'occurrences du couple de résidus correspondant dans l'ensemble des données. Lorsqu'il y a au moins 10 occurrences, la case est pleine. S'il y en a 5, la case est à moitié pleine. Ceci permet de représenter la confiance que l'on a dans chacune des valeurs de la matrice.

2.8 L'utilisation du programme

Le programme prend en entrée les spectres ainsi que les séquences des peptides correspondants. Il existe plusieurs formats pour présenter les données relatives aux pics des spectres MS/MS : PKT, PKM, DTA, MGF, mzXML, etc. Chacun d'entre eux contient la valeur m/z et l'intensité pour chaque pic ainsi que le m/z du précurseur et sa charge.

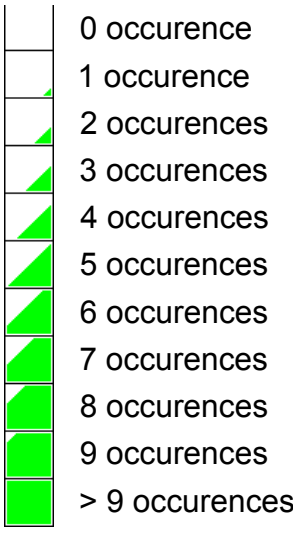


Figure 2.1 – Légende des cartes de fragmentation par paires Le remplissage des casses de la matrice est proportionnel au nombre d'occurrences du couple de résidus correspondant dans l'ensemble des séquences peptidiques.

D'autres informations peuvent être disponibles dans certains formats de fichier mais celles-ci ne sont pas utilisées. Dans la version alpha actuelle, seul le format de fichier DTA, dont le format est décrit dans le tableau 2.II, est accepté. Le fichier contenant la séquence correspondante au spectre décrit dans un fichier DTA porte l'extension SEQ. Le fichier contient une seule ligne sur laquelle est indiquée la séquence peptidique avec le code à une lettre. Le programme StatPeaks génère un certain nombre de fichiers aux premiers desquels les fichiers THC et IDE. Le premier contient le spectre théorique correspondant à la séquence indiquée dans le fichier SEQ. Celui-ci a le même format que le fichier DTA sans la première ligne. Le deuxième contient la liste des pics théoriques pour lesquels sont indiqués les pics observés correspondants. Le type d'ion fragment, la charge, l'intensité absolue, l'intensité relative par rapport au pic d'intensité maximale et par rapport à la somme des intensités, la perte de neutre, la séquence du fragment correspondant s'il s'agit d'un fragment interne, l'acide aminé s'il s'agit d'un ion immonium

sont renseignés.

1823.78	2
113.34	654
125.10	345
238.91	1120
315.67	619
450.73	8415
517.33	10765
599.12	312
610.08	9371
780.76	6541
980.36	3201
1100.95	1201

Tableau 2.II – Exemple de fichier DTA. La première ligne contient la masse et la charge du peptide. Les lignes suivantes contiennent les valeurs m/z (à gauche) et la valeur de l'intensité associée (à droite).

Le programme dispose d'une interface utilisateur permettant son utilisation sans connaissance préalable des instructions de ligne de commande ou du système d'exploitation Linux. Celle-ci permet de spécifier les statistiques à calculer et renseigne sur l'avancement des différentes tâches en cours.

2.9 L'implémentation

StatPeaks est développé dans le langage de programmation C++. La librairie Boost, un ensemble de bibliothèques écrit en C++ sous licence libre, a été utilisée notamment pour l'utilisation des expressions régulières et la génération des nombres aléatoires utilisés par le rééchantillonnage bootstrap. L'interface utilisateur a été conçue en Gtk+ (version 2.22.0) qui a le double avantage d'être sous licence libre et d'être multiplateforme. Le développement de l'interface a été facilité par l'utilisation de l'application Glade (version 3.6.6). La génération de certains fichiers images est basée sur l'utilisation bas niveau (pixel par pixel) de la librairie libpng (version 1.4.5). Le développement est fait

dans l'environnement de développement intégré Code::Blocks. Il existe une version graphique de StatPeaks pour le système d'exploitation Windows et une version ligne de commande pour Linux. StatPeaks est au moment de la rédaction de ce mémoire en version alpha.

2.10 Conclusion

Avec StatPeaks, nous disposons d'un programme bioinformatique permettant d'analyser la fragmentation d'une population de peptides à partir d'un grand nombre de spectres. Il ne requiert aucune connaissance en informatique. Dans le prochain chapitre, nous utilisons ce programme pour l'étude de la fragmentation des peptides du CMH-I.

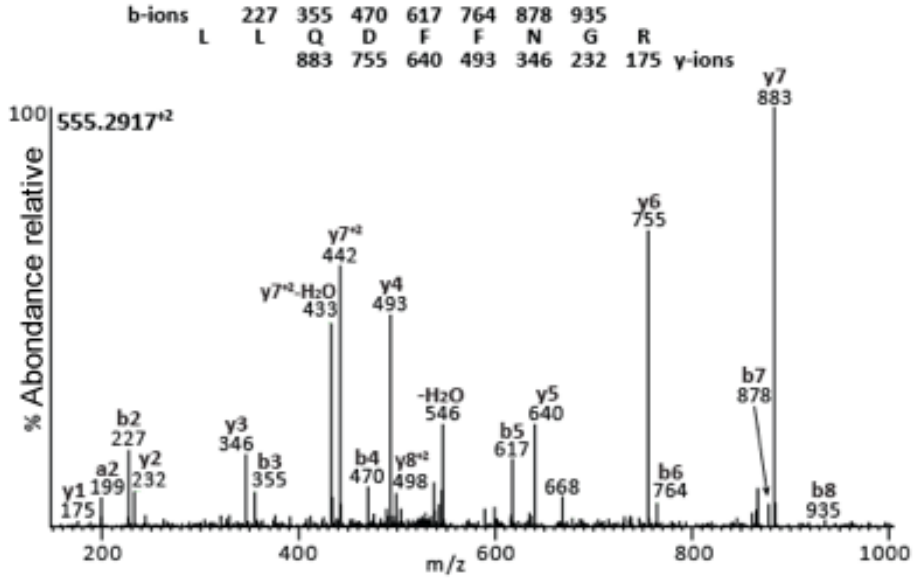
CHAPITRE 3

LA FRAGMENTATION DES PEPTIDES DU CMH DE CLASSE I

L'identification des peptides par la spectrométrie de masse en tandem repose sur la connaissance des ions fragments produits par la fragmentation. À ce jour, aucune description exhaustive de la fragmentation des peptides du CMH-I n'a été faite. Or, la composition en acides aminés de ces peptides affecte leur fragmentation et par voie de conséquence la qualité du spectre obtenu. Habituellement, dans les analyses protéomiques, les protéines sont digérées par la trypsine. Cette enzyme est une endoprotéase qui hydrolyse les liaisons peptidiques en position C-terminale des acides aminés Lys et Arg. Ces acides aminés ayant un pKa > 7 ont la particularité d'être protonés en milieu acide (PH = 3) lors des analyses LC-MS/MS. Tous les peptides trypsiques ont donc une charge du côté N-terminal (sur l'amine) et une charge du côté C-terminal. En conséquence, la fragmentation et l'identification de ces peptides sont grandement facilitées. La distribution de la charge à travers la séquence peptidique est d'une grande importante pour la fragmentation. Comme le montre la figure 3.2, le spectre MS/MS CID d'un peptide trypsique est de bonne qualité. La charge localisée en C-terminale du peptide trypsique conduit à un nombre de fragments *y* uniformément répartis. On peut aussi observer des ions *a* et *b* même s'ils ont des intensités moindres. La complémentarité des ions *y* et *b* (ou *a*) facilite l'assignation des pics et conduit à de bons scores MASCOT et à un séquençage *de novo* plus aisé.

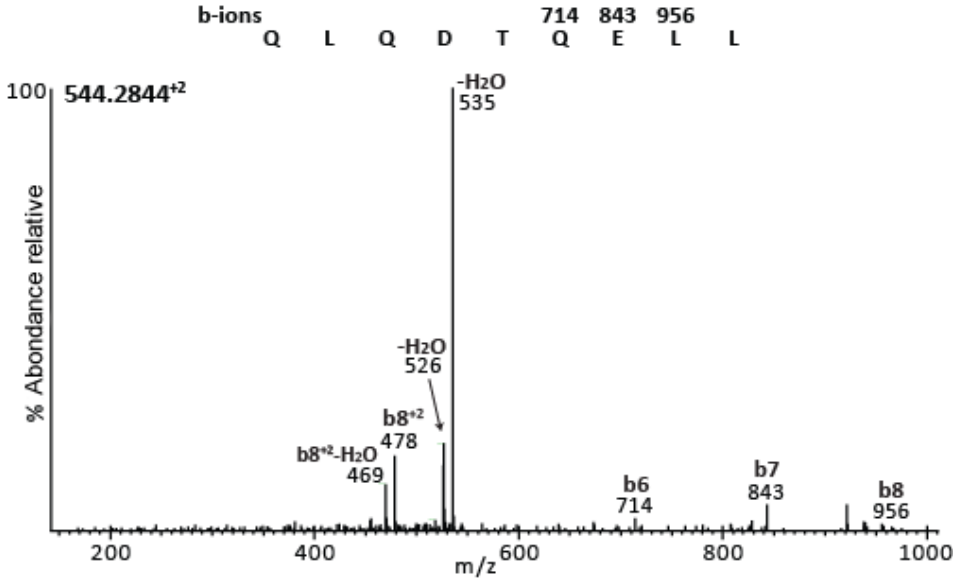
En revanche, les peptides du CMH-I se terminent rarement par un acide aminé basique mais le plus souvent par des acides aminés aliphatiques (ex. : Leu, Ile ou Val). Leur fragmentation en est donc affectée. La localisation de la charge en position N-terminal

Figure 3.1 – Spectre MS/MS d'un peptide trypsique de la protéine HSP70 de souris



est favorable aux ions *b*. La faible quantité de fragments conduit à des scores MASCOT relativement faibles et rend le séquençage *de novo* plus difficile.

Figure 3.2 – Spectre MS/MS d'un peptide du CMH de classe I de la myosine de souris



Dans ce chapitre, on s'intéresse à caractériser le profil particulier de fragmentation

des peptides du CMH-I et à le comparer avec celui des peptides trypsiques. Il existe deux approches pour déterminer et comprendre les relations entre la séquence peptidique et le spectre produit :

1. l'approche du point de la chimie qui repose sur l'investigation des processus chimiques mis en oeuvre lors de la fragmentation,

2. et l'approche statistique qui se base sur l'analyse d'un ensemble de spectres MS/MS pour identifier les facteurs influençant la fragmentation tels que les propriétés des peptides sans pour autant comprendre les processus chimiques impliqués.

C'est cette seconde approche que nous utilisons ici. Nous utilisons le programme StatPeaks décrit précédemment dans le chapitre 2. Cependant, lorsque des processus chimiques connus viennent corroborer nos investigations, ils sont mentionnés. On s'est intéressé à deux modes de fragmentations : CID (collision induced dissociation) et HCD (Higher energy Collision Dissociation). On met en évidence que ce dernier mode de fragmentation permet d'obtenir des spectres plus informatifs.

3.1 La librairie de peptides synthétiques du CMH de classe I

Un certain nombre d'études visant à identifier les différents processus de fragmentation des peptides trypsiques ont été effectuées par le passé [37]. Elles ont permis notamment de mettre en évidence l'influence de chacun des acides aminés sur la fragmentation. Ici, on s'intéresse à effectuer le même type d'analyse sur les peptides du CMH-I. Pour ce faire, nous disposons d'une librairie de peptides synthétiques de séquences connues. Celle-ci nous a été fournie par le Dr. Alexandro Sette (La Jolla Inst. Allergy & Immunol.)

3.1.1 La description de la librairie

La librairie de peptides synthétiques dont nous disposons contient un total de 625 peptides répartis en 5 groupes de 125 peptides pour lesquels nous disposons des séquences peptidiques. 2 groupes correspondent à des peptides provenant de la souris pour les allèles H2-Db et H2-Kb, et 3 groupes pour des peptides provenant de l'homme pour les allèles HLA-A*01, HLA-A*02 et HLA- A*03. Les peptides synthétiques sont des peptides antigéniques de protéines endogènes de l'homme et de la souris ou de protéines provenant de différents pathogènes listés en annexe VII. Cette librairie est de taille modeste au regard de celles qui ont été utilisées dans d'autres études [37]. La plupart d'entre elles se basent sur plusieurs milliers de spectres. De plus, seul un sous-ensemble de la librairie de 625 peptides pourra être exploité pour les analyses, comme nous l'expliquerons plus loin. Pour la fragmentation CID, nous disposons d'un total de 490 peptides. Quant à la fragmentation en HCD, nous n'en disposons que de 419. Les peptides correspondent à des ORFs connus ou prédits. Ils ont été sélectionnés en fonction de leur affinité prédite de liaison avec la molécule du CMH-I. Tous les peptides sont considérés comme ayant une bonne affinité ($IC_{50} \leq 500$). Nos analyses se basent sur l'hypothèse selon laquelle l'ensemble des peptides synthétiques est représentatif de la population de peptides effectivement présentés à la surface de la cellule. Nous avons vérifié cette hypothèse en comparant la composition en acides aminés des peptides synthétiques avec ceux répertoriés dans la littérature. Pour ce faire nous avons utilisé l'ensemble des peptides du CMH-I correspondant aux 5 groupes répertoriés dans la base de données IED [83]. Les fréquences de chacun des acides aminés sont corrélées (spearman $r = 0.9, p \leq 0.05$). Nous avons également comparé la fréquence de chacun des couples d'acides aminés parmi les peptides de notre librairie et ceux répertoriés dans la base de données de l'IED.

On peut constater que pour la majorité des couples (70,75%), nous avons un rapport de fréquence compris entre 0,5 et 2 (figure 3.3). Par conséquent, on peut dire que notre librairie de peptides est représentative du répertoire des peptides du CMH-I effectivement présentés à la surface des cellules.

	L	I	V	A	S	F	G	T	N	Y	E	P	D	K	M	R	Q	H	C	W
L	1,0	1,8	2,0	1,8	1,1	1,4	3,7	3,3	1,0	1,0	0,6	1,0	1,0	1,2	0,5	1,3	1,4	1,3	0,2	0,4
S	1,5	3,5	4,8	1,9	1,1	1,7	3,2	1,8	2,0	1,4	6,4	5,4	1,8	2,1	1,7	1,2	2,1	1,6	0,4	
F	1,0	2,3	1,6	3,8	1,0	3,9	4,0	2,5	0,7	1,7	1,3	1,8	2,1	2,7	1,5	3,2	2,0	1,8	1,2	
V	1,0	2,5	1,4	1,6	1,1	2,1	3,1	1,9	0,8	1,6	1,2	1,1	0,5	1,6	1,0	4,8	1,2	0,7		
A	1,6	1,7	1,4	0,9	1,3	2,3	3,0	2,1	1,9	1,1	1,5	0,8	1,9	3,4	1,0	2,9	2,8	1,0	1,9	
I	1,0	2,1	1,2	1,3	1,2	2,4	3,0	1,2	1,3	0,9	1,1	1,5	0,4	0,7	0,5	1,9	2,0	1,2	1,1	0,0
G	1,2	1,7	0,9	1,5	1,3	1,3	3,0	1,1	1,5	0,6	1,5	0,7	1,2	1,9	0,7	0,5	1,0	2,0	1,2	
T	1,2	2,4	2,8	1,0	0,8	2,6	4,0	0,9	1,1	1,9	1,7	1,5	1,8	1,2	0,4	1,3	0,9	3,2	0,8	0,0
N	0,8	2,4	2,3	1,6	0,7	0,4	1,9	1,4	0,5	0,4	1,9	2,0	0,7	1,2	1,0	1,6	0,9	1,0	0,6	0,3
Y	1,3	5,3	4,2	1,5	1,7	2,1	1,5	1,8	0,6	1,9	3,6	4,8	2,6	1,3	1,4	2,1	0,9	2,1		0,5
P	1,5	3,1	3,8	1,2	0,5	2,2	1,3	0,9	1,1	1,9	1,0	7,5	1,2	1,3	1,4	0,9	1,4	1,5	0,6	
E	1,7	1,1	0,8	1,1	1,2	1,7	4,7	1,6	1,2	1,3	1,1	1,8	1,3	0,8	0,6	0,9	1,0	1,8		
D	0,8	2,6	1,8	1,1	1,8	2,4	2,2	2,4	1,3	0,6	1,4	2,3	1,6	1,9	2,4	4,0	1,4	2,1		
K	2,7	3,1	4,5	2,0	1,1	1,6	2,7	0,9	1,7	1,7	0,9	2,7	2,1	1,6	1,5		3,5		0,5	1,1
R	2,6	3,6	1,8	0,7	1,1	1,4	3,5	1,2	1,0	1,4	3,5	4,6	1,4	1,0	1,3	2,9	3,0	1,1	4,2	
Q	0,5	1,3	1,3	0,3	0,9	1,8	1,5	1,0	1,1	0,7	0,7	1,5	1,3	1,2	0,5	0,5	1,6	0,7	0,4	
M	1,3	1,6	3,2	1,8	0,6	1,5	1,7	1,1	1,0	1,8	2,5	1,1	1,4	0,7	1,4	2,9	2,0	2,0		1,6
H	2,1	3,7	2,0	1,9	2,0	1,7	0,8	1,9	0,7	0,4	0,5	1,8	1,2	1,4		3,8	1,1	1,6		
C	0,8			1,2	0,4	1,2	1,3		0,5	0,7	0,5		1,1	3,8	0,0	2,7	1,6	1,6	3,2	
W	0,1	0,3		0,0	0,0		0,6	0,8		0,8	1,1		0,3		0,0	1,3		1,6		

Figure 3.3 – Rapport de fréquence de chacun des couples d'acides aminés entre les peptides synthétiques du CMH-I et ceux de la base IED. Le rapport de fréquence est compris entre 0,5 et 2 pour la majorité des couples d'acides aminés.

3.1.2 La représentativité des acides aminés

La figure 3.4 montre pour chacun des 400 couples d'acides aminés possibles, leur fréquence dans la librairie de peptides synthétiques dont nous disposons. On peut constater que nous avons une bonne représentativité pour les couples impliquant les acides aminées suivants : Leu, Ser, Thr, Val, Tyr, Lys, Asn, Ala, Phe et Glu. En revanche, peu de couples contenant les acides aminés Trp et Cys sont représentés. Ces deux acides aminés sont effectivement moins fréquents dans les séquences peptidiques du protéome avec une fréquence d'abondance respective de 1,15% et 1,5% [3]. La représentativité des

acides aminés dans notre librairie a un impact sur les analyses statistiques. Par exemple, la précision de l'évaluation de l'influence de chacun des acides aminés est affectée par celle-ci.

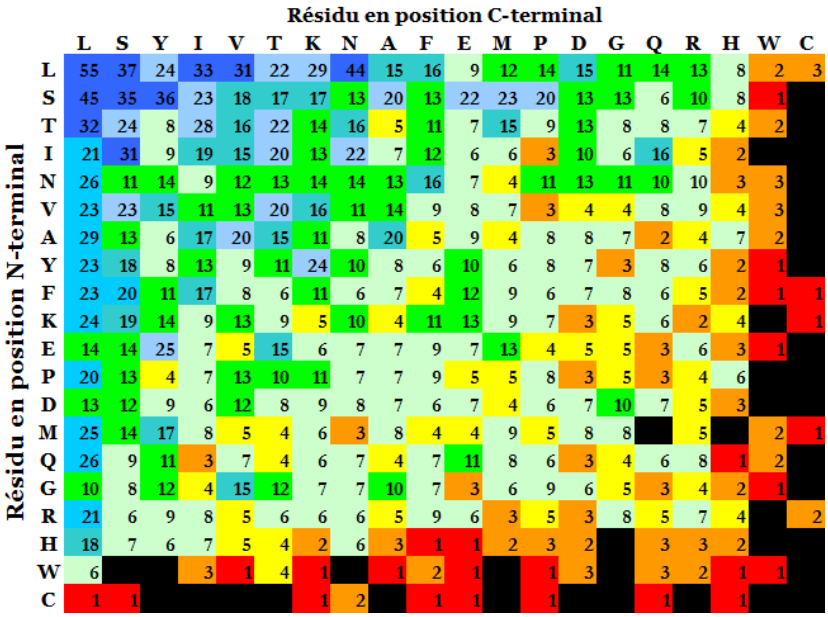


Figure 3.4 – Fréquence de chacun des 400 couples d'acides aminés parmi les peptides synthétiques. Les couples impliquants les acides aminés Leu, Ser, Thr, Val, Tyr, Lys, Asn, Ala, Phe et Glu sont les plus représentés. Alors que ceux contenant Trp ou Cys sont plus rares.

3.1.3 La construction des librairies de spectres MS/MS

La construction des librairies de spectres CID et HCD nécessite deux étapes :

1. générer les spectres MS/MS,

2. associer à chacun des spectres la séquence du peptide correspondant.

Pour la deuxième étape, nous avons utilisé le moteur de recherche MASCOT. Les échantillons contiennent inévitablement des contaminants. De plus, la synthèse peptidique

peut conduire à plusieurs types d'erreur : épimères, insertions d'acides aminés additionnels, délétion d'acides aminés, troncatures de séquences et modifications de la chaîne latérale des acides aminés. Les analyses de la librairie ont mis en évidence un nombre important d'erreurs de synthèse. Parmi les plus fréquentes, il y a les insertions et les troncatures. L'utilisation de MASCOT permet de ne conserver que les spectres pour lesquels les séquences peptidiques associées correspondent effectivement aux séquences connues des peptides du CMH-I.

3.1.3.1 La librairie de spectres CID de peptides du CMH-I

Les peptides synthétiques ont été analysés par nanoLC-MS/MS sur le spectromètre de masse LTQ-Orbitrap XL de la compagnie Thermo Fisher Scientific [24]. Les spectres de masse conventionnels ont été acquis avec l'analyseur Orbitrap avec une résolution de 60000 (pour un m/z de 400) alors que les spectres MS/MS ont été obtenus dans la trappe ionique (LTQ) de résolution d'environ 1000. Les données ont été analysées avec le logiciel Xcalibur de Thermo Fisher Scientific (version 2.1) et les listes de pics ont été générées par le logiciel MASCOT distiller (version 2.3.2.0). L'identification des séquences associées à chacun des spectres MS/MS a été faite à l'aide du logiciel MASCOT (version 2.1.1) avec une base de données restreinte à la liste des 625 séquences des peptides synthétiques. La tolérance en masse pour le précurseur est fixée à 0,02 Da alors que celle pour les fragments est de 0,5 Da. Les recherches ont été faites avec aucune enzyme et avec l'oxydation de la méthionine comme modification variable. Un total de 490 sur 625 peptides ont pu être identifiés avec un score MASCOT supérieur à 25, soit 78,4 %. Le tableau 3.I présente les résultats de la recherche MASCOT. Plusieurs spectres sont associés à un même peptide. Pour chacun des peptides identifiés, seul le spectre ayant le meilleur score MASCOT est conservé. On dispose donc de 490 spectres pour lesquels

on connaît la séquence correspondante.

Tableau 3.I – Statistiques de la recherche MASCOT pour les peptides synthétiques du CMH-I en fragmentation CID avec une base restreinte aux 625 séquences.

	HLA-A01	HLA-A02	HLA-A03	H2-Db	H2Kb
Score MASCOT moyen	48,75	45,4	49,26	45,26	41,11
Écart-type	17,18	15,17	16,67	16,6	15,42
Nombre total de spectres	748	782	1101	910	507
Nombre de spectres assignés	529	523	685	635	335
Pourcentage de spectres assignés	70,7	66,9	62,2	69,8	66,1
Nombre de peptides identifiés	90/125	86/125	105/125	109/125	100/125
Pourcentage de peptides identifiés	72	68,8	84	87,2	80

3.1.3.2 La librairie de spectres CID de peptides trypsiques

Un ensemble de spectres MS/MS CID de peptides trypsiques a été constitué afin de comparer leur fragmentation avec celle des peptides synthétiques du CMH-I avec ceux-ci. Les peptides proviennent de digestats trypsiques de protéines issues de l'étude du phagosome chez la souris [79]. Les données ont été analysées par la même méthode que pour la librairie de peptides du CMH-I. Nous avons retenu des spectres associés à des peptides d'une longueur comprise entre 8 et 12 résidus afin que le facteur taille du peptide n'introduise pas de biais dans les analyses. L'ensemble est exempt de doublons et de spectres associés à des sous-séquences de séquences déjà présentes dans cet ensemble, afin de ne pas biaiser les analyses par une sur-représentation des séquences identiques ou similaires. Nous nous sommes ainsi assurés que notre échantillon est représentatif de l'ensemble des peptides trypsiques possibles et que chacune des observations peut être considérée comme indépendante. De plus, n'ont été retenus que les spectres pour lesquels les identifications étaient associées à un score MASCOT supérieur à 50 pour réduire le taux de faux positifs. L'ensemble contient 2328 spectres MS/MS CID associés à des séquences uniques.

3.1.3.3 La librairie de spectres HCD de peptides du CMH-I

Les peptides synthétiques ont été analysés par nanoLC-MS/MS sur le spectromètre de masse LTQ-Orbitrap Velos de la compagnie Thermo Fisher Scientific. Les spectres MS/MS suite à l'activation par collision des ions précurseurs par HCD et les fragments correspondants ont été analysés par l'Orbitrap. Les données ont été analysées par la même méthode que pour la librairie de spectres CID. La tolérance de masse pour le précurseur est fixée à 0,01 Da alors que la tolérance pour les fragments est de 0,02 Da. Les recherches ont été faites avec aucune enzyme et avec l'oxydation de la méthionine comme modification variable. 415 peptides sur 625 ont pu être identifiés, soit 66,4 %. Le tableau 3.II présente les résultats de la recherche MASCOT. Plusieurs spectres sont associés à un même peptide. Pour chacun des peptides identifiés, seul le spectre ayant le meilleur score MASCOT est conservé. On dispose donc de 415 spectres pour lesquels on connaît la séquence correspondante.

Tableau 3.II – Statistiques de la recherche MASCOT pour les peptides synthétiques du CMH-I en fragmentation HCD avec une base restreinte aux 625 séquences.

	HLA-A01	HLA-A02	HLA-A03	H2-Db	H2Kb
Score MASCOT moyen	41,4	37,96	56,37	39,38	36,38
Écart-type	14,16	15,52	15,71	13,77	10,92
Nombre total de spectres	666	759	1543	795	815
Nombre de spectres assignés	518	542	1010	597	568
Pourcentage de spectres assignés	77,8	71,4	65,5	75,1	69,7
Nombre de peptides identifiés	79/125	72/125	99/125	77/125	88/125
Pourcentage de peptides identifiés	63,2	57,6	79,2	61,6	70,4

3.1.3.4 La comparaisons des librairies de spectres CID et HCD des peptides du CMH-I

On peut noter que la fragmentation HCD a conduit à 15,3% moins d'identifications que la fragmentation CID. Il se pourrait que ce soit dû à la sensibilité moindre du HCD.

Néanmoins rien n'est sûr. Il faudrait probablement procéder à d'autres analyses pour confirmer ou infirmer cette hypothèse. Les analyses en mode HCD ont été menées environ une année après celles faites en mode CID. La conservation des échantillons pourrait être incriminée et notamment les cycles de gel-dégel. Comme le montre la figure 3.5 une large proportion des peptides détectés en fragmentation HCD le sont en fragmentation CID.

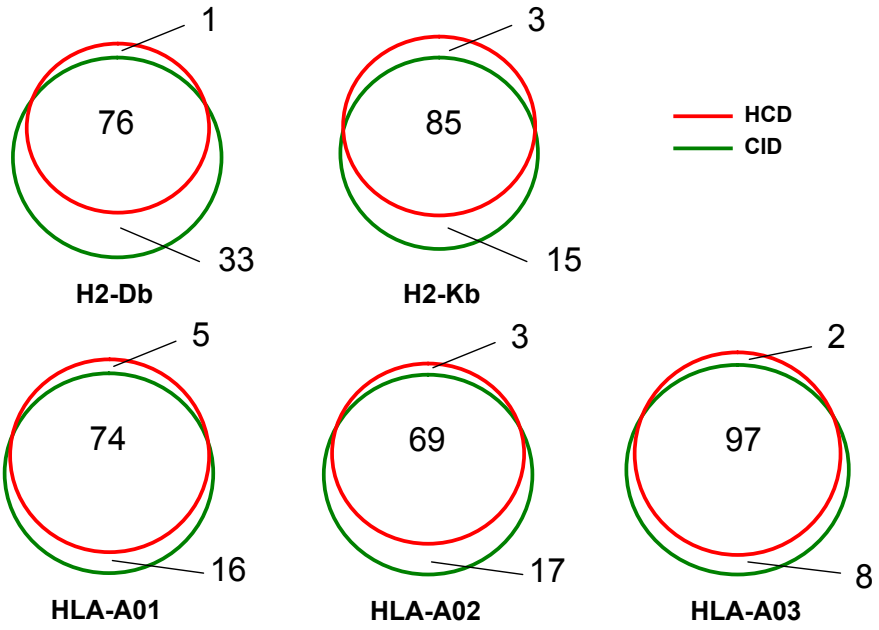


Figure 3.5 – Comparaison de la proportion de peptides identifiés par fragmentation CID et fragmentation HCD. Les peptides identifiés dans l'un et l'autre des modes de fragmentation sont sensiblement les mêmes. Cependant, un plus grand nombre de peptides sont identifiés en mode CID.

3.2 La composition des peptides synthétiques du CMH-I

Il convient de jeter un coup d'oeil à la composition des peptiques du CMH-I et de la comparer à celle des peptides trypsiques. On compare également la fréquence de chacun des acides aminés parmi ces peptides avec celle dans le protéome. Pour ce faire, on utilise la base de données Swiss-Prot [3]. La figure 3.6 met en évidence un certain nombre de

différences.

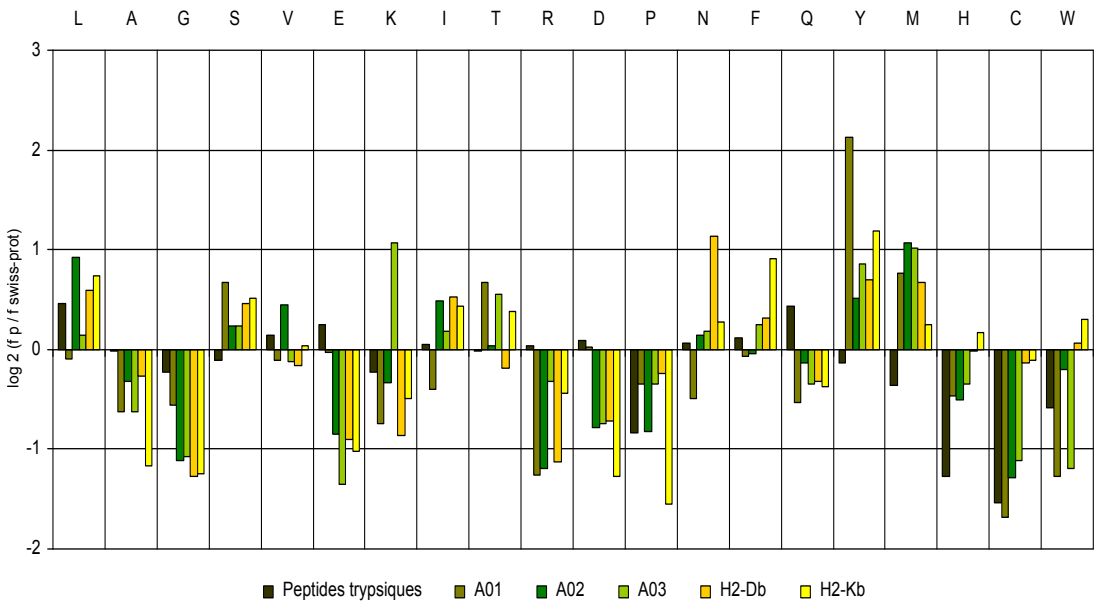


Figure 3.6 – Comparaison de la composition des peptides trypsiques et des peptides du CMH-I pour les allèles HLA-A*01, HLA-A*02, HLA-A*03, H2-Db et H2-Kb avec la fréquence des acides aminés dans la base de données Swiss-Prot.

L'acide aminé Met (M) est plus abondant (avec un rapport supérieur à 2) parmi les peptides HLA-A*01, HLA-A*02, HLA-A*03, H2-Db et dans une moindre proportion dans les peptides H2-Kb que parmi les peptides trypsiques et le protéome. La proportion d'acide aminé Lys (K) parmi les peptides synthétiques du CMH-I est équivalente à celle trouvée parmi les peptides trysiques ainsi que dans la base Swiss-prot. On peut néanmoins noter qu'il y en a plus parmi les peptides HLA-A*03. En fait, l'acide aminé Lys se trouve principalement en position C-terminale des peptides trypsiques d'une longueur inférieur ou égale à 12 résidus, rarement ailleurs dans la séquence. Parmi les peptides du CMH-I, cet acide aminé peut se retrouver n'importe où dans la séquence. De plus, pour les peptides HLA-A*03, cet acide aminé est fréquemment retrouvé en posi-

tion C-terminale en plus de pouvoir être présent ailleurs dans la séquence [42]. On note que l'acide aminé Tyr (Y) est très présent parmi les peptides HLA-A*01 (abondance = 12,41%). Ce constat corrobore le motif établi pour cet allèle qui montre une présence importante de tyrosine en position C-terminal [42]. L'acide aminé Cys est globalement moins présent dans les peptides trypsiques et du CMH-I que dans la base Swiss-prot. Ceci peut notamment être expliqué par le fait que les spectres résultant de la fragmentation de ces peptides sont plus difficilement interprétables ou que cet acide aminé est modifié. On peut supposer par conséquent que la présence de cet acide aminé parmi les peptides du CMH-I identifiés par spectrométrie de masse est sous-évaluée. Il y a également une plus faible proportion de Proline parmi les peptides trypsiques et les peptides du CMH-I. De plus, l'acide aminé Arg est moins fréquent parmi les peptides du CMH-I. Les peptides synthétiques du CMH-I contiennent une plus faible proportion des acides aminés Glu (E) et de Gly (G) surtout pour les peptides HLA-A*02, HLA-A*03, H2-Db, H2-Kb. C'est également la même tendance pour l'acide aminé Asp (D) sauf pour les peptides HLA-A*01 qui d'après la littérature ont souvent ce résidu en position P3 [42].

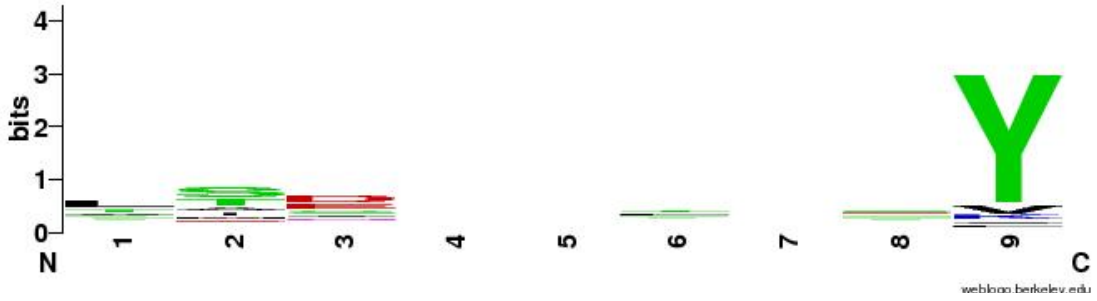


Figure 3.7 – Logo pour les peptides HLA-A*01 de 9 résidus. La position en position C-terminale constitue un résidu d'ancrage alors que la position 3 contient majoritairement les acides aminés Asp (D) ou Glu (E).

La figure 3.7 met en évidence clairement le résidu d'ancrage en position C-terminale des peptides HLA-A*01. En position 3, ils présentent souvent un acide aminé Asp (D)

ou Glu (E). Pour les peptides HLA-A*02, il y a une plus grande proportion d'acides aminés Leu (L), Ile (I) Val (V) en position C-terminale comme le montrent la figure 3.8. En position P2, on note une présence importante d'acides aminés Leu (L) et Met (M).

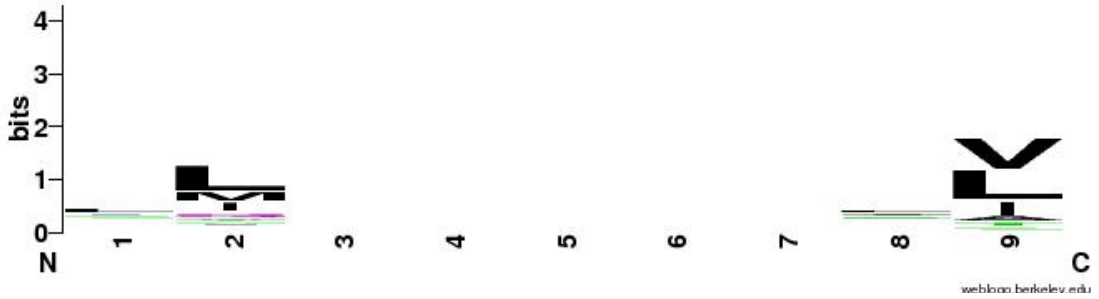


Figure 3.8 – Logo pour les peptides HLA-A02 de 9 résidus. Les positions C-terminale et P2 constituent des résidus d'ancrage.

Les peptides HLA-A*03 sont connus pour se terminer, dans une grande proportion, par les acides aminés basiques Lys (K), Arg (R) ou l'acide aminé aromatique Tyr (Y) (Figure 3.9)[42]. Les acides aminés Lys (K) et Glu (Q) sont isobariques ; leurs masses ne diffèrent que de 0.04 u. La connaissance de l'isoforme HLA peut fournir un indice pour distinguer ces deux acides aminés lorsque l'ambiguïté concerne le résidu en position C-terminal.

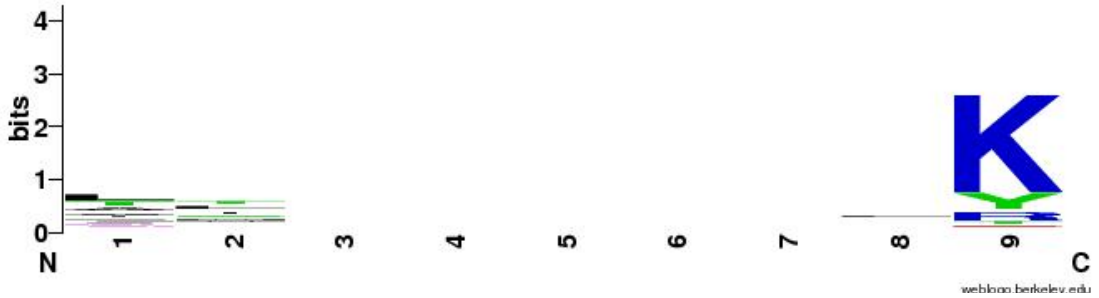


Figure 3.9 – Logo pour les peptides HLA-A*03 de 9 résidus. En position C-terminal, on constate la présence majoritaire de Lys (K), Tyr (Y) et Arg (R).

Les peptides H2-Db ont une position d'ancrage en P5 avec un acide aminé Asp (N)

et se terminent par une méthionine (M), isoleucine (I), leucine (L) comme le montre la figure 3.10. Les peptides H2-Kb ont majoritairement un acide aminé Leu (L), Phe (F) ou Val (V) en position C-terminal.

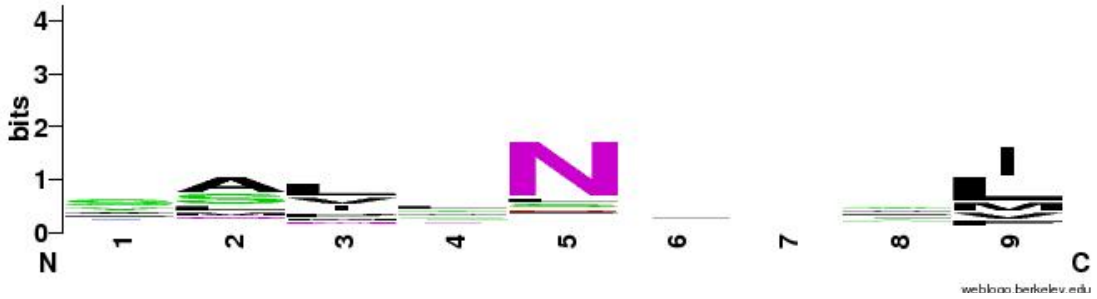


Figure 3.10 – Logo pour les peptides H2-Db de 9 résidus. Les positions P5 et C-terminale constituent des positions d'ancrage.

Figure 3.11 – Logo pour les peptides H2-Kb de 9 résidus. La position C-terminale est une position d'ancrage.

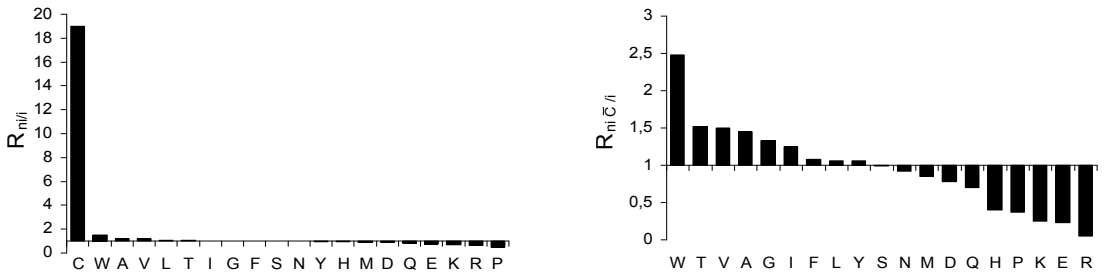
À l'exception des peptides HLA-A*03, les peptides synthétiques du CMH-I se terminent rarement en C-terminale par un acide aminé porteur de charge. On peut donc *a priori* présumer une plus faible proportion de fragments *y* pour ces peptides. Comme il a été précisé plus tôt, un certain nombre de peptides n'ont pas été identifiés par MASCOT. L'analyse de la composition de ces peptides permet de mettre en évidence quelques faits intéressants. Comme le montre la figure 3.12 (a), une large proportion (52%) de ces peptides contiennent un acide aminé Cys (C). Ce résidu peut subir un certain nombre de

modifications qui peuvent compromettre l'identification des peptides si celles-ci ne sont pas prise en compte lors de la recherche. Il y a notamment :

1. la réduction et l'alkylation lors de la préparation des échantillons,
2. la formation de ponts disulfide : les groupes thiol de deux résidus cystéine peuvent se combiner pour former un pseudo-dimère avec une perte de deux protons,
3. le résidu cystéine peut être converti en acide sulfinique ou sulfonique dans des conditions oxydatives,
4. la béta-élimination correspondant à la perte d'un groupe H2S d'un résidu cystéine [31], etc.

Cependant, les analyses faites n'ont pu mettre en évidence avec certitude l'une de ces modifications. D'autres modifications pourraient avoir lieu. Certains spectres pourraient mettre en évidence l'addition de sodium mais des recherches supplémentaires sont nécessaires pour le confirmer. Il se pourrait aussi que cet acide aminé ait une influence négative sur l'ionisation du précurseur. Il a été décidé de ne pas mener davantage d'investigations sur ce sujet étant donnée la faible fréquence de l'acide aminé Cys parmi la population des peptides du CMH-I.

Parmi les peptides non identifiés par MASCOT dans le cas de la fragmentation CID, on constate une fréquence des acides aminés H, P, K, E, R moindre que parmi les peptides identifiés par MASCOT (3.12 (b)). Les résidus H, K et R favorisent la fragmentation des peptides car ils sont porteurs de charge positive. Ces observations peuvent expliquer en partie pourquoi ces peptides ne sont pas identifiés. Pour ce qui est des peptides non observés, nous soupçonnons qu'une des raisons possibles serait la mauvaise synthèse de ces peptides ou encore la formation de produits de réactions secondaires.



(a) (b)

Figure 3.12 – Composition des peptides synthétiques du CMH-I non identifiés par MASCOT. (a) Rapport de l'occurrence de chacun des acides aminés parmi les peptides synthétiques non identifiés par MASCOT par l'occurrence parmi les peptides identifiés ($R_{ni/i}$). Les peptides non identifiés contiennent un proportion beaucoup plus grande de cystéine ($R_{ni/i} \simeq 19$). (b) Rapport de l'occurrence de chacun des acides aminés sauf la cystéine parmi les peptides synthétiques non identifiés par MASCOT ne contenant pas de cystéine par l'occurrence parmi les peptides identifiés ($R_{ni\bar{C}/i}$). Ces derniers contiennent une plus faible proportion d'arginine, de lysine, d'histidine et de proline $R_{ni\bar{C}/i} < 0,5$.

3.3 L'analyse comparée de la fragmentation des peptides trysiques et des peptides synthétiques du CMH-I

3.3.1 La complexité des spectres

On s'est intéressé à la corrélation entre la longueur du peptide et le nombre de pics dans le spectre MS/MS. Le coefficient de Spearman est égal à 0.6075 pour les peptides trypsiques. Bien que la figure 3.13 montre que le nombre de pics ait une légère tendance à augmenter avec la longueur du peptide, il est impossible de prévoir à partir de la seule longueur du peptide le nombre de pics, et ce même de façon très approximative. Plusieurs facteurs influent sur la complexité du spectre. Le spectromètre utilisé, le mode de fragmentation (CID ou HCD) et la composition du peptide sont parmi ceux-ci. La fragmentation en mode HCD produit une plus grande variété de fragments qu'en mode CID. En effet, le mode HCD conduit à davantage de fragmentations successives des ions et donc à la présence d'un plus grand nombre de fragments dans la partie des basses masses du spectre. Un nombre plus important de fragments apporte plus d'information

mais complexifie le spectre et l'analyse en aval, ce qui *in fine* peut compliquer la tâche des algorithmes de séquençage *de novo* s'ils ne sont pas adaptés. On verra plus loin que certains algorithmes conçus pour les spectres HCD utilisent une étape de simplification du spectre.

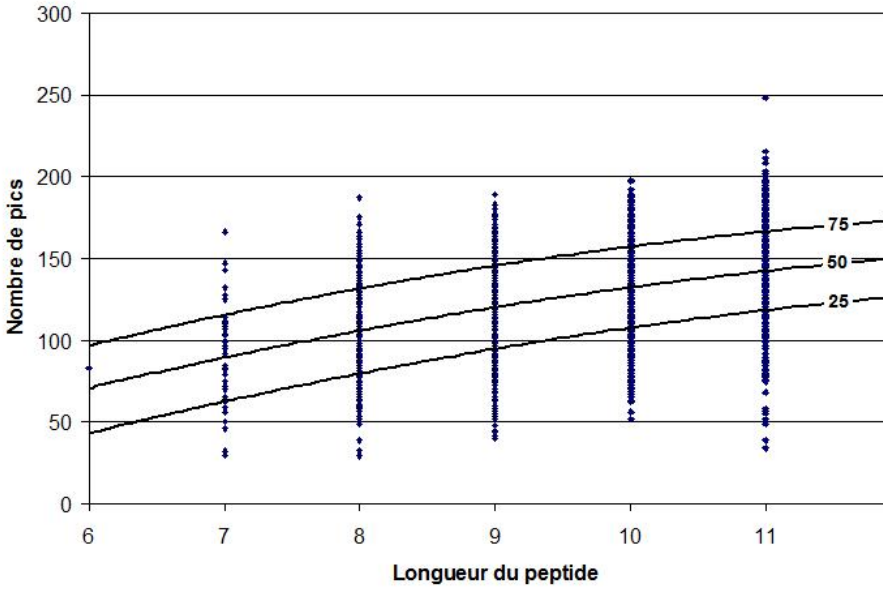


Figure 3.13 – Complexité des spectres par rapport à la longueur des peptides

On obtient la même tendance pour les peptides synthétiques du CMH-I. Néanmoins, on peut constater que le nombre de pics observables pour les peptides synthétiques du CMH-I est inférieur au nombre de ceux observables pour les peptides trypsiques. La complexité est donc moindre, mais l'information l'est nettement aussi.

Tableau 3.III – Nombre moyen de pics par rapport à la longueur des peptides en fragmentation CID

Longueur	Peptides trypsiques	Peptides CMH-I
8	135,76	58,38
9	141,29	60,6
10	157,28	69,86

3.3.1.1 Les ions fragments observés

La composition des peptides a une grande influence sur la nature et la quantité des ions fragments observés. On constate des différences notables entre les peptides trypsiques et les peptides synthétiques du CMH-I (Annexe II). En fragmentation CID, la proportion des ions fragments pour les peptides synthétiques du CMH-I est inférieure, pour chacun des types *y*, *b*, et *a*, à celle des peptides trypsiques. Plus de 80% des fragments *y* sont observés pour les peptides trypsiques. Cela signifie que pour un spectre associé à un peptide trypsique d'une longueur de 10 résidus, on observe en moyenne $80\% \times (10-1) = 7,2$ fragments *y*. Pour les peptides synthétiques du CMH-I, seuls 58% des fragments *y* sont observés. Pour les fragments *b*, la proportion est sensiblement la même (57,2% pour les peptides trypsiques contre 52,2% pour les peptides synthétiques du CMH-I). Ceci s'explique par le fait que les peptides synthétiques du CMH-I contiennent une charge en position N-terminal comme les peptides trypsiques. Par conséquent, la proportion d'ions *b* est moins affectée. Les spectres de peptides trypsiques contiennent aussi un plus grand nombre de fragments *a*. L'annexe II rapporte les résultats pour un plus grand nombre de types de fragments et renseigne sur l'intervalle de confiance bootstrap. On note aussi que la proportion de fragments internes est a peu près le double pour les peptides trypsiques.

La figure 3.15 montre la proportion des fragments pour des niveaux d'intensité relative calculée par rapport à l'intensité du fragment le plus abondant. On peut voir que les spectres des peptides synthétiques du CMH-I contiennent davantage de fragments *y* ayant des intensités relatives inférieures à 10%. Globalement, les fragments *y* dans les spectres des peptides trypsiques sont plus abondants. Pour les fragments *b*, il n'y a pas des différences significatives à l'exception de la proportion des fragments *b* ayant une

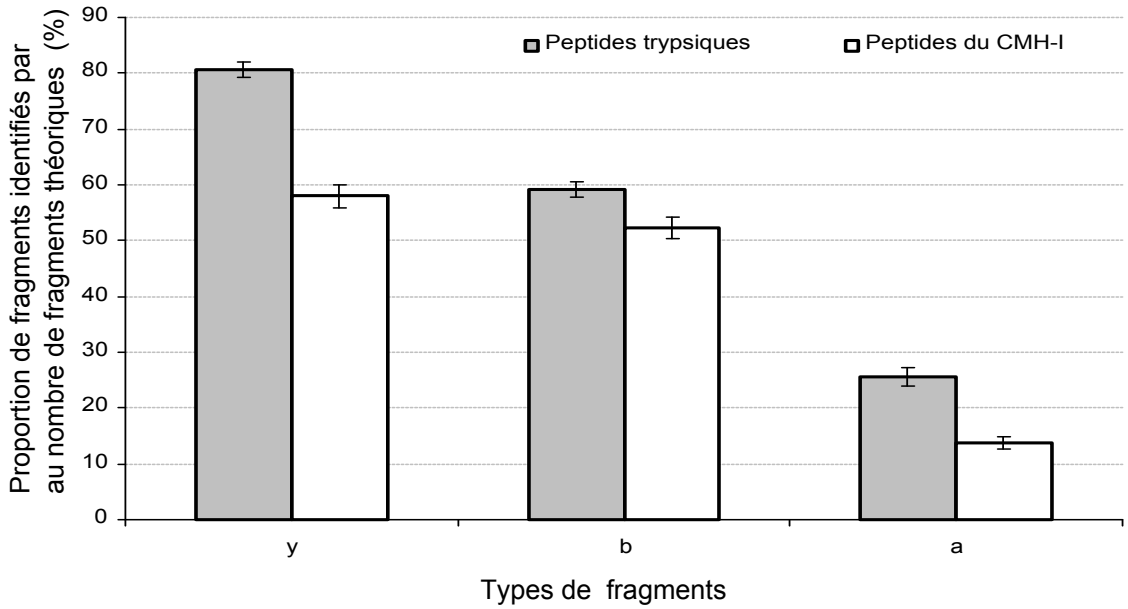


Figure 3.14 – Comparaison de la proportion des fragments *y*, *b*, et *a* observés par rapport au nombre de fragments théoriques. Les spectres de peptides trypsiques contiennent une plus grande proportion de fragments *y*, *b*, et *a* que ceux des peptides du CMH-I.

intensité relative comprise entre 90% et 100%. En fait, il arrive plus fréquemment dans les spectres des peptides synthétiques du CMH-I qu'un fragment *b* soit le plus abondant. Ce résultat est explicable par le fait que les peptides synthétiques du CMH-I sont principalement chargés du côté N-terminal. L'intensité des fragments *a* est globalement inférieure à 10%. Il n'est pas impossible qu'un proportion des fragments *a* identifiés comme tel correspond à du bruit.

Le profil de fragmentation des peptides synthétiques du CMH-I est également différent de celui des peptides trypsiques. Comme le montre la figure 3.16, les ions *y* ont des intensités relatives supérieures dans les spectres de peptides trypsiques (de 30% à 60% supérieurs). En revanche, pour les ions *b*, on observe, dans une moindre mesure, une tendance inverse. Pour les spectres de peptides du CMH-I, l'intensité du pic augmente avec la longueur du fragment *b*.

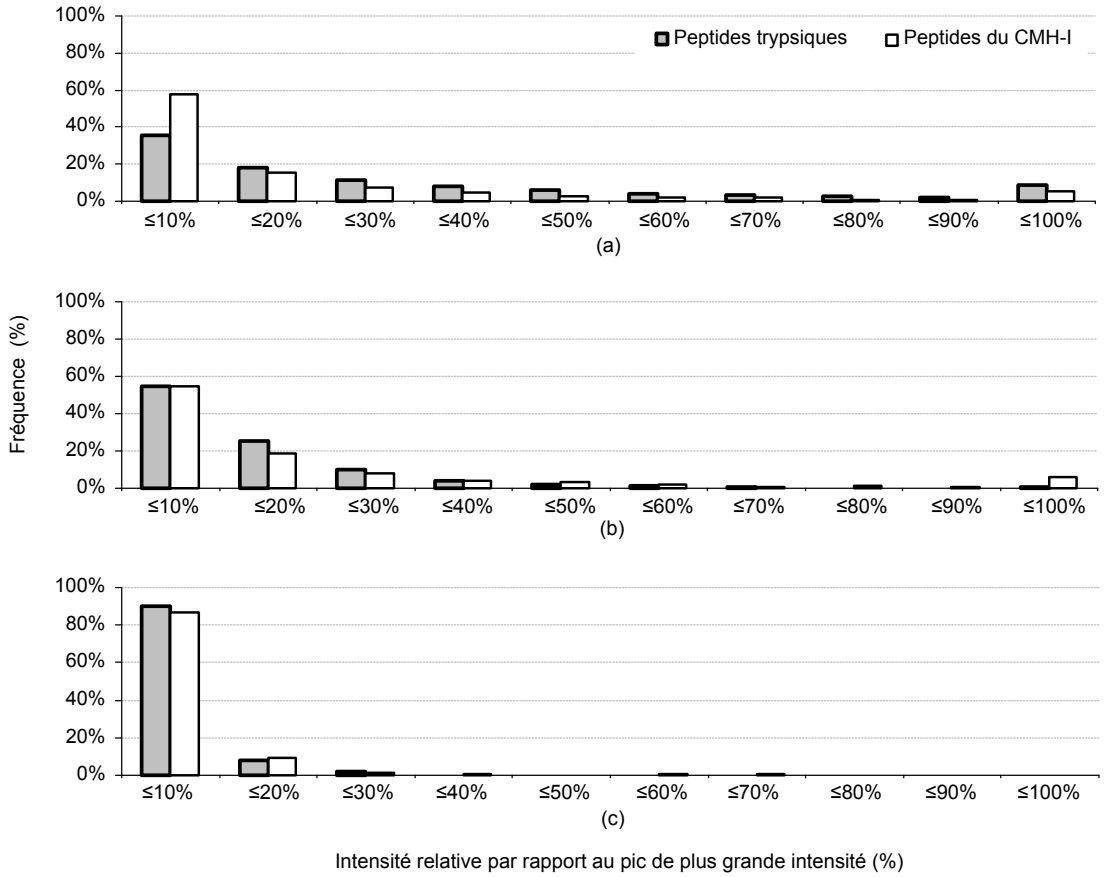


Figure 3.15 – Comparaison de la distribution de l'intensité des fragments *y*, *b*, et *a* observés entre les spectres CID pour les peptides trypsiques et les peptides synthétiques du CMH-I. (a) Les fragments y. (b) Les fragments *b* (c) Les fragments a.

3.3.1.2 L'incomplétude des spectres

La comparaison de l'incomplétude entre les spectres CID des peptides du CMH-I et des peptides trypsiques montre clairement que les derniers conduisent dans une grande proportion à des spectres complets (68%) ou presque complets (29%). En comparaison, seuls 7,7% des spectres CID des peptides du CMH-I sont complets et 45,5% sont presque complets. Cela signifie que pour chacun des 46,8% des spectres des peptides du CMH-I, il y a un grand nombre de séquences peptidiques correspondantes possibles.

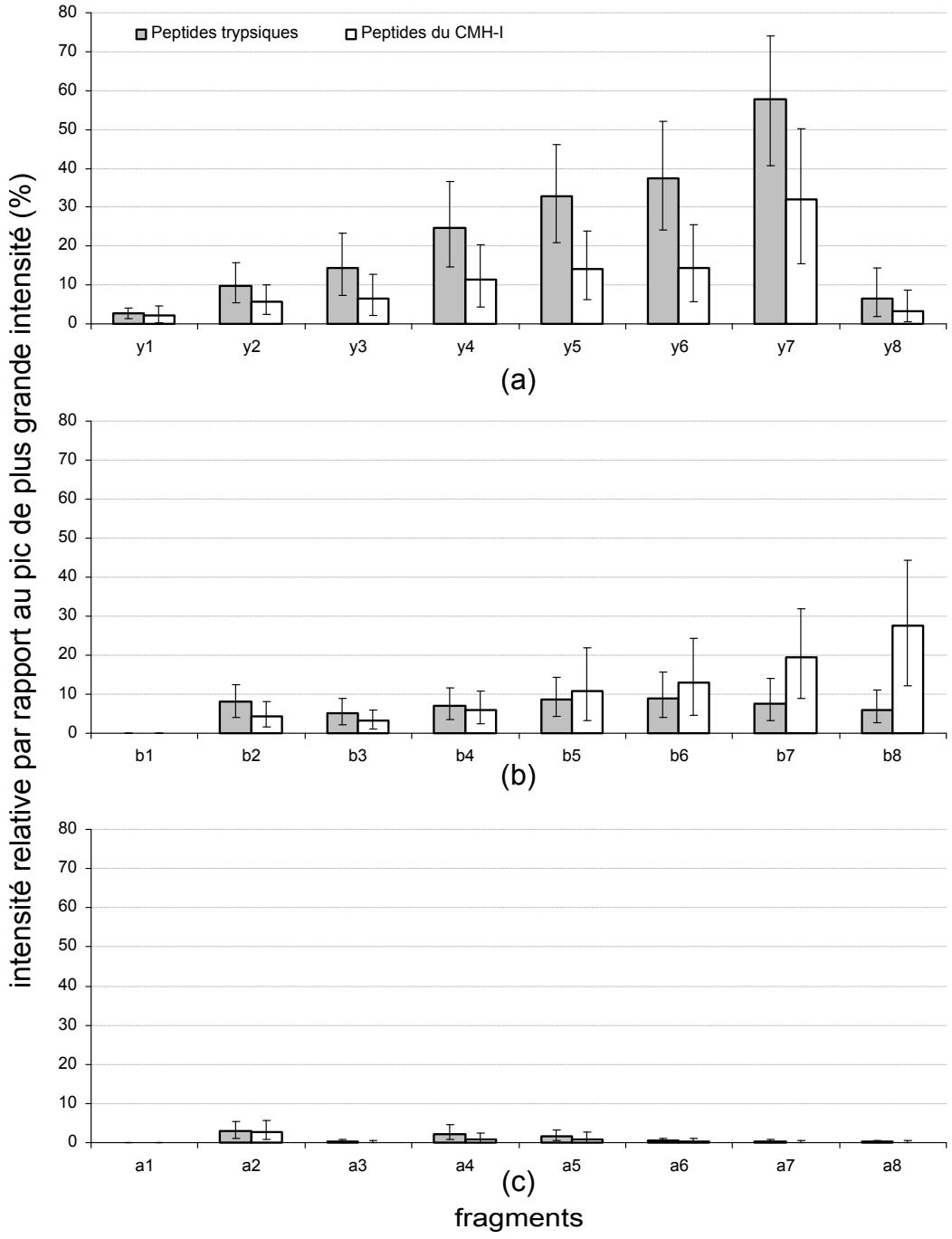


Figure 3.16 – Distribution des ions des fragments des peptides trypsiques et des peptides synthétiques du CMH-I de 9 résidues. (a) Les fragments *y*. (b) Les fragments *b* (c) Les fragments *a*.

Le nombre de combinaisons d'acides aminés possibles augmente de façon exponentielle avec le m/z séparant deux ions fragments. Il est quasiment impossible d'identifier avec une confiance raisonnable la véritable séquence peptidique associée à un spectre ayant une incomplétude supérieur à 0,3 avec une algorithme de séquençage *de novo*.

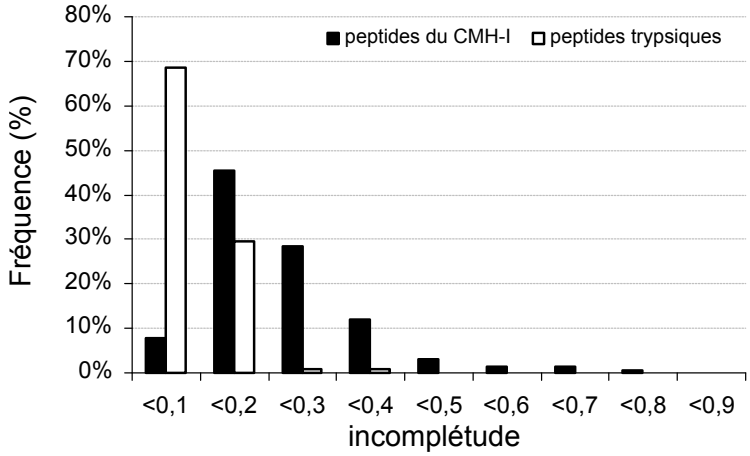


Figure 3.17 – La comparaison de l'incomplétude des spectres CID des peptides synthétiques du CMH-I et des peptides trypsiques

Toutes ces observations expliquent pourquoi l'identification des peptides du CMH-I par MASCOT est plus difficile que pour celle des peptides trypsiques. Les spectres plus pauvres en fragments rendent le séquençage *de novo* compliqué voire impossible dans certains cas tant le nombre de combinaisons de séquences peptidiques possibles est grand. De plus, la plupart des programmes de séquençage *de novo* ont été développé pour les peptides trypsiques qui montrent des profils de fragmentation différents de ceux des peptides du CMH-I. Cependant, des études ont montré que la fragmentation en mode HCD conduit à un plus grand nombre de fragments [33] [5] [12]. Dans la prochaine section, nous montrons les bénéfices qu'apporte ce mode de fragmentation pour les peptides du CMH-I.

3.4 La comparaison des fragmentations CID et HCD des peptides du CMH-I

Nous nous intéressons ici à comparer la fragmentation des peptides synthétiques du CMH-I par deux modes de fragmentation différents : CID et HCD. Le deuxième mode que propose le spectromètre LTQ Oribtrap Velos avec des performances inégalées d'après les concepteurs nous est apparu très prometteur pour l'identification des peptides synthétiques du CMH-I, et notamment leur séquençage *de novo*.

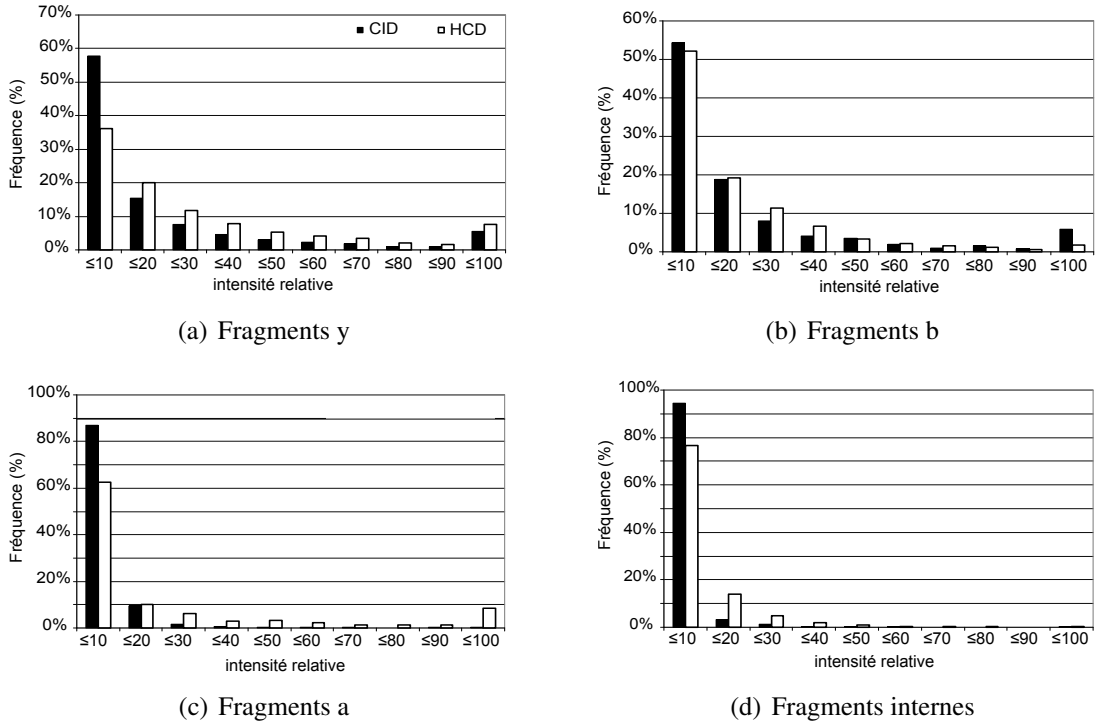
3.4.1 L'intensité des fragments

La comparaison de l'intensité des seuls fragments les plus couramment observés *y*, *b* et *a* entre la fragmentation CID et HCD permet de mettre en évidence que cette dernière conduit à des fragments plus abondants. Près de 45% des fragments dans le mode CID ont une intensité relative inférieure à 5% alors qu'en mode HCD, la proportion est de moins de 30%. Pour les fragments d'intensité comprise entre 5% et 10%, on peut observer la même tendance dans une moindre proportion. Pour les intensités supérieures, on observe l'inverse. On constate qu'il y a une proportion non négligeable des ions *a* dans les spectres HCD qui ont une intensité relative comprise entre 95% et 100%. On observe une plus large proportion de fragments internes en mode HCD ayant des intensités intermédiaires (entre 20% et 50%).

3.4.2 Les profils de fragmentation

L'analyse des profils de fragmentation, telle que montrée par la figure 3.19 permet de mettre en évidence plusieurs caractéristiques qui distinguent les deux modes de fragmentation CID et HCD :

1. les spectres HCD montrent une abondance en fragments y_1 et y_2 nettement plus



(a) Fragments y

(b) Fragments b

(c) Fragments a

(d) Fragments internes

Figure 3.18 – Comparaison des distributions de l'intensité des ions fragments *y*, *b* et *a* identifiés pour la fragmentation CID et HCD. (a) Dans le mode de fragmentation HCD, les fragments *y* se situent dans des intensités relatives par rapport à l'intensité maximale supérieures que dans le mode CID. (b) Les fragments *b* ont des intensités légèrement plus élevés dans le mode HCD. (c) Dans le mode CID, une plus grande proportion de fragments *a* ont des intensités inférieures à 10% alors qu'environ 9% des fragments *a* dans le mode HCD ont des intensités entre 90 et 100%. (d) La plupart des fragments internes en mode CID a de faible intensité. Une plus large proportion de fragments internes en mode HCD ont des intensités intermédiaires.

importante que dans les spectres CID,

2. les spectres CID montrent une augmentation d'intensité avec la longueur du fragment alors que ce n'est pas le cas avec les spectres HCD,

3. les spectres HCD montrent une abondance importante en fragments a_2 et b_2 alors que les spectres CID en contiennent peu.

L'ensemble des intensités relatives pour les peptides HLA-A*01 de 9 résidus se trouve en annexe III. Les mêmes observations peuvent être faites avec les peptides HLA-

A*02, HLA-A*02, H2-Db et H2-Kb. Les fragments abondants de faible masse tels que y_1, y_2, a_2 et b_2 permettent de fournir des informations supplémentaires. Le mode normal des trappes ioniques, CID, présentent moins de fragments de faible masse. L'intensité des fragments *b1* est très faible dans les deux modes de fragmentation. Ceci est probablement dû à l'instabilité relative des ions b_1 [30]. On peut noter qu'il y a presque 10% des fragments *a* qui ont une intensité comprise entre 90% et 100%. Cela signifie qu'il y a des spectres HCD pour lesquels il y a au moins un fragment *a* qui est proche de l'intensité maximale. Ceci n'est pas du tout le cas avec les spectres CID. Les spectres des peptides ITEMLQKEY (Figure 3.20) et LSNFGAPSY (Figure 3.21) illustrent les observations faites sur l'ensemble des données.

3.4.3 L'incomplétude des spectres

La comparaison de l'incomplétude met évidence que la fragmentation HCD permet d'obtenir une plus grande proportion de spectres fournissant toutes les informations nécessaires au séquençage. En effet, 49,9% des spectres HCD des peptides synthétiques du CMH-I contient toutes les informations sur la fragmentation (Figure 3.22). Ce qui signifie que tous les clivages le long de la chaîne peptidique sont représentés (par des ions *b* ou/et *y*). Alors qu'il n'y en a que 7,7% avec la fragmentation CID (Figure 3.22). Pour 45,5% des spectres CID il manque un clivage. Les spectres HCD avec seulement un clivage manquant comptabilisent un total de 24.3%. Ensemble, les spectres HCD contenant presque toute l'information sur le clivage comptabilisent 74,3% des spectres. Ce total n'est que de 53,2% pour les spectres CID. Ces résultats montrent clairement le gain d'information obtenu avec la fragmentation HCD.

Les figures 3.20(a) 3.20 (b) montrent respectivement le spectre en fragmentation CID et le spectre en fragmentation HCD du même peptide ITEMLQKEY. On peut

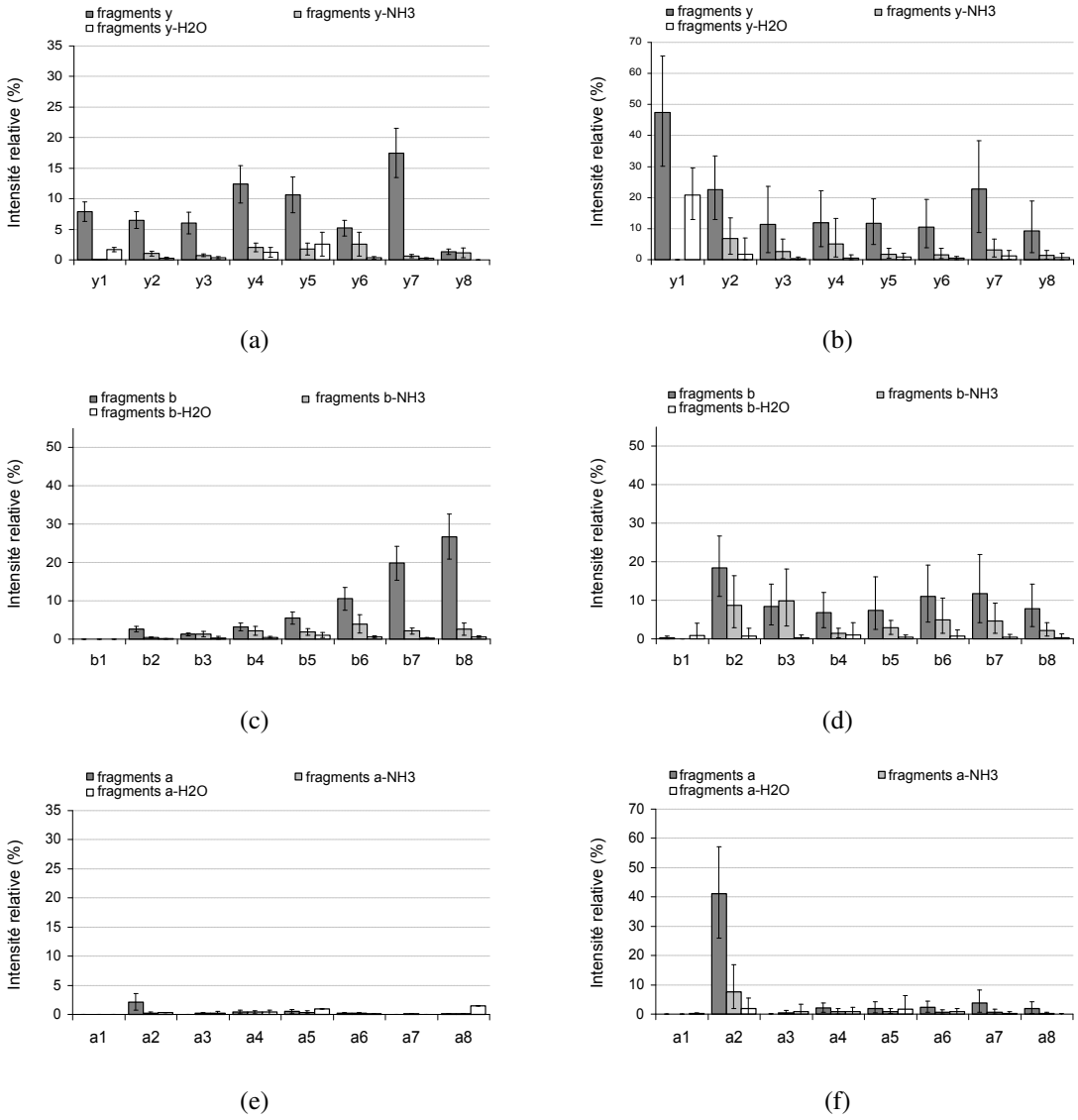
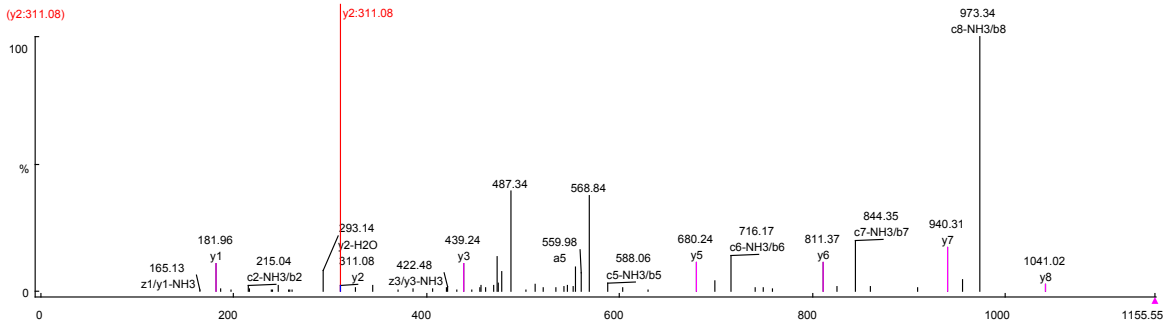


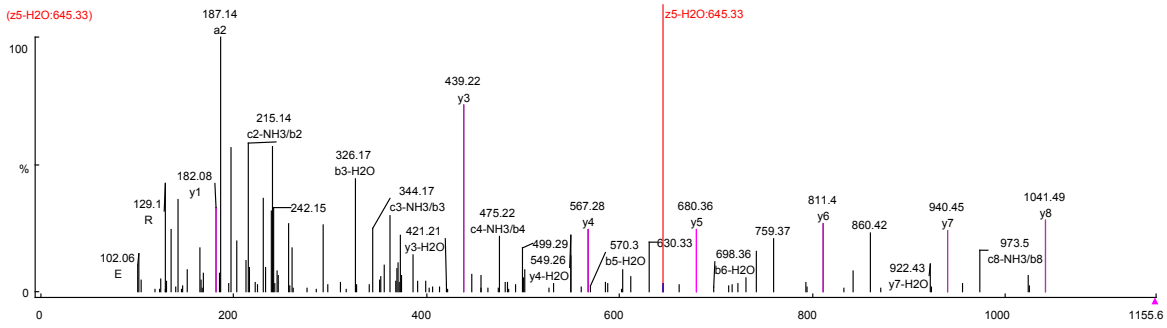
Figure 3.19 – Profil des ions fragments *y* **pour les peptides HLA-A01 dans les deux modes de fragmentation** CID (a) et HCD (b)

constater que le spectre HCD contient un plus grand nombre de pics et fournit plus d'informations.

Pour illustrer l'importance de la complétude pour l'interprétation d'un spectre prenons l'exemple du peptide qui a pour séquence MLNRVQILM et pour lequel le spectre HCD a une incomplétude de 0,56. Les seuls fragments *y* observés sont : y_1, y_2, y_7 et y_8



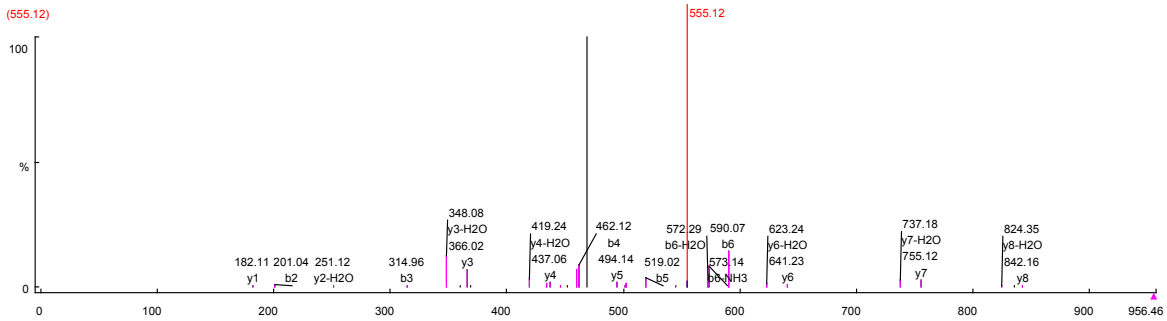
(a) CID



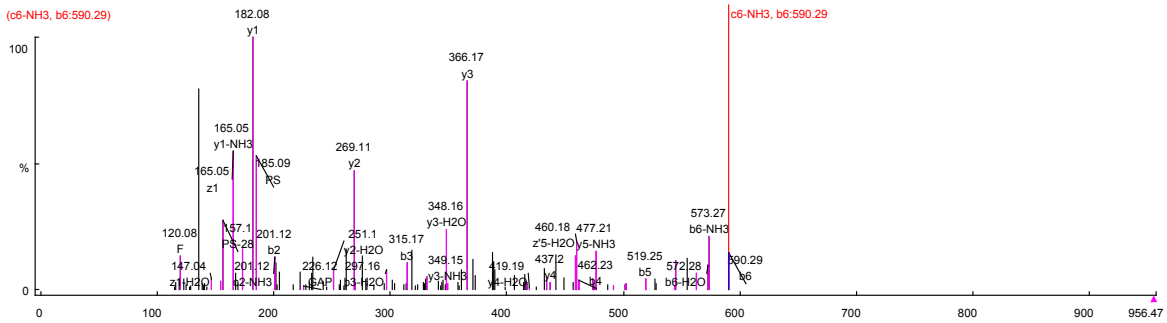
(b) HCD

Figure 3.20 – Comparaison des spectres CID et HCD pour le peptide ITEMLQKEY. (a) spectre CID du peptide ITEMLQKEY doublement chargé. Le spectre contient une soixantaine de pics. L'intensité des ions fragments *b* croit à mesure que la taille des fragments augmente. Dans une moindre mesure, l'intensité des ions fragments *y* croit à mesure que la taille des fragments augmente à l'exception de l'ion fragment y_8. On n'observe pratiquement pas d'ions *a*. (b) spectre HCD du peptide ITEMLQKEY doublement chargé. Le spectre HCD contient le double de pics que le spectre CID. L'intensité des ions fragments (b_2, b_3, b_4) décroit à mesure que la taille des fragments augmente. Les ions fragments *y* ont une intensité supérieure que dans la fragmentation CID. Les intensités des ions *y* sont homogènes. On note que l'ion fragment a_2 à une intensité relative de 100%.

alors qu'on observe aucun fragment *b*. Ne tenons pas compte ni des fragments internes ni des ions immoniums. La différence de masse entre les fragments y_2 et y_7 est de 610,356. Le nombre de séquences possibles correspondant à cette masse est 32002 avec une tolérance sur la masse des fragments de +/- 0,02 Da. Autant dire que la détermination de la séquence est impossible. La présence de l'ion immonium de l'acide aminé Q (sachant que cet acide aminé n'est pas dans le reste de la séquence) permet de réduire ce nombre à 6955. Nous verrons plus tard l'importance des ions immoniums.



(a) CID



(b) HCD

Figure 3.21 – Comparaison des spectres CID et HCD pour le peptide LSNFGAPSY. (a) spectre CID du peptide LSNFGAPSY. L'intensité de l'ion y_1 est faible. Il y a peu de pics dans les basses masses du spectre. (b) spectre HCD du peptide LSNFGAPSY. Le spectre HCD contient beaucoup plus de pics que le spectre CID. L'intensité des ions y_1 et b_2 est nettement supérieure que dans le spectre CID. Il contient plus de fragments dans les basses masses du spectre.

La corrélation entre la complétude en CID et HCD est faible (La corrélation de Spearman donne $r = 0,21$, et elle est comprise entre 0,11 et 0,30 pour intervalle de confiance de 95%). Pour 22% des peptides synthétiques, les spectres HCD sont moins complets que les spectres CID. Cependant pour la majorité (68%) les spectres HCD sont plus complets. La comparaison de la composition de ces peptides ne montrent pas de différences significatives à l'exception d'une plus grande proportion d'acides aminés Trp (rapport 3,51) et Met (2,79) parmi les peptides produisant des spectres plus informatifs en mode CID et une plus faible proportion de His (rapport 0,43).

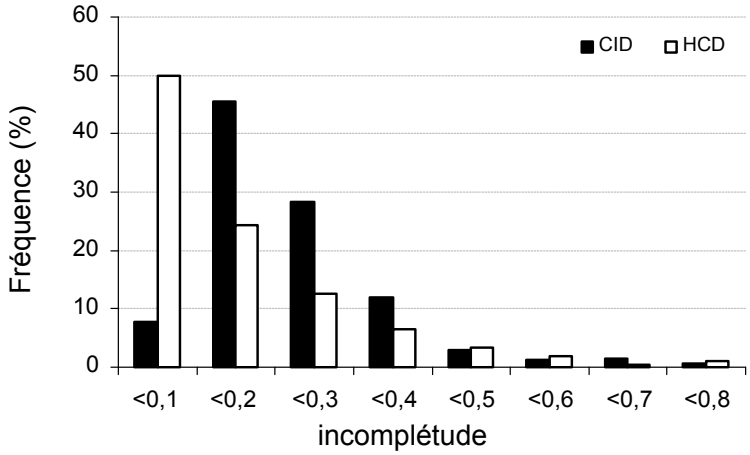


Figure 3.22 – La comparaison de l'incomplétude des spectres CID et HCD pour les peptides synthétiques du CMH-I. La proportion de spectres ayant une incomplétude inférieure à 0,1 (pour la majorité égale à 0) est plus importante avec la fragmentation en mode HCD qu'en mode CID. Globalement, l'incomplétude est plus importante pour les spectres CID.

3.4.4 Les ions immoniums

L'absence d'un ion fragment *y* ou *b* peut conduire à une ambiguïté non seulement sur l'orientation des deux résidus adjacents au site de clivage manquant mais aussi sur l'identité de ces deux derniers. Pour une masse donnée, on peut avoir plusieurs combinaisons possibles d'acides aminés. Par exemple, une différence de masse de 229 peut correspondre à plusieurs combinaisons d'acides aminés : N-D, Q-T, K-T, A-A-S, A-G-T, D-G-G. Nous avons déterminé tous les compositions d'acides aminés possibles pour les masses comprises entre 114 et 250 u (Annexe IV). Les ions immoniums peuvent être très informatifs sur la présence d'un acide aminé dans la séquence peptidique. Leur présence a donc un intérêt évident pour le séquençage *de novo*. Ils apparaissent dans un spectre MS/MS à des masses inférieurs à 200 Da. Le tableau 3.IV fournit la fréquence des ions immoniums F_l pour les spectres CID et HCD. Les spectres CID contiennent rarement d'ions immonium ; pour ainsi presque pas. En revanche, les spectres HCD en

contiennent fréquemment. Pour donner une idée intuitive de ce à quoi correspond la valeur de F_l, on peut dire qu'un spectre HCD pour un peptide de 10 résidus contient en moyenne un peu de moins de 2 ions immoniums ($\simeq 17{,}75\% \times 10$). La figure 3.23 montre pour chacun des acides aminés, la fréquence des ions immoniums. L'absence des ions immoniums en-dessous de 100 u est expliquée par la gamme de masse du spectromètre de masse (100 à 2000 u). Sous 100 u il y aurait une perte de sensibilité trop importante due à la lecture à la fois en haute masse et en basse masse. De plus, il est connu que tous les acides aminés non pas la même propension à produire des ions immoniums. Par conséquent, alors que la présence d'ions immoniums est informative, l'absence ne l'est nullement. Au vu des observations, les ions immoniums de F, H et W ont une probabilité d'être observés supérieure à 85%. Pour F et Q, elle est respectivement supérieure à 70% et 60%. Elle est de plus ou moins 50% pour R, E et Y.

Tableau 3.IV – Proportion d'ions immoniums dans les spectres CID et HCD. Le ratio F_l correspond au nombre d'occurences de ions immoniums divisé par la longueur du peptide. IC_{95} est l'intervalle de confiance à 95%.

Mode de fragmentation	Ratio F_l	IC_{95}
CID	0,55%	+/-8%
HCD	17,75%	+/- 12%

3.4.5 Les fragments internes

Les fragments internes résultent du clivage de deux liaisons de la chaîne peptique. Ceux-ci ont perdu les côtés N et C terminaux. Comme ils correspondent à des séquences de juste quelques résidus, ils se situent principalement dans la première partie du spectre. Ces pics peuvent aider à confirmer la séquence ou une sous-séquence du peptide surtout

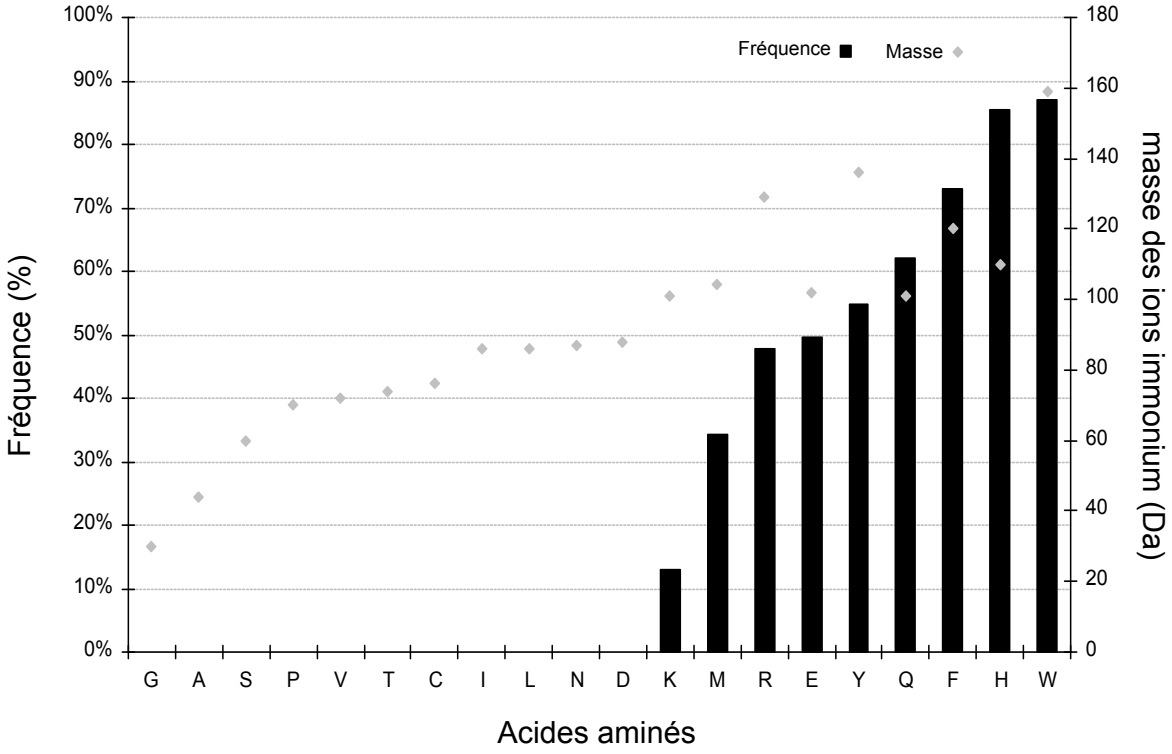


Figure 3.23 – Proportion d'ions immoniums pour chacun des acides aminés. Les losanges représentent la masse de chacun des ions immoniums. Aucun des ions immoniums ayant une masse inférieur à 100 Da n'est détecté.

lorsqu'il manque des ions *b* et *y*. La présence d'un acide aminé Pro favorise les fragments internes. De plus, ceux-ci sont environ 4 fois plus fréquents dans les spectres HCD que dans les spectres CID comme le montre le tableau 3.V. La différence est significative. Cette observation et celle faite sur la distribution d'intensité des fragments internes (figure 3.18) nous montrent que le fragmentation en mode HCD conduit à un nombre de fragments internes significativement plus importants et d'abondance plus importante.

Tableau 3.V – Proportion de fragments internes dans les spectres CID et HCD. Le ratio $\bar{f}_{interne}$ correspond au nombre d'occurences de fragments internes divisé par la longueur du peptide moins 1.

Mode de fragmentation	$\bar{f}_{interne}$	IC_{95}
CID	28%	+/-7,23%
HCD	117%	+/-5,14%

3.4.6 La précision de masse

La précision de masse est la mesure de la proximité entre la valeur observée et la vraie valeur. L'erreur de masse implique une composante aléatoire ou stochastique et une autre composante systématique ou biais. Cette dernière peut être évaluée et corrigée, notamment par ce que l'on appelle la calibration. La précision en masse relève d'une grande importante pour l'interprétation des spectres et peut être déterminante pour le séquençage *de novo*. Comme déjà mentionné, plusieurs combinaisons d'acides aminés peuvent correspondre à une même différence de masse (Annexe IV). Le concept de peptides homéométriques, introduit par le Dr Pevzner et son équipe, permet de mettre en évidence l'apport évident de la haute précision en masse [5]. Ces derniers sont des peptides de séquences différentes mais de spectres similaires. Ils démontrent que plus la précision en masse est bonne, moins de séquences peuvent correspondre à un spectre donné.

La figure 3.24 représente la distribution de l'erreur de masse sur les fragments *y*, $y - H_2O$ et $y - NH_3$ en mode HCD. Celle-ci se situe en-dessous de 20ppm soit 0,02 Da pour un peptide d'une masse de 1000 Da. Les résultats sont semblables pour les autres types de fragments. Par comparaison, la figure 3.25 montre que l'erreur de masse en mode CID est bien supérieure. Moins de 10% des fragments observés coïncidant avec les fragments théoriques se situent dans un intervalle inférieur à 0,02 Da pour les spectres CID. La proportion est de plus de 80% pour les spectres HCD.

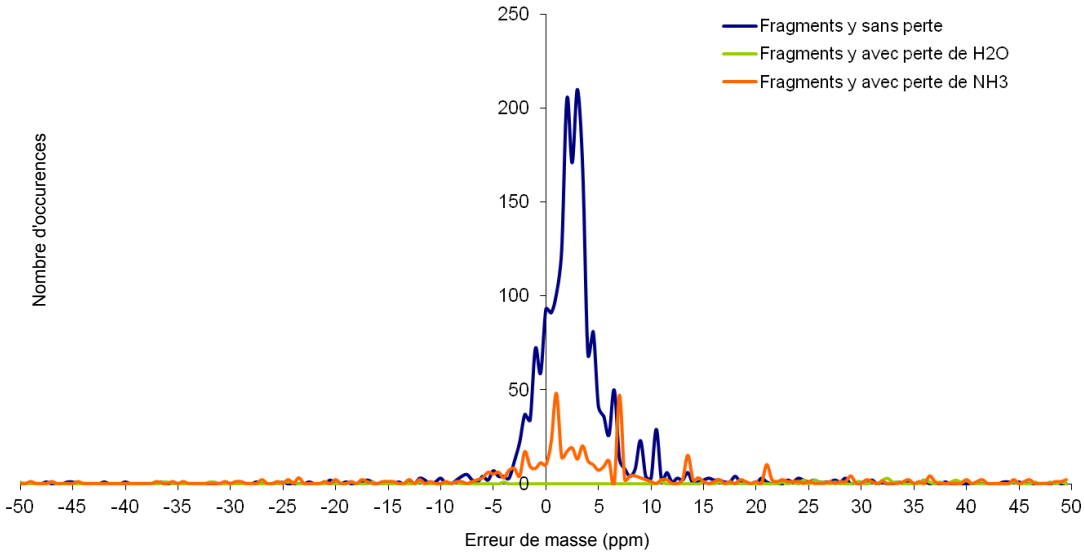


Figure 3.24 – Erreur de masse sur les fragments y**,** $y - H_2O$ **et** $y_N H3$ **en mode HCD.** L'erreur de masse est en-dessous de 20 ppm.

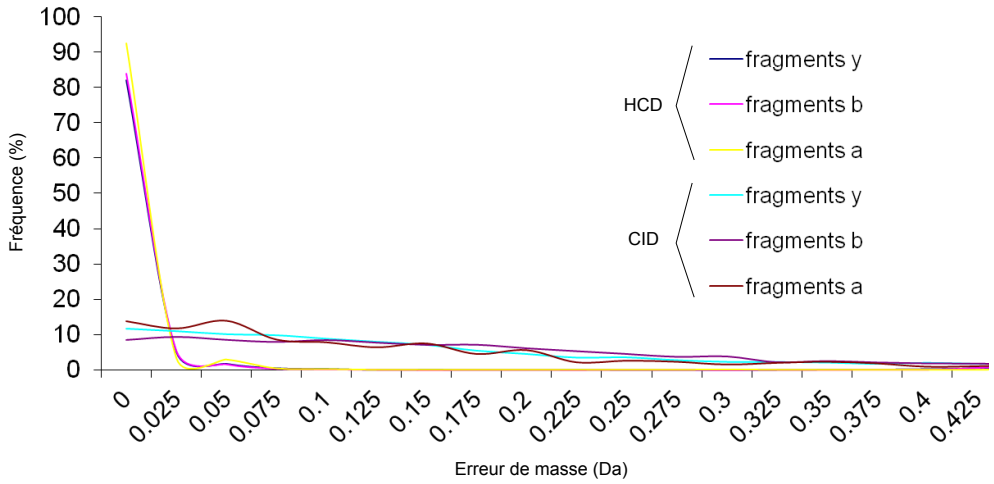


Figure 3.25 – Comparaison de l'erreur de masse entre le mode HCD et CID pour les fragments y**,** $y - H_2O$ **et** $y_N H3$

3.5 L'influence de la composition des peptides sur la fragmentation

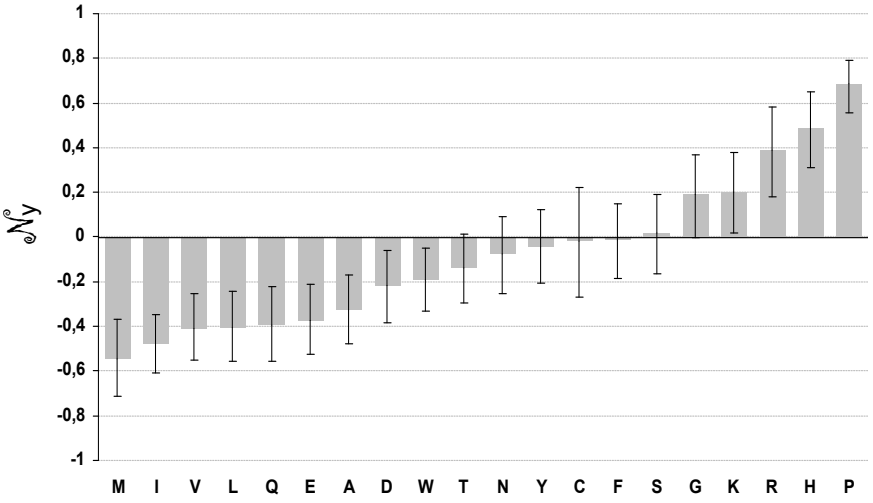
Plusieurs études se sont intéressées à déterminer l'influence de chacun des acides aminés sur la fragmentation des peptides trypsiques. Ici, on fait le même type d'analyse sur les peptides du CMH-I. On montre l'influence de chacun des acides aminés adjacents au site de clivage et distants de celui-ci.

3.5.1 Le N-biais

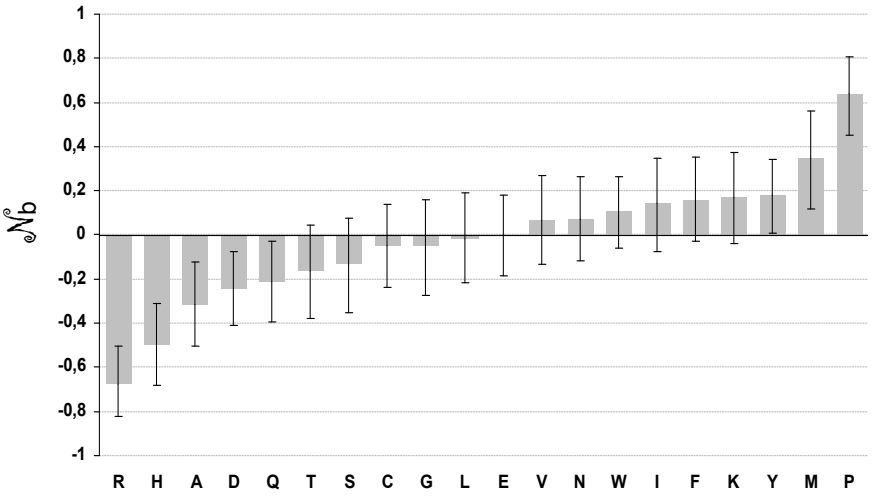
On s'intéresse ici au N-biais uniquement pour les spectres HCD. On met en évidence que la protonation et la rupture de la liaison amide située en N-terminal de l'acide aminé Pro est favorable. En effet, la valeur N-biais pour cet acide aminé approche 0,7 pour les ions *y* et 0,6 pour les ions *b* comme le montre la figure 3.26. Ces ions, résultant de la fragmentation du coté N-terminal, domine en général le spectre alors que ceux issus de la rupture du côté C-terminale sont peu observés. La structure de la proline fait obstacle au clivage de la liaison du côté C-terminale en empêchant l'attaque de la fonction carbonyle en N-terminal [41]. D'autres acides aminés ont un N-biais opposé tels que les acides aminés Met (M) , Ile (I), Val (V), Leu (L) et Gln (Q) et Glu (N). Les acides aminés Arg (R) et His (H) montrent respectivement un N-biais de 0,38 et 0,49 pour les ions *y*. Alors que pour les ions *b*, les valeurs de N-biais sont opposés. L'acide Met (M) montre un N-biais supérieur à 0,3 pour les ions *b*. Pour les autres résidus, l'intervalle de confiance relativement important ne permet pas de faire ressortir un effet évident.

3.5.2 L'influence des résidus adjacents au site de clivage

Les cartes de couples de résidus permettent de faire plusieurs observations. On peut voir sur la figure 3.27 que pour les couples X-P où $X \in A, N, D, E, Q, H, I, L, K, M, F, S, W, Y, V$



(a)



(b)

Figure 3.26 – Le N-biais pour les peptides du CMH-I en fragmentation HCD. (a) N-biais pour les ions y correspondant à chacun des acides aminés. (b) N-biais pour les ions b correspondant à chacun des acides aminés. L'intervalle de confiance bootstrap à 95% montré par les lignes noires.

(colonne P), on a principalement des carrés foncés. L'acide aminé Pro est favorable à la fragmentation du côté N-terminal dans les deux modes de fragmentation CID et HCD. Ceci confirme ce que le calcul du N-biais mettait en évidence.

En revanche, les couples C-P et R-P conduisent à des fragments de plus faible abon-

dance. Les lignes correspondantes à I, L et V sont globalement plus foncées que les autres. Les couples I-X, L-X et V-X sont en effet favorables à la fragmentation. Les couples A-X montrent dans une moindre mesure une propension à produire des fragments abondants. Le clivage en C-terminale de la glutamine (Q-X) est relativement important. Les couples N-X impliquant l'autre acide aminé avec un groupe amide ne produit pas de fragments aussi abondants. Les couples G-X, R-X, H-X (sauf H-P) et P-X produisent des fragments *y* de faible abondance surtout en CID. La figure 3.28 montre que les couples I-X, L-X et V-X produisent des fragments *b* relativement abondants en CID. La figure 3.29 met en évidence la différence significative entre la fragmentation CID et HCD pour les fragments *a*. Globalement, les couples I-X, L-X et V-X produisent les ions fragments les plus abondants.

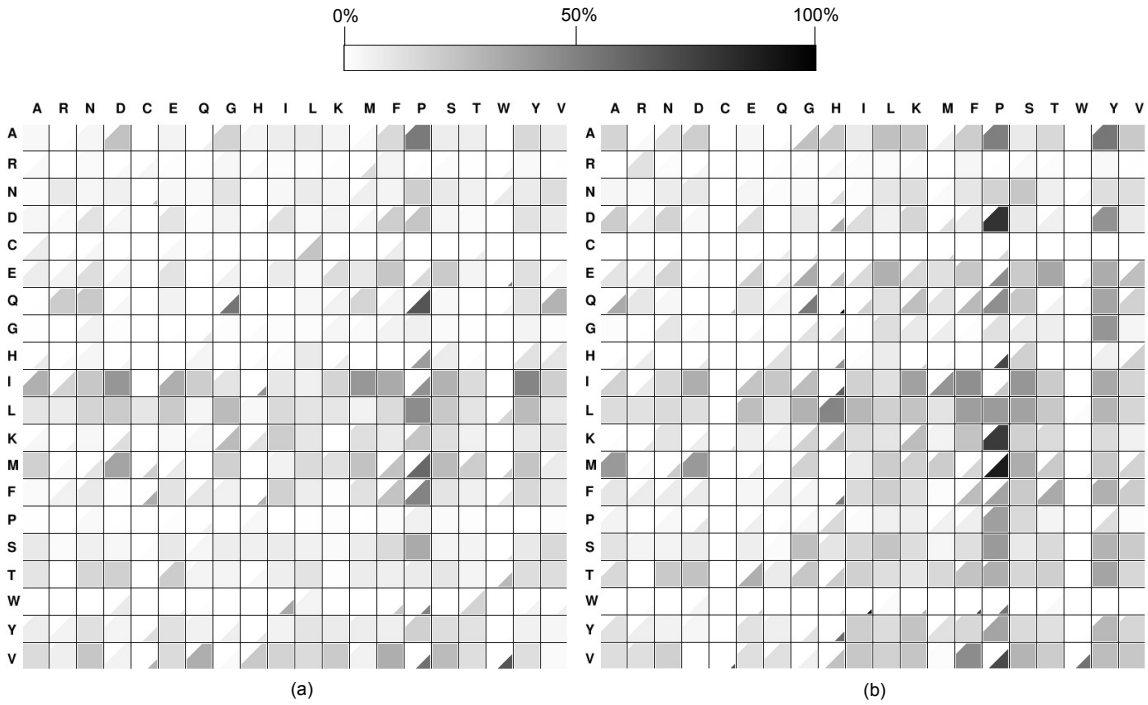


Figure 3.27 – Matrice d'intensité pour les ions *y* en fonction des deux résidus adjacents au site de clivage pour l'ensemble des peptides du CMH-I. (a) CID (b) HCD

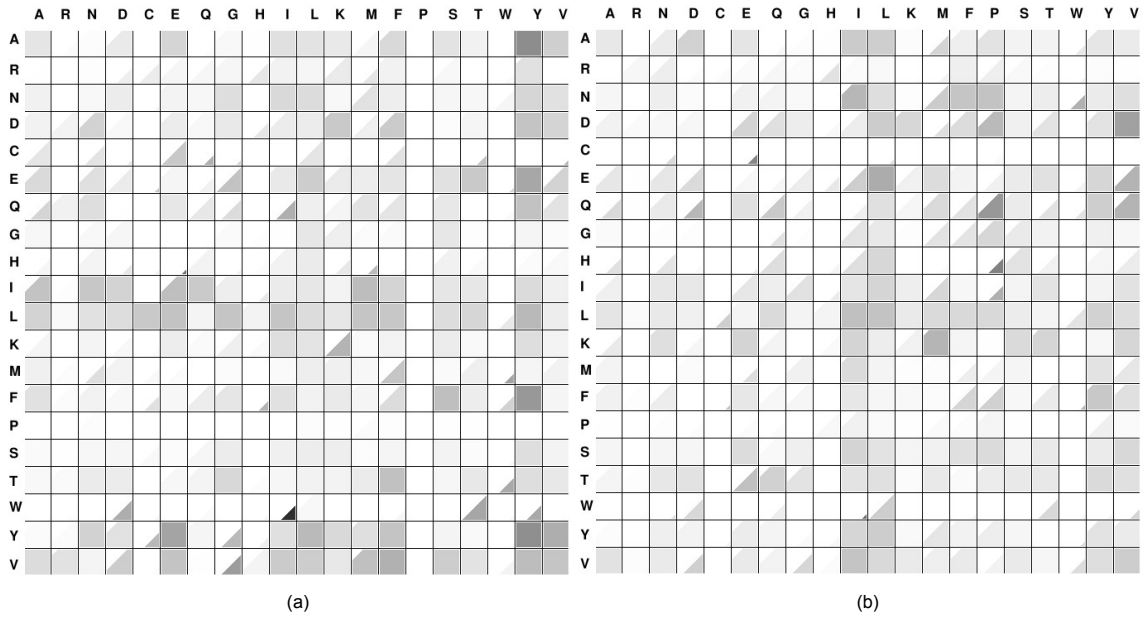


Figure 3.28 – Matrice d'intensité pour les ions *b* en fonction des deux résidus adjacents au site de clivage pour l'ensemble des peptides du CMH-I. (a) CID (b) HCD

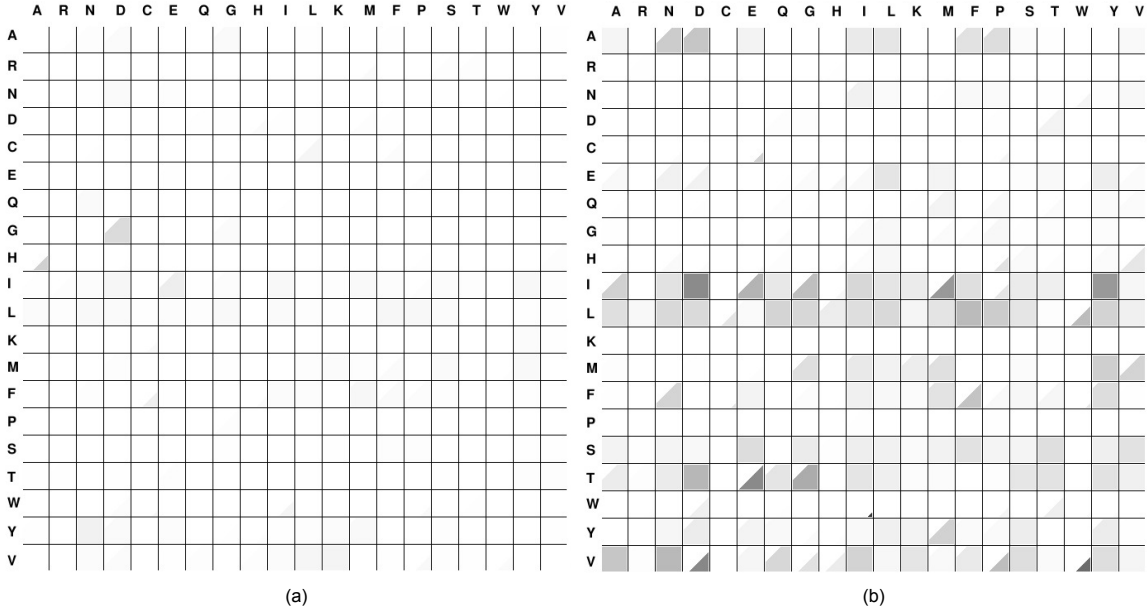


Figure 3.29 – Matrice d'intensité pour les ions *a* en fonction des deux résidus adjacents au site de clivage pour l'ensemble des peptides du CMH-I. (a) CID (b) HCD

3.5.3 L'influence des résidus non-adjacents au site de clivage

L'influence de chacun des résidus sur la fragmentation lorsqu'ils ne sont pas adjacents au site de clivage a été évalué. Pour les ions *y* et *b*, la fréquence des acides aminés est calculée en divisant leur occurrence dans les ions fragments observés par leur occurrence dans l'ensemble des séquences des peptides synthétiques. En fragmentation HCD, on peut constater que les acides aminés Arg, Lys et His se retrouvent préférentiellement dans les ions fragments *y* (figure 3.30). Alors que les résidus Ala, Asp, Gln et Asn se trouve plus souvent parmi les ions fragments *b*. On constate que l'acide aminé Met, plus fréquent parmi les peptides du CMH-I que dans le protéome, est le moins favorable à l'observation des ions *y* et *b*. En fragmentation CID, on trouve également le même résultat.

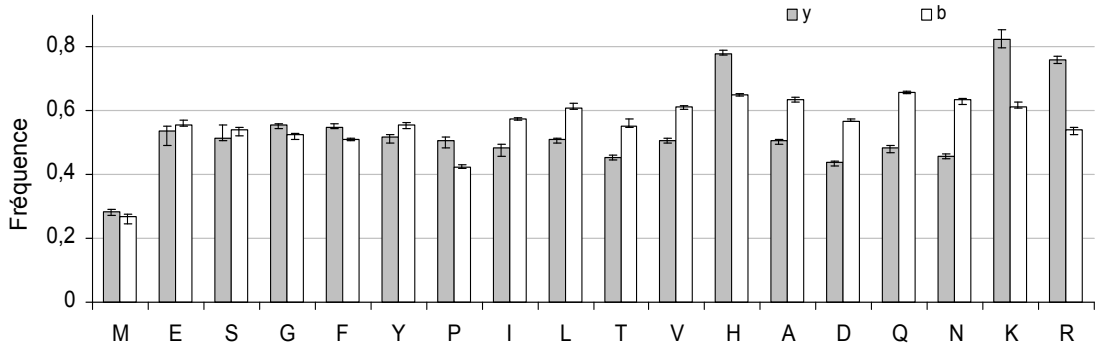


Figure 3.30 – Fréquence de chacun des acides aminés distants du site de clivage dans les ions fragments.

3.6 Une revue des acides aminés

Dans cette section, on passe en revue un certain nombre d'acides aminés pour lesquels quelques observations pertinentes ont été faites et dont les algorithmes de séquençage *de novo* devraient tenir compte.

3.6.1 Glycine

Deux glycines consécutives (G-G) conduit souvent à l'absence de clivage entre les deux résidus. C'est notamment le cas pour les peptides synthétiques du CMH-I suivant : LLPRVVGGK, AALLDGGNM, RGGVNTFLI, WFSQRGGSY, ATGTDMPGGY. Notons que 5 peptides synthétiques du CMH-I contenant un sous-séquence G-G n'ont pas été assignés par MASCOT parce qu'ils ont conduit à une fragmentation trop pauvre en information : VMCGGSLYV, LLFNILGGWV, SFGGASCCLY, QLYLGGMSYY, TIFVTGGL. De plus, l'asparagine est isomérique à deux glycines. L'absence de clivage entre les deux glycines ne permet pas de discriminer avec une asparagine.

3.6.2 Sérine

La sérine présente souvent des pertes en eau. On peut noter la présente d'un pic à une masse inférieure de 18 Da à la masse du fragment sans perte. On remarque parfois aussi la présente d'autres pics distants de 18 Da du fragment avec perte d'eau. Il se produit des pertes de molécules d'eau distantes du site de clivage.

3.6.3 Proline

Cet acide aminé est associé avec des ions fragments abondants. Nos analyses le confirment aussi avec les peptides synthétiques du CMH-I et ce quelque soit la fragmentation utilisé. Cette propriété est peu affectée par les autres acides aminés composant le peptide. À tel point que le pic associé au clivage en position N-terminal est facilement identifiable grâce à son intensité. Habituellement, les peptides contenant une proline à n'importe quelle position à l'exception de celles terminales ont une grande proportion de fragments internes avec la proline comme premier résidu. On n'observe pas cela avec

les peptides synthétiques du CMH-I.

3.6.4 Valine, leucine, isoleucine

Les acides aminés Val, Leu et Ile ont en commun une chaîne aliphatique. Et on remarque aussi qu'ils semblent avoir la même propension à favoriser la fragmentation quand ils sont situés du côté N-terminal du site de clivage.

3.6.5 Asparagine et acide aspartique

L'asparagine a une masse (114.043) très proche de celle de l'acide aspartique (115.027). Sans une bonne précision de masse, il est possible de les confondre. Il faut noter aussi que cet acide aminé est isomérique à deux glycines. De plus lors de modifications post-traductionnelles des protéines, certains acides aminés telle que l'asparagine peut subir une déamination convertissant ce dernier en acide aspartique ou en acide iso-aspartique. Les manipulations des échantillons peuvent aussi conduire à cette déamination [26]. Une erreur sur l'identification du pic monoisotopique peut conduire à une ambiguïté. Ces remarques sont valables aussi pour l'acide aspartique.

3.6.6 Lysine

La lysine (128.095) est isobarique à la glutamine (128.059). Une bonne précision de masse est donc nécessaire. Dans le cas des peptides trypsiques, il ne peut y avoir d'ambiguïté entre la lysine et la glutamine en position C-terminal. Pour les peptides auxquels nous nous intéressons, nous n'avons aucune connaissance *a priori* sur la nature de l'acide aminé en position C-terminal. Si on a connaissance de l'isoforme de la molécule du CMH-I, on peut avoir un indice. Les spectres de peptides du CMH-I ayant une glutamine en N-terminal présentent des ions fragments abondants avec des valeurs

de m/z inférieurs de 17u par rapport aux fragments a et b [29]. La lysine est isobarique de glycine-alanine. La lysine influe sur la proportion respective des fragments *b* et *y* en fonction de sa position dans la séquence.

3.6.7 Glutamine

La précision en masse est un atout pour distinguer la glutamine et la glutamate qui ont des masses très proches. La glutamine est isoméroque de glycine-alanine.

3.6.8 Méthionine

La méthionine semble être défavorable à l'observation des ions *y* et *b* lorsque cet résidu n'est pas adjacent au site de clivage.

3.6.9 Histidine

La présence d'une histidine conduit souvent (85% en mode HCD) à l'observation des ions immoniums à masse de 110 m/z. Ceci est donc une bonne indication de la présence d'une histidine dans la séquence peptidique car aucun autre acide aminé ne conduit à des ions à cette masse. L'histidine influe sur la proportion respective des fragments *b* et *y* en fonction de sa position dans la séquence.

3.6.10 Phénylalanine

Comme pour l'histidine, la présence de l'ion immoninum à un m/z de 120 permet de confirmer la présence d'un moins une phénylalanine dans la séquence peptidique.

3.6.11 Tyrosine

On peut observer la présence d'ions immoniums associé à la tyrosine dans plus de la moitié des cas en mode HCD.

3.6.12 Arginine

L'arginine influe beaucoup sur la fragmentation. S'il est en position C-terminal, il va favoriser la présence des ions *y*. En revanche, s'il est du côté N-terminal, il va favoriser les ions *b*. L'arginine est isobarique à valine-glycine.

3.6.13 Tryptophane

Cet acide aminé est très peu fréquent dans nos données. Cependant, il l'est également dans l'ensemble du protéome.

3.7 Conclusion

La comparaison de la fragmentation en mode CID des peptides du CMH-I avec les peptides trypsiques habituellement utilisés dans l'étude du protéome fait ressortir plusieurs faits intéressants :

1. les peptides du CMH-I conduisent à moins de 10% de spectres complets alors qu'avec les peptides trypsiques on atteint presque 70% ;
2. la fragmentation des peptides du CMH-I donnent lieu à moins de fragments *b*, *y* et *a* ;
3. les ions fragments des peptides du CMH-I sont globalement moins abondants ;
4. le profil de fragmentation des peptides du CMH-I est différent.

La fragmentation des peptides à laquelle on s'intéresse ici est nettement moins informative que celle des peptides trypsiques. On comprend aisément pourquoi ces peptides sont plus difficiles à identifier et à séquencer. De plus, le profil de fragmentation est différent. Par conséquent, les algorithmes basés sur l'apprentissage de données spectrales de peptides trypsiques seront certainement moins performants, à niveau d'information égal, avec les peptides du CMH-I.

Néanmoins, on montre que la fragmentation en mode HCD permet d'obtenir des spectres plus informatifs qu'en mode CID. La moitié des spectres HCD sont complets et presque 25% sont quasiment complets avec un seul clivage manquant. Les ions immoniums, les fragments internes et les ions fragments de basse masse apportent également des informations supplémentaires pour l'interprétation des spectres. De plus, la précision en masse du mode HCD permet de réduire le nombre de séquences candidates possibles pour un spectre donné. Tous ces éléments devraient permettre d'améliorer le séquençage *de novo*.

Le profil de fragmentation des peptides du CMH-I différent des peptides généralement utilisés pour la conception des algorithmes de séquençage *de novo* montre qu'il serait souhaitable de les amender spécifiquement pour ces peptides.

CHAPITRE 4

L'ÉVALUATION DES ALGORITHMES DE SÉQUENCAGE *DE NOVO* POUR LES PEPTIDES DU CMH-I

Le séquençage *de novo* présente un intérêt évident lorsqu'on s'intéresse à des séquences peptidiques mutées ou pour lesquelles aucun gène n'est séquencé. Il peut permettre aussi d'identifier des pathogènes ou des nouvelles souches de virus par exemple [7]. Il existe une diversité de programmes de séquençage *de novo* qui, pour certains, ont fait leur preuve. Il n'est pas aisé de savoir lequel d'entre eux est le plus adapté pour un cas particulier. Il existe peu d'études indépendantes qui ont évalué les performances des programmes existants. Il ressort néanmoins d'une étude relativement exhaustive que quelques uns se démarquent : Peaks, pepNovo, novoHMM [61]. Cependant, celle-ci ne s'est intéressée qu'aux peptides trypsiques. À ce jour, aucune d'entre elles ne s'est penché sur les peptides du CMH-I. Comme on l'a vu précédemment, les peptides du CMH-I ont des caractéristiques de fragmentation différentes des peptides trypsiques. La moindre qualité des spectres issus de la fragmentation des peptides du CMH-I et la plus faible quantité d'information fournie pénalisent leur séquençage par rapport aux peptides trypsiques. S'ajoute à cela le fait que la plupart des algorithmes de séquençage *de novo* ont été conçus ou entrainés pour les peptides trypsiques. On s'est néanmoins basé sur l'étude de Pevtsov et al. [61] pour faire une pré-sélection des programmes les plus susceptibles de montrer des performances acceptables avec les peptides auxquels on s'intéresse. D'autres programmes, non évalués de façon indépendantes, ont récemment été développés spécifiquement pour la fragmentation en mode HCD. On s'est intéressé également à ces derniers. Pour évaluer les différents algorithmes, un programme a été

développé. Celui-ci compare les séquences estimées par les méthodes de séquençage *de novo* avec les véritables séquences associées à chacun des spectres MS/MS. On montre de façon non équivoque que le programme Peaks est plus performant que les autres. Sa conception en explique en partie la raison. Après l'évaluation comparative des différents programmes retenus, on s'attardera davantage sur le programme montrant les meilleures performances.

4.1 Les programmes évalués

L'étude antérieure de Pevtsov et al. [61] et la compréhension des algorithmes mis en oeuvre permettent de faire une présélection. On peut s'épargner une étude exhaustive de l'ensemble des programmes existants qui serait fastidieuse et ne présenterait que peu d'intérêt. Certains programmes ont des performances médiocres (AUDENS) alors que d'autres sont spécifiquement conçus pour les peptides trypsiques au point qu'ils ne retournent que des séquences se terminant par un acide aminé Lys ou Arg (novoHMM). On s'est donc arrêté sur les algorithmes suivants : **Lutefisk, pepNovo, Peaks, pNovo** et **Vonode**. Les deux derniers ont été développés récemment et ont été conçus spécialement pour la fragmentation HCD.

4.1.1 Lutefisk

Le programme Lutefisk [78] est basé sur une approche de théorie des graphes pour la fragmentation en mode CID. Il a été développé pour les peptides trypsiques. Par conséquent, on peut s'attendre à ce qu'il donne de moins bons résultats avec les peptides du CMH-I. Le programme convertit tous les ions en ions *b* correspondants. La liste d'ions *b* obtenus est convertie en graphe. Il parcourt ensuite le graphe à partir de la po-

sition N-terminal pour déterminer les séquences candidates. Il commence par trouver tous les ions *b* possibles distants d'une masse d'un acide aminé à partir de la position N-terminal et ainsi de suite. Le score est basé sur le nombre de pics correspondant à des fragments connus et leur intensité. Une analyse de corrélation croisée est ensuite faite sur les meilleures séquences candidates. La stratégie utilisée par ce programme le rend plus sensible aux clivages manquants par exemple. La version utilisée pour notre analyse est LutefiskXP v1.0.5 (`http://sourceforge.net/projects/lutefiskxp`).

4.1.2 PepNovo

Le programme PepNovo [25] est basé sur l'utilisation des graphes. Il cherche le chemin asymétrique ayant le meilleur score à l'aide d'un algorithme de programmation dynamique. Il utilise un test d'hypothèse qui compare deux hypothèses concurrentes pour un graphe *S* et une masse *m* d'un possible site de clivage. La première hypothèse, appelée hypothèse CID, concerne la véracité pour une masse *m* d'un clivage dans le peptide qui conduit au spectre *S*. D'après cette hypothèse, il y a des règles qui régissent la fragmentation. Par exemple, Il y a certaines combinaisons de fragments et intensités qui sont plus probables que d'autres. Un modèle renseigne sur la probabilité de détecter un ensemble d'intensités de fragments observés étant donnée une masse *m* correspondant à un site de clivage dans le peptide dont résulte le spectre *S*. Un ensemble de spectres accompagnés des séquences peptidiques correspondantes est utilisé pour l'entrainement du modèle. L'hypothèse concurrente est l'hypothèse de pics aléatoires. Elle présume que les pics dans le spectre résultent d'un phénomène aléatoire. Le score pour une masse *m* et un spectre *S* correspond au logarithme du rapport de vraisemblance de ces deux hypothèses. Cet algorithme est dépendant des données d'apprentissage. Étant donné que celui-ci a été entraîné sur des spectres CID de

peptides trypsiques, il est avant tout adapté pour ce type des spectres. PepNovo ne retourne pas nécessairement la séquence complète mais privilégie l'exactitude de la sous-séquence retournée. La version utilisée est PepNovo+ (Release 2010225 : `http://proteomics.ucsd.edu/Software/PepNovo.html`).

4.1.3 Peaks

Peaks [47] utilise une approche originale par rapport à la majorité des programmes basés sur la théorie des graphes. Le programme repose sur l'utilisation d'un score de pénalité/récompense. Peaks procède en 4 étapes :

1. Prétraitement des données MS/MS brutes incluant le filtrage du bruit, le centrage des pics et la déconvolution des ions doublement et triplement chargés,

2. Détermination des séquences candidates possibles,

3. Sélection des meilleurs candidats,

4. Calcul d'un score de confiance global pour la séquence du peptide et d'un score de confiance individuel pour chacun des acides aminés.

La détermination des candidats consiste à calculer les 10000 meilleures séquences possibles pour une masse de précurseur donnée. La précision en masse sur le précurseur relève d'une grande importance pour cette étape. Les ions a, b, c, x, y, $b-H_20$, $y-H_20$, $b-NH_3$, $y-NH_3$ sont pris en compte. Plus il y a des pics calculés coïncidant avec des pics observés, plus la séquence est considérée comme probable. Pour chaque masse m, l'algorithme calcule la pénalité/récompense qu'un ion y ou b ait une masse m. S'il y a un pic proche de m, la récompense est égale au logarithme de l'intensité du pic multiplié par

un facteur proportionnel à l'erreur de masse entre m et la masse du pic observé, et multiplié par un facteur reflétant la coexistence des ions x, $y - H_20$, y-NH_3 ou a, c, b-H_20, b-NH_3. S'il n'y a pas de pic à proximité de m, une pénalité est affectée. La meilleure séquence est celle pour laquelle les ions y et b maximisent la valeur récompense moins la valeur pénalité. À la troisième étape, le score est recalculé pour l'ensemble des 10000 séquences en tenant compte notamment des ions immoniums et des fragments internes. Pour terminer, Peaks calcule un score de confiance pour les k meilleurs candidats (k étant fixé par l'utilisateur). Le score est une mesure de vraisemblance pour chaque peptide. Cette approche qui diffère de celles basées sur les graphes permet de contourner plus aisément le cas où des pics sont manquants comme par exemple c'est souvent le cas avec des glycines successives. Peaks pourra déterminer deux sous-séquences dans un même peptide avec un niveau très élevé de confiance séparés par un segment de faible confiance. On a également vu dans le chapitre 3 que contrairement aux peptides trypsiques, les peptides du CMH-I conduisent plus souvent à des spectres pour lesquels certains clivages sont manquants. Il faut noter que la stratégie de Peaks repose grandement sur la masse du précurseur. Une bonne précision de masse ne peut que conduire à une meilleure confiance et à la réduction de taux de faux positifs. Comme Peaks se base sur la masse du précurseur, autrement dit la masse du peptide, il va donc proposer des solutions en accord avec cette masse. Bien qu'il puisse arriver que des spectres soient très informatifs pour une sous-séquence du peptide, Peaks va proposer une solution globale au détriment de l'exactitude pour la région du spectre de qualité. Ces remarques faites, Peaks semble être tout adapté pour des peptides de petite taille et une bonne précision de masse sur le précurseur et ne pâtit pas autant que les autres méthodes de l'absence de pics.

4.1.4 pNovo

Le programme pNovo [12] utilise les informations supplémentaires fournies par les spectres HCD tels que les fragments internes et les ions immonium. De plus, il tire profit de la haute précision en masse. Comme nombre de programmes de séquençage *de novo*, il est basé sur l'utilisation de la théorie des graphes. Il y a une étape de prétraitement qui consiste en la détermination de la charge des pics, la transformation des intensités absolues en rangs relatifs [5], la suppression dans le spectre des ions immoniums afin de simplifier son interprétation et enfin l'extraction des k pics les plus intenses pour la construction du graphe spectral. La construction du graphe spectral suit les étapes suivantes :

1. Chaque pic est séparé en un maximum de 6 sommets correspondants aux différents types d'ions : y, b, $y-NH3$, $y-H2O$, a, et y doublement chargé. Par exemple, s'il y a un pic à un m/z de 796,54 dans un spectre pour lequel la masse du précurseur est 1387,76 Da, et que les ions y et b sont pris en considération, alors deux sommets, localisés à un m/z de 796,54 et un m/z de 591,22 sont générés. À chaque sommet est associé un poids proportionnel à l'intensité du pic correspondant. L'association sommet-pic est générée seulement si la probabilité d'observation du pic est supérieur à 0,1.

2. Si plusieurs sommets sont associés à des fragments de même masse (à une tolérance de masse près), ils sont fusionnés. Le nouveau poids associé au sommet fusionné correspond à la somme des poids de chaque sommet.

3. Il connecte ensuite les sommets pour lesquels la différence de masse correspond à celle d'un acide aminé ou la somme de plusieurs résidus. Le poids de chaque arc

correspond à la somme des poids des sommets reliés par celle-ci.

4. Enfin, à chaque arc sont assignés des nouveaux poids en tenant compte de la précision de masse et l'observation des ions immoniums et des fragments internes. Le poids est multiplié par une pénalité corrélée à la différence avec la masse théorique.

Une fois le graphe construit, Il cherche les chemins asymétriques ayant les meilleurs scores à l'aide d'un algorithme de recherche en profondeur d'abord avec une stratégie d'élagage [12] . Le calcul du score repose sur la différence de masse entre les pics observés et les pics théoriques. Le programme est donc très sensible à la précision de masse. En conséquence, une mauvaise calibration peut nuire à l'élucidation de la séquence. Le programme exécutable nous a été directement livré par Hao Chi (Institute of Computing Technology, Chinese Academy of Sciences, Beijing).

4.1.5 Vonode

Le programme Vonode [58] a été développé pour tirer profit de la haute précision en masse. Le calcul du score repose davantage sur la précision (ou erreur) de masse que l'intensité des pics comme dans la plupart des autres programmes. Grâce à la haute précision en masse, les ions fragments avec de faible intensité peuvent être pris en considération et correspondre à des ions *b* et *y* à l'instar de ceux d'intensité plus importante. Vonode utilise un nouveau type de graphe dans lequel les ions observés ne sont transformés qu'en un seul et unique sommet et 4 types d'arcs sont utilisées pour représenter les différentes relations possibles entre des ions fragment adjacents. Un résidu est flanqué par les ions y_h, y_{h+1}, b_{k-1} et b_k. Un arc relie l'ion y_h au prochain ion y_{h+1} (ou b_{k-1} à b_k). Une arc relie y_h au prochain ion b_{k-1} complémentaire. Une arc relie y_{h+1} à l'ion

complémentaire précédent b_k. Enfin, une arc relie deux ions *b* et *y* complémentaires. La résolution du graphe est plus complexe que dans les autres programmes mais le calcul du score est simplifié par la prise en compte uniquement de la masse des fragments. Vonode cherche à trouver une sous-séquence ("sequence tags") correspondant parfaitement à la vraie séquence. Autrement dit, Vonode retourne rarement la séquence complète. Il fonctionne en 4 étapes :

1. la construction du graphe spectral avec les 4 types d'arcs,
2. la construction du graphe de la séquence à partir du graphe spectral,
3. la recherche des sous-séquences,
4. et enfin le calcul des scores.

Pour plus de détails, on peut se référer à la publication dans laquelle l'algorithme est expliqué [58] . Il y a 3 remarques importantes à retenir :

1. les performances sont affectées par l'erreur de masse,
2. les séquences retournées sont majoritairement des sous-séquences,
3. les séquences inverseées des séquences candidates sont à prendre en considération (par exemple, si la séquence ANDREWS est retournée, il faut tenir compte de la séquence SWERDNA).

La version utilisée est disponible à l'adresse suivante : `http://compbio.ornl.gov/Vonode/`.

4.2 L'algorithme de comparaison des algorithmes de séquençage *de novo*

Évaluer les performances d'un algorithme de séquençage *de novo* consiste à mesurer la proximité des séquences proposées par celui-ci à la séquence réelle associé au spectre. Le problème se résume donc à calculer la similarité entre deux séquences nucléiques, d'une part la véritable séquence du peptide (que nous appellerons séquence réelle) et la séquence (ou les séquences) retournée(s) par l'algorithme (que nous appellerons séquence *de novo*).

Les séquences réelles et *de novo* sont en fait des chaînes où chacun des acides aminés est symbolisé par une lettre appartenant à un alphabet $\sum$ de 20 lettres ($\sum$ = {A, C, D, E, F, G, H, K, M, N, P, Q, R, S, T, V, W, J, Y}). Pour mesurer la similarité de deux séquences S_1 et S_2, on utilise ce qu'on appelle une fonction de similarité $\alpha(s_1, s_2)$ (l'inverse de la fonction de distance utilisée par ailleurs). L'algorithme de comparaison de séquences est inspiré de l'algorithme de Needleman-Wunsch [54] couramment utilisé en bioinformatique. Ce dernier effectue un alignement global maximal des deux séquences peptidiques en garantissant de trouver l'alignement de score maximal. Cependant, il faut tenir compte de spécificités liées à la spectrométrie de masse et la fragmentation. Car ici on ne s'intéresse pas à évaluer l'ensemble de la chaine de traitement conduisant *in fine* à la détermination de la séquence mais uniquement l'algorithme de séquençage *de novo* qui est tributaire de la qualité et l'informativité du spectre qui lui est fourni.

4.2.1 La matrice de similarité des acides aminés

L'algorithme d'alignement et de comparaison de séquences utilise une matrice de similarité des acides aminés. Celle-ci renseigne, comme son nom l'indique, sur la similarité entre chacun des acides aminés. Contrairement aux cas plus complexes où l'on

cherche à mesurer une similarité de nature phylogénique ou à prédire la structure secondaire d'une protéine, on s'intéresse ici à mesurer la similarité du point de vue de la masse. La similarité $s(a_1, a_2)$ traduit tout simplement la différence en masse de deux acides aminés a_1 et a_2 :

$$s(a_1, a_2) = 100 \times \left[1 - k \times \frac{|m(a_1) - m(a_2)|}{max[m(a_1), m(a_2)]}\right] \tag{4.1}$$

La matrice de similarité est ainsi construite par l'application de cette formule. La similarité entre deux résidus de même masse est égale à 100. Quelques soient les performances des algorithmes de séquençage *de novo* existants, ils ne font pas la distinction entre les acides aminés Ile et Leu. Par conséquent $s(L,I) = s(I,L) = 100$. Les ions immoniums associés à ces deux acides aminés ont également la même masse. Seuls les fragments *d* et *w*, peu abondants, peuvent permettre de discriminer entre l'un et l'autre mais ils ne sont pas utilisés par les algorithmes. Par conséquent, les séquences ALIEN et ALLEN, par exemple, sont considérées identiques dans notre évaluation. De plus, si la précision de masse est moins bonne que 0.1 Da, il est impossible de distinguer entre les acides aminés Lys et Asp avec les seuls fragments *y* et *b*. La valeur de *k* a été fixée à 1,38 de façon à ce que la plus faible valeur de similarité entre deux acides aminés soit égale à 4,3 (Trp *versus* Gly).

4.2.2 La permutation des résidus adjacents

Il arrive fréquemment que des résidus adjacents soient intervertis à cause de l'absence de l'ion fragment correspondant au clivage de la liaison peptidique les reliant. Cette absence peut être due au fait que le peptide n'a pas été fragmenté entre les deux résidus ou que le fragment n'a pas été détecté. L'algorithme de comparaison tient compte

de cette erreur et n'affecte pas le score de similarité puis que l'algorithme de séquençage *de novo* n'est pas responsable de celle-ci. Par exemple, les séquences ANDREWS et NADREWS seront considérées comme parfaitement similaires.

L'algorithme calcule donc la similarité entre la séquence *de novo* et toutes les séquences contenant les permutations générées à partir de la séquence réelle. Par exemple, pour la séquence ALIEN, on aura les séquences suivantes : ALIEN, ALINE, ALNIE, AILEN, AILNE, LAIEN, LAINE, LAEIN. Le nombre maximum de séquences de longueur l contenant des permutations de résidus adjacents, noté $N(l)$, se calcule par la formule de récurrence suivante :

$$N(l) = \begin{cases} 1 \, si \, l = 1 \\ 2 \, si \, l = 2 \\ sinon \, N(l-1) + N(l-2) \end{cases} \tag{4.2}$$

Il peut y en avoir moins si deux résidus adjacents sont identiques.

Dans la présente étude, nous disposons de spectres et des séquences correspondantes. Le programme StatPeaks fournit dans le fichier IDE les fragments *y* et *b* présents dans le spectre. Par conséquent, une autre approche possible est de générer à partir de ces fragments observés et identifiés, toutes les séquences possibles. Cette alternative à la génération de séquences avec permutation telle décrite plus haut demande une étape supplémentaire. Elle n'a pas été retenue et n'est pas décrite ici.

4.2.3 Le calcul des scores

Pour évaluer les algorithmes, 3 scores différents sont calculés afin de tenir compte des spécificités de chacun d'entre eux.

4.2.3.1 Le score de similarité

On s'intéresse à déterminer le degré de similarité entre la séquence réelle et la séquence *de novo*. Pour ce faire, on utilise un algorithme d'alignement. L'algorithme inspiré de l'algorithme de Needleman-Wunsch utilise une pénalité de trou égale à 0. On désigne la séquence réelle par S_r de longueur n et la séquence *de novo* par S_i de longueur m. L'algorithme construit une matrice à deux dimensions. Le nombre de lignes est égal à la longueur de S_r alors que le nombre de colonnes est égal à la longueur de S_i. Les rangs des lignes et des colonnes sont respectivement notés i et j. Pour chacun des résidus de S_r est associée une colonne et pour chacun des résidus de S_i est associé une ligne. Au fur et à mesure de la progression de l'algorithme, un score optimal F_{ij} est calculé pour les i premiers résidus de S_r et les j premiers résidus de S_i. Une fois la matrice complétée, la valeur de F_{nm} correspond au score maximal et celui-ci est divisé par la plus petite longueur des deux séquences. On définit la similarité par :

$$\alpha(S_r, S_i) = \frac{F_{nm}}{min(n,m)} \tag{4.3}$$

Notons que si la séquence *de novo* est une sous-séquence de la séquence réelle, on considère que la similarité est totale. Par exemple, pour les séquences S_r = ALDENTE et S_i = DENTE, nous obtenons un score maximal de 500 car il y a 5 résidus identiques et 2 trous au coût de 0. La similarité $\alpha(S_r, S_i)$ est égale à 100.

S_r :	A	L	D	E	N	T	E
S_i :	-	-	D	E	N	T	E

Ce score permet d'évaluer l'exactitude de la séquence retournée sans tenir compte de l'exhaustivité. Pour certaines applications, il est plus important d'obtenir une sous-séquence exacte qu'une séquence se voulant exhaustive mais comportant des erreurs. La

longueur des peptides du CMH-I varie entre 8 et 12 résidus pour la plupart. Obtenir une sous-séquence de 7 résidus peut être suffisant pour permettre de procéder à un "blast" et identifier la protéine source sans ambiguité dans la plupart des cas. Il est recommandé aux moins 6 résidus pour la recherche avec l'algorithme *blastp* [21].

4.2.3.2 Le taux d'identité

Le taux d'identité *Id* de séquences se mesure par le nombre d'acides aminés identiques obtenu après alignement. À partir de l'alignement effectué avec l'algorithme décrit plus haut, on compte le nombre d'acides aminés identiques entre la séquence *de novo* et la séquence réelle en tenant compte de leur emplacement dans la séquence. Dans l'exemple suivant, 5 résidus sur 7 sont identiques, on a donc une identité de 5/7 = 71,43%.

S_r : R G **D E N T E**
S_i : N V **D E N T E**

4.2.3.3 La plus longue sous-séquence correcte

Une des approches fréquemment rencontrée dans la littérature pour évaluer les algorithmes de séquençage *de novo* est l'utilisation du pourcentage de sous-séquences contigües correctes [25] [23]. Tous les algorithmes ne prédisent pas des peptides complets correspondant nécessairement à la masse du précurseur. C'est le cas notamment de PepNovo et de Vonode. Il arrive fréquemment que les pics à proximité des positions terminales ne soient pas détectés ou soient confondus avec le bruit. Comme vu au chapitre 3, les fragments y_{n-1} et b_1 pour des peptides du CMH-I de n résidus sont souvent de faible intensité. Par conséquent, les sous-séquences correctes se situent le plus sou-

vent dans le milieu du peptide. Dans l'exemple ci-dessous, la plus longue sous-séquence correcte est DENTE d'une longueur de 5 résidus.

S_r : **A** **L** R **D** **E** **N** **T** **E** G
S_i : **A** **L** N **D** **E** **N** **T** **E** V

4.3 L'évaluation des 5 algorithmes

Dans cette section sont présentés les résultats de comparaisons faites sur les 5 programmes que nous avons retenus. Par de soucis concision, ne sont exposés que les résultats des analyses menées avec les peptides du CMH-I. Pour procéder aux tests, nous avons utilisé les librairies de spectres décrites au chapitre 3. Pour les peptides du CMH-I, nous disposons de 490 spectres en mode CID et 415 en mode HCD. Pour les peptides trypsiques, nous disposons de 2328 spectres CID.

Pevzner et al. ont calculé l'abondance des peptides homéométriques [5] et leur impact sur l'élucidation des spectres de masse. Même si une haute précision en masse permet d'en diminuer le nombre, il en subsiste une quantité importante notamment avec les peptides du CMH-I pour lesquels il y a plus souvent des clivages manquants. Cela signifie que plusieurs séquences peptidiques peuvent correspondre à un même spectre. On les appelle des *séquences équiprobables*. L'abondance de celles-ci oblige de tenir compte des k premières séquences candidates retournées par le programme de séquençage *de novo*. Dans les analyses, les 5 premières séquences sont prises en compte (k = 5).

En mode CID, pour chacun des programmes, nous avons utilisé les paramètres par défaut. Aucune enzyme n'a été spécifiée quand ce paramètre était ajustable. La tolérance sur la masse du précurseur a été fixée à 20 ppm (ou 0,02 Da) et celle sur les fragments

à 0,5 Da. En mode HCD, la tolérance sur la masse du précuseur a été fixé à 10 ppm (ou 0,01 Da) et 0,02 Da pour les fragments.

4.3.1 La comparaison des scores de similarité

Les algorithmes Lutefisk et PepNovo ont été développés pour le mode CID. Leurs performances sont donc optimales pour ce mode de fragmentation. On peut constater que PepNovo est l'algorithme qui donne les meilleurs résultats en mode CID avec les peptides synthétiques du CMH-I.

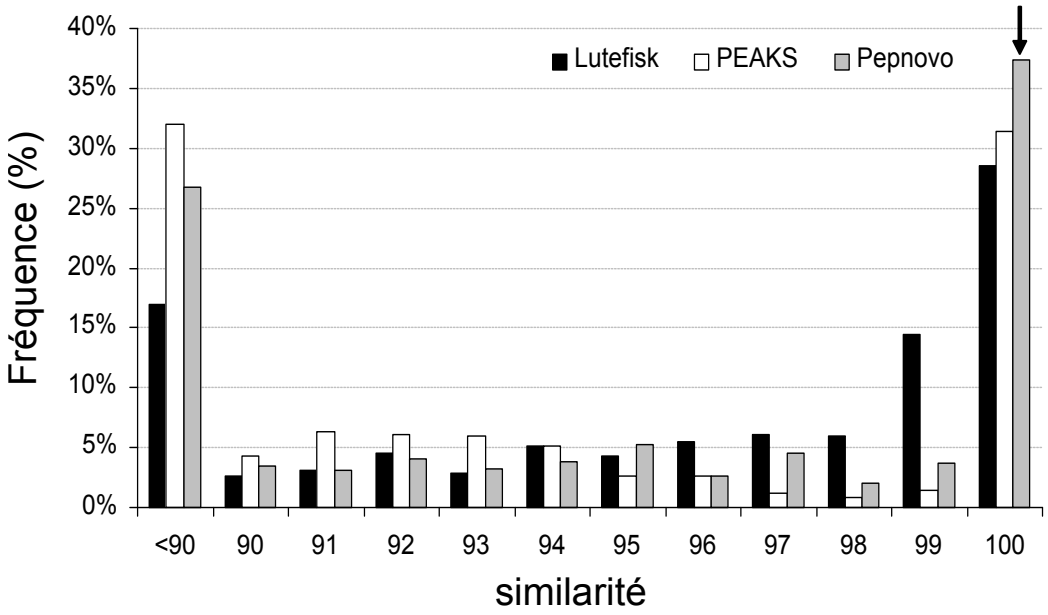


Figure 4.1 – Comparaison des distributions de score de similarité pour les programmes Lutefisk, Peaks et Pepnovo pour les peptides du CMH-I en mode CID. Près de 37% des spectres CID sont correctement séquencés avec le programme PepNovo. Peaks donne de moins bons résultats avec seulement environ 32% des spectres correctement séquencés. Lutefisk est sensiblement moins bon que les deux premiers programmes pour ce qui est de séquençage avec un score de similarité de 100. Cependant, ce dernier fournit moins de séquences éloignées de la vraie séquence avec seulement 17%.

En mode HCD, Peaks donne de meilleurs résultats que les autres programmes. Ces résultats sont probablement attribuables à l'utilisation par Peaks d'informations supplémentaires fournies par les spectres HCD tels que les ions immoniums et les fragments

internes ainsi qu'à la bonne précision en masse fournie. Plus de 50% des peptides du CMH-I sont correctement séquencés par Peaks. PepNovo est deux fois moins performant, alors que Lutefisk fournit des résultats intermédiaires. Les résultats de pNovo sont surprenants étant donné que cet algorithme a été développé spécifiquement pour les spectres HCD. Le fait que ce programme ait été développé avant tout pour les peptides trypsiques est une partie de l'explication d'après les auteurs [11]. En annexe V, on montre la même comparaison mais en ne tenant compte que de la première séquence. On peut constater que le nombre de séquences similaires est sensiblement moindre. Par exemple avec Peaks, on obtient 41,61% en ne tenant compte que de la première séquence retournée et 51,3% avec les 5 premières, soit une différence de 17,8%.

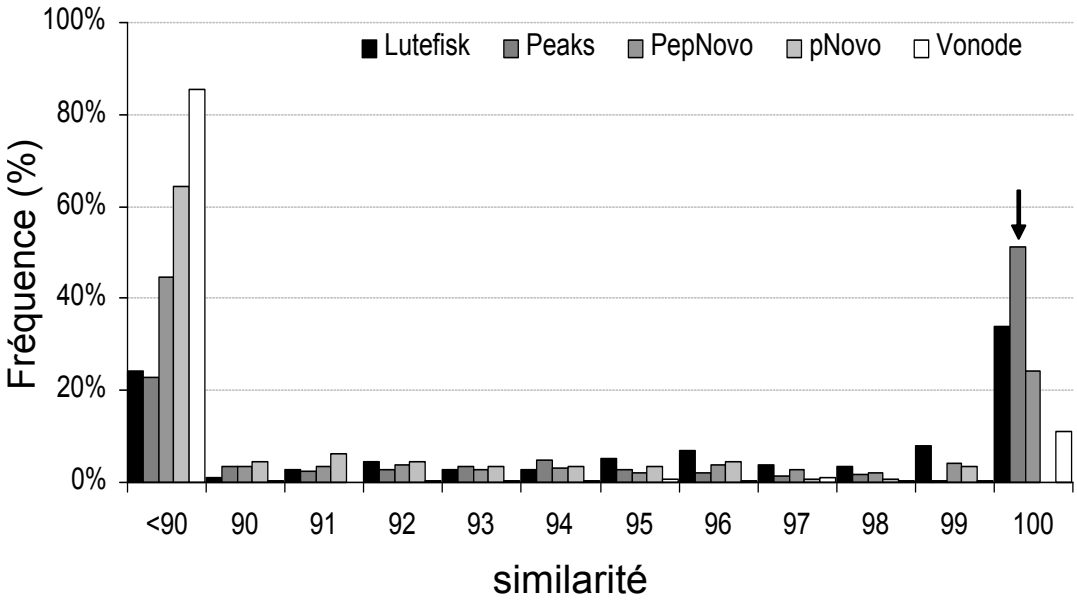


Figure 4.2 – Comparaison des distributions de score de similarité pour les programmes Lutefisk, Peaks, Pepnovo, pNovo et Vonode pour les peptides du CMH-I en mode HCD. Plus 50% des spectres HCD sont correctement séquencés avec le programme Peaks. Lutefisk arrive second, suivi de PepNovo. pNovo ferme la marche avec 0% de spectres correctement séquencés.

4.3.2 La comparaison des taux d'identité

Quand on s'intéresse au taux d'identité en mode CID, Peaks montre de meilleurs résultats que PepNovo. Ces résultats ne sont pas contradictoires avec ceux exposés précédemment mais complémentaires. Pour aucun des spectres CID, PepNovo ne retourne la séquence entière mais seulement une sous-séquence. La plupart d'entre elles ont une longueur comprise entre 70 et 80% de la longueur de la séquence réelle. Le programme Peaks retourne systématiquement des séquences pour lesquelles la masse correspond à la masse du précurseur. Donc les séquences *de novo* qui ont une similarité égale à 100 ont des taux d'identité de 100%. Ces résultats illustrent les fonctionnement différents des deux programmes. En mode HCD, Peaks montre des résultats nettement meilleurs que les autres programmes.

Tableau 4.I – Comparaison des distributions de score de similarité pour les programmes Lutefisk, Peaks et Pepnovo pour les peptides du CMH-I en mode CID.

	Lutefisk	Peaks	PepNovo
$0 \leq Id < 10$	4,9%	1,2%	2,4%
$10 \leq Id < 20$	6,1%	3,7%	9,4%
$20 \leq Id < 30$	11,4%	5,3%	11,2%
$30 \leq Id < 40$	14,9%	10,8%	13,3%
$40 \leq Id < 50$	12,0%	9,8%	11,8%
$50 \leq Id < 60$	17,8%	10,6%	11,0%
$60 \leq Id < 70$	14,9%	12,0%	14,5%
$70 \leq Id < 80$	10,2%	8,6%	**21,4**%
$80 \leq Id < 90$	3,7%	5,9%	4,1%
$90 \leq Id < 100$	0,2%	0,6%	0,8%
$= 100$	3,9%	**31,4**%	0,0%

4.3.3 La comparaison des plus longues sous-séquences correctes

Si on compare les performances des différents algorithmes en fonction du critère de la plus longue sous-séquence correcte, Peaks donne encore les meilleurs résultats. Avec

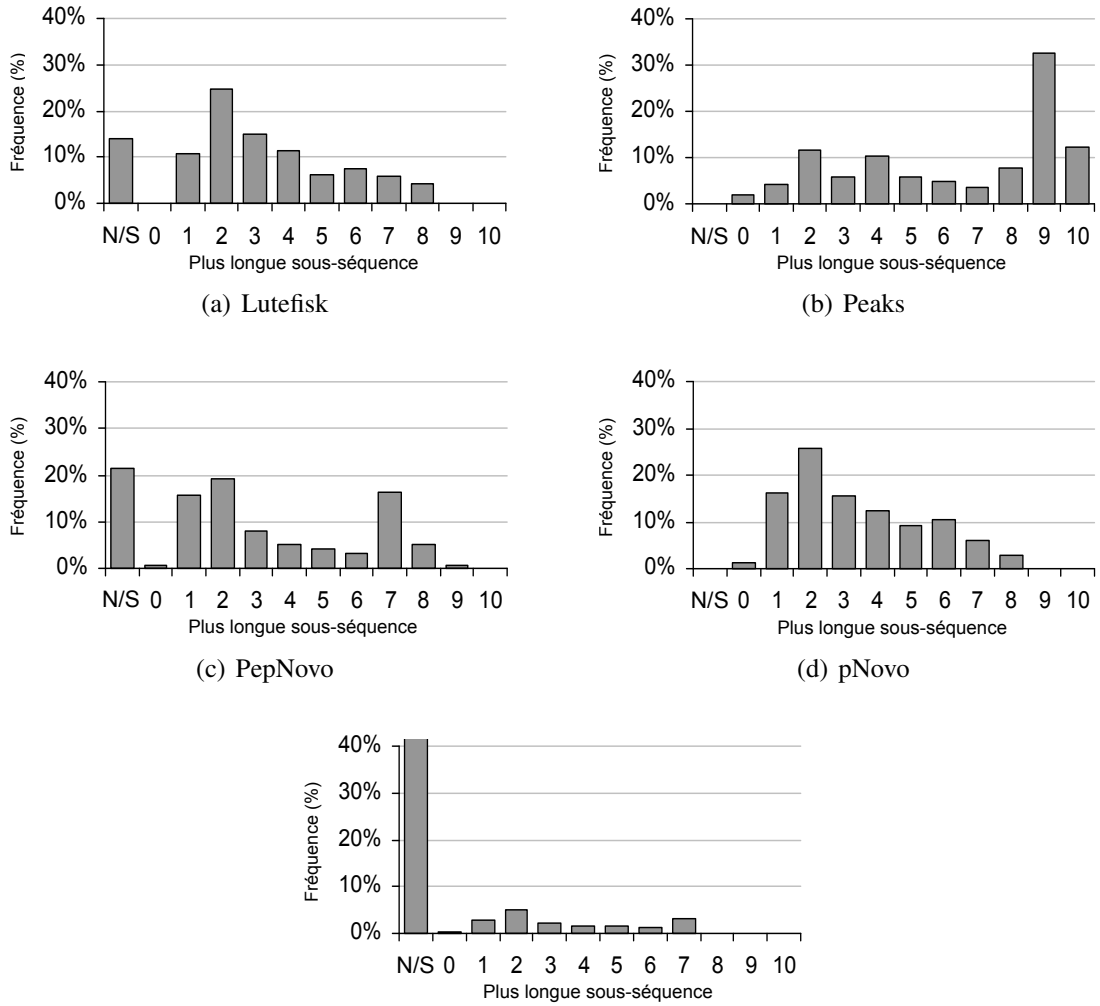
Tableau 4.II – Comparaison des distributions de score de similarité pour les programmes Lutefisk, Peaks, Pepnovo, pNovo et Vonode pour les peptides du CMH-I en mode HCD.

	Lutefisk	Peaks	PepNovo	pNovo	Vonode
$0 \leq Id < 10$	13,5%	0,2%	22,4%	2,7%	82,4%
$10 \leq Id < 20$	3,6%	2,2%	10,4%	7,0%	1,0%
$20 \leq Id < 30$	9,4%	3,4%	11,1%	15,2%	4,6%
$30 \leq Id < 40$	19,3%	5,5%	12,8%	16,9%	2,7%
$40 \leq Id < 50$	13,7%	7,0%	10,1%	14,0%	2,9%
$50 \leq Id < 60$	17,1%	7,0%	6,3%	16,1%	2,2%
$60 \leq Id < 70$	12,0%	9,4%	8,9%	15,9%	3,1%
$70 \leq Id < 80$	8,4%	9,6%	12,3%	9,4%	1,2%
$80 \leq Id < 90$	2,7%	3,1%	5,5%	2,9%	0,0%
$90 \leq Id < 100$	0,2%	1,2%	0,2%	0,0%	0,0%
$= 100$	0,0%	**51,3**%	0,0%	0,0%	0,0%

celui-ci, plus de 60% des séquences *de novo* contiennent une sous-séquence d'au moins 6 résidus corrects (figure 4.3). Lutefisk, PepNovo et pNovo sont moins bons et donnent des résultats relativement proches. Vonode ne retourne une séquence candidate que pour une petite fraction des spectres.

4.4 L'analyse détaillée de Peaks

Il ressort clairement des analyses exposées plus haut que le programme Peaks fournit les meilleurs résultats avec les peptides du CMH-I. On va donc s'attarder davantage sur ce programme. Il n'y a pas d'algorithme de séquençage *de novo* performant qui ne fournisse pas de score permettant d'évaluer la vraisemblance des solutions proposées. La particularité de Peaks est que la somme des scores de l'ensemble des séquences proposées est égale à 100. Ce qui signifie qu'en théorie s'il propose deux séquences équiprobables, le score associé à chacune des solutions est égale à 50. En pratique Peaks propose toujours plus de séquences. La figure 4.4 montre la distribution du nombre de séquences *de novo* en fonction du score Peaks lorsque l'on tient compte uniquement de



(a) Lutefisk

(b) Peaks

(c) PepNovo

(d) pNovo

(e) Vonode

Figure 4.3 – Comparaison des distributions des plus longues sous-séquences correctes pour les programmes Lutefisk, Peaks, Pepnovo, pNovo et Vonode pour les peptides du CMH-I en mode HCD.

la première solution proposée d'une part et des 20 premières d'autre part. On peut voir 4 pics correspondants approximativement aux scores 12, 24, 49 et 99. La courbe de distribution est construite de telle sorte que 12 signifie entre 12 et 13, 24 entre 24 et 25, etc.

On peut voir un pic vers le score de 99 (donc correspondant à des scores compris entre 99 et 100). Les séquences *de novo* pour lesquelles le score est supérieur à 99

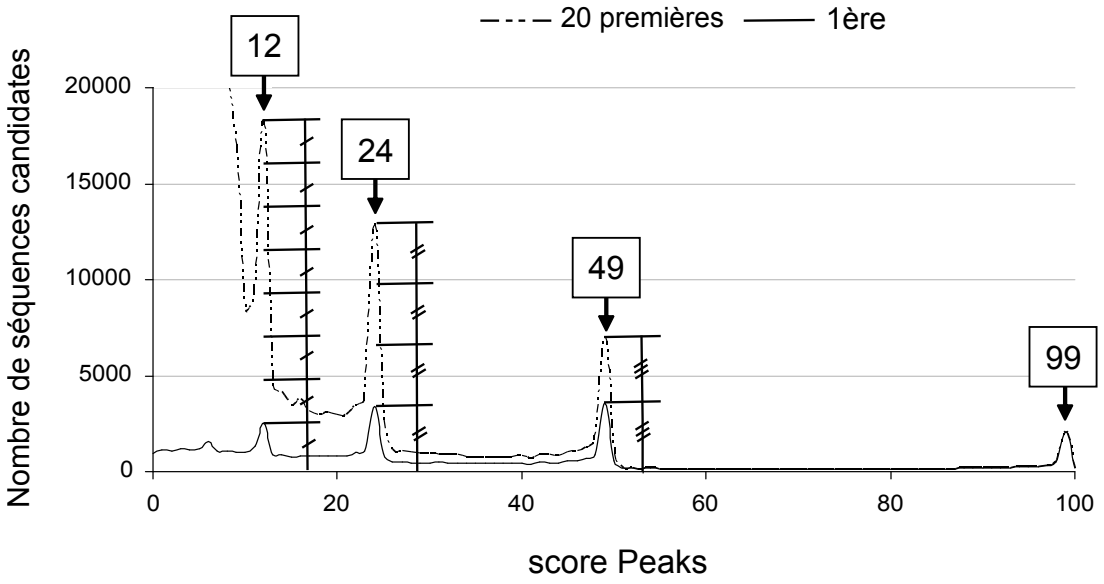


Figure 4.4 – Distribution du nombre de séquences *de novo* en fonction du score Peaks calculée à partir de 2400 spectres HCD. Il y a des pics à proximité des scores 12, 24, 49. Ces scores correspondent à des peptides homéométriques. Le pic à un score de 99 correspond aux spectres pour lesquels seule une séquence candidate a une forte plausibilité d'être la bonne séquence.

sont toujours exactes (100%) à condition de spécifier au programme les modifications chimiques post-translationnelles pertinentes. Dans le cas des peptides synthétiques du CMH-I, nous n'avons pas pris en compte de l'oxydation de la méthionine, ce qui a conduit à un certain nombre de faux positifs pour lesquels les scores Peaks sont supérieurs à 99 (voir en annexe VI. Ceci illustre l'importance de tenir compte des modifications post-transcriptionnelles car elles peuvent conduire Peaks à proposer une séquence erronée et ce avec un score élevé. Les autres séquences candidates qui suivent ont des scores proches de 0 de telle sorte que la somme des scores soit égale à 100.

Les séquences *de novo* avec un score proche de 49 sont pour la plupart équiprobables. Par exemple, Peaks propose pour le spectre du peptide YLLEKSRAI deux séquences *de novo* avec le même score comme le montre le tableau 4.III.

La séquence *de novo* correspondant à la séquence réelle est la première. Mais dans

Tableau 4.III – Séquences *de novo* proposées par Peaks pour le spectre du peptide YLLEKSRAI

séquence *de novo*	score Peaks
YLLEKSRAL	49,634987
YLLEKSARL	49,634987
YLLEKWKL	0,12109855
YLLEKRGTL	0,0753777
YLLEKGRTL	0,0753777

d'autres cas tout aussi fréquents, il s'agit de la deuxième séquence proposée. En fait, les 2 premières séquences *de novo* sont équiprobables. Les séquences suivantes sont beaucoup moins probables. On peut constater que la hauteur du pic correspondant au score de 49 en tenant compte des 20 premières séquences *de novo* est le double de celle qui correspond à la prise en compte d'uniquement la première solution.

Le pic à un score de 24 est quatre fois plus haut quand on prend en compte les 20 premières séquences *de novo* que lorsqu'on prend uniquement la première séquence. En fait, il s'agit des cas où Peaks propose 4 séquences équiprobables. C'est le cas par exemple avec le spectre du peptide SLYQYVRL. Quand on examine le spectre, on constate que le clivage entre les acides aminés Ser et Leu est manquant. La masse cumulée de ces deux acides aminés est égale à 200,116 Da. Il se trouve que cette masse peut correspondre à plusieurs combinaisons comme montré dans l'annexe IV : A-E, C-P, I-S, L-S et T-V. Cependant grâce à la haute en précision en masse sur les fragments (< 0,02), on réduit les combinaisons possibles à : I-S, L-S et T-V. Étant donné le spectre, Peaks ne peut privilégier l'une des combinaisons et ne peut pas plus l'établir l'ordre des deux résidus. Par conséquent, il propose 4 séquences équiprobables. Il propose également la séquence SLY**GA**YVRL où la masse de G-A correspond à celle de Q. Il existe en effet une probabilité très faible mais néanmoins non nulle que le pic qui correspondrait au clivage entre les acides aminés G et A n'ait pas été détecté.

Tableau 4.IV – Séquences *de novo* proposées par Peaks pour le spectre du peptide SLYQYVRL

séquence *de novo*	score Peaks
SLYQYVRL	24,989687
TVYQYVRL	24,989687
LSYQYVRL	24,989687
VTYQYVRL	24,989687
SLY**GA**YVRL	0,005113327

De la même manière, le pic à un score de 12 corresponds au cas où Peaks propose 8 séquences équiprobables. Ce qui explique pourquoi on a 8 fois plus de séquences *de novo* avec une score compris entre 12 et 13 lorsqu'on tient compte des 20 premières séquences.

On a classifié les spectres en fonction de la distribution des scores Peaks pour chacune des 20 premières séquences *de novo* comme montrée par la figure 4.5. L'ensemble 8 correspond au cas idéal où Peaks propose une solution avec un score > 90. Dans ce cas, seule la première séquence *de novo* peut être prise en compte. L'ensemble 4 est proche de l'ensemble 8 mais la première solution a un score compris seulement entre 60 et 90. L'ensemble 7 correspond au cas où deux séquences *de novo* équiprobables sont proposées. Il convient dans ce cas de tenir compte de ces deux solutions en déterminant par exemple une séquence consensus comme YLLEKS[RA|AR][I|L], où [RA|AR] signifie que RA et AR sont possibles. L'ensemble 3 correspond à 4 solutions équiprobables. Il convient ici aussi de tenir compte des 4 solutions surtout si celles-ci ont un taux d'identité élevé entre elles. Les ensembles 5 et 2 contiennent des spectres pour lesquels des solutions proposées sont équiprobables mais sont en concurrences avec d'autres ayant des scores non négligeables. L'ensemble 2 inclut aussi les spectres pour lesquels 8 solutions équiprobables sont proposées. L'ensemble 6 correspond à des spectres peu informatifs. Peaks construit donc des séquences *de novo* à partir de la masse du précurseur et des

quelques pics permettant de confirmer la présence de certains acides aminés et ou de petites sous-séquences. Ce cas de figure est le plus défavorable. L'ensemble 1 contient des séquences *de novo* avec des scores plus élevés que dans l'ensemble 6 mais les spectres ont une incomplétude supérieure à 0,3. On constate que dans la plupart des cas la première solution proposée n'est pas correcte. Ces analyses montrent que l'utilisation d'un seuil sur le score n'est pas pertinente avec Peaks. Le cas où l'on a deux séquences *de novo* équiprobables avec un score proche de 49 est plus facilement exploitable qu'avec un spectre pour lequel la première séquence *de novo* a un score de 60 et que le score décroit progressivement avec le rang des séquences proposées. Dans le premier cas, il est souvent possible de construire un séquence consensus informative. Dans le deuxième cas, les séquences sont souvent trop divergentes.

4.5 Le filtrage des séquences *de novo*

L'association du séquençage *de novo* et de la recherche de similarité de séquences dans les bases de données de protéines permet de combiner les avantages de l'un avec ceux de l'autre. En effet, le séquençage *de novo* permet d'identifier les peptides porteurs de mutations et de délétions d'acides aminés. La recherche dans les bases de données permet de lever des ambiguïtés liées à la fragmentation des peptides et à leur séquençage. On a vu qu'il est impossible dans la plupart des cas de discriminer entre une leucine et une isoleucine. D'autre part, il arrive fréquemment qu'il y ait des clivages manquants. Lorsqu'ils sont nombreux, le nombre de séquences possibles est trop important. Lorsqu'il est limité à 1, 2 voire 3, le programme de séquençage *de novo* propose souvent plusieurs séquences équiprobables. Pevzner et son équipe ont introduit le concept de peptides homéométriques. Ces peptides ont des séquences différentes mais des spectres

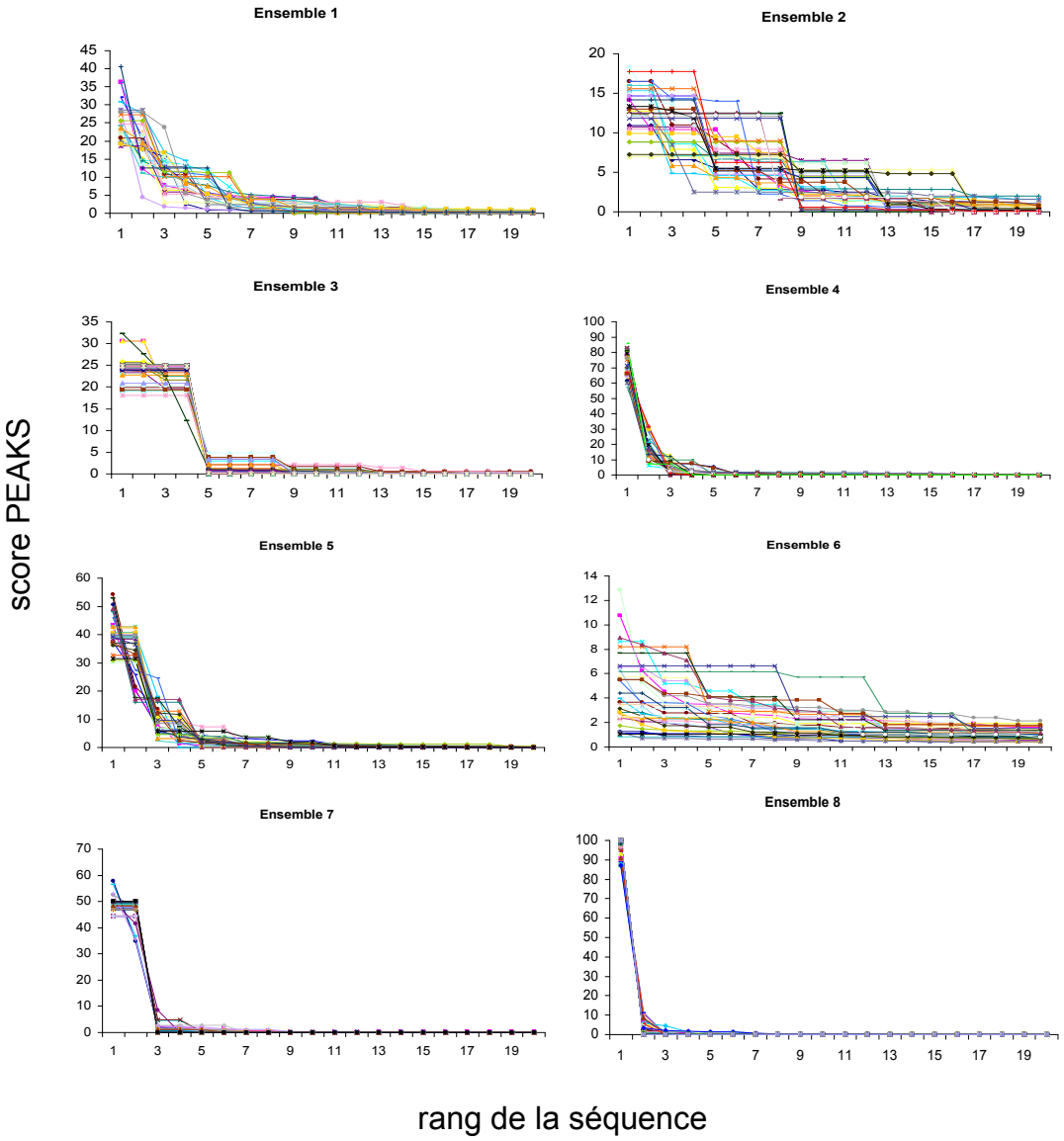


Figure 4.5 – Classification des spectres en fonction de la distribution des scores des séquences *de novo* retournées par Peaks

similaires, voire identiques [5]. Le spectre obtenu ne permet pas de déterminer lequel des peptides possibles a réellement été fragmenté. Bien que la haute précision en masse diminue de façon significative le nombre de peptides homéométriques (réduction d'un facteur 20 entre une précision de 0,5 Da et 0,0075 Da [5]), la probabilité de peptides homéomé-

triques reste non négligeable (on la détermine par extrapolation à 5% pour une précision en masse de 0,02 Da). Les peptides du CMH-I conduisant à un plus grand nombre de spectres incomplets majorent de cette probabilité. En effet, 11,7% des séquences *de novo* retournées par Peaks avec les peptides synthétiques du CMH-I contiennent une ambiguité sur l'ordre de deux résidus adjacents. Ces constats impliquent que les programmes d'identification utilisant des bases de données ou procédant par séquençage *de novo* ne peuvent rarement fournir qu'une seule solution. Le filtrage des séquences *de novo* par l'utilisation du programme *Blast* (version 2.2.14) permet d'identifier la bonne séquence parmi les solutions équiprobables proposées. Nous avons construit deux bases de données à partir des fichiers FASTA contenant les séquences des protéines d'une part de la souris et d'autre part de l'homme auxquelles ont été ajoutées les séquences des pathogènes d'où proviennent les peptides antigéniques synthétiques. Les bases de données pour les allèles HLA-A01, HLA-A02 et HLA-A03 de l'homme et pour les allèles H2-Db et H2-Kb pour la souris contiennent respectivement 195400 et 45640 séquences peptidiques. En annexe VII se trouve l'ensemble des organismes pour lesquels l'ensemble des protéines fait partie de deux bases de données.

On définit les termes suivants :

1. vrai positif : séquence *de novo* appartenant aux séquences des peptides synthétiques retenue par le filtre ;

2. vrai négatif : séquence *de novo* n'appartenant pas aux séquences des peptides synthétiques rejetée par le filtre ;

3. faux positif : séquence *de novo* n'appartenant pas aux séquences des peptides synthétiques retenue par le filtre ;

4. faux négatif : séquence *de novo* appartenant aux séquences des peptides synthétiques rejetée par le filtre.

Une séquence *de novo* est retenue par le filtre si elle a une identité de 100% avec une séquence peptidique de la base de données utilisée par *Blast* et qu'elle est de même longueur.

Tableau 4.V – Performance du séquençage *de novo* avec Peaks et un filtre basé sur l'utilisation de Blast.

	Taux
Vrais positifs (VP)	51,3 %
Vrais négatifs (VN)	42 %
Faux positifs (FP)	2,1 %
Faux négatifs (FN)	4,5 %

On dessine la courbe ROC qui représente le taux de vrais positifs (équation 4.4) (ou la sensibilité) en fonction du taux de faux positifs (équation 4.5) (ou 1 - spécificité) en faisant varier le seuil d'identité entre la séquence *de novo* et la plus proche séquence peptidique trouvée dans la base de données. La courbe ROC de l'association **Peaks-Filtre basé sur Blast** met évidence que l'on peut obtenir simultanément des valeurs élevées de sensibilité et de spécificité. La mesure de l'identité a donc un fort pouvoir discriminateur entre vrais positifs et faux positifs. La comparaison des courbes ROC (figure 4.6) montre également que l'identité est un meilleur discriminateur que le score Peaks.

$$Taux\ de\ vrais\ positifs\ (TVP) = \frac{VP}{VP+FN} \tag{4.4}$$

$$Taux\ de\ faux\ positifs\ (TFP) = \frac{FP}{FP+VN} \tag{4.5}$$

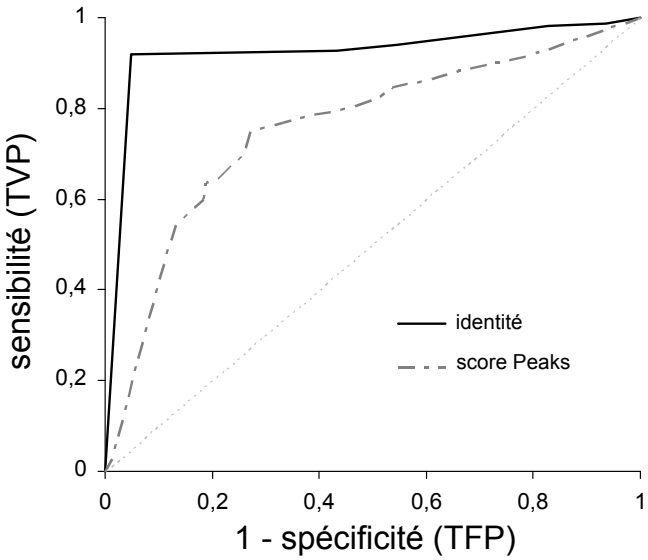


Figure 4.6 – Courbe ROC pour l'association Peaks-Filtre Blast. L'identité est un meilleur discriminateur que le score Peaks.

4.6 Conclusion

L'analyse comparative des 5 programmes **Lutefisk, pepNovo, Peaks, pNovo** et **Vonode** démontre sans équivoque les meilleures performances de Peaks avec les peptides du CMH-I. Toutefois, le programme PepNovo montre d'aussi bons résultats que Peaks avec des spectres CID de peptides trypsiques (données non montrées). Avec les peptides du CMH-I en mode CID, il fonctionne relativement bien aussi. Mais l'utilisation conjointe de la fragmentation en mode HCD et de Peaks fournit clairement des résultats les plus intéressants. On peut y voir au moins deux explications. Peaks est moins affecté par l'absence de clivage entre deux résidus, plus fréquents avec les peptides du CMH-I. En outre, Peaks tire profit de la meilleure précision en masse en diminuant significativement le nombre de séquences possibles pour une masse de précurseur donné. Le calcul du score Peaks utilise notamment les informations relatives aux ions immoniums et aux fragments internes nettement plus abondants dans les spectres HCD. Cependant,

la composition des peptides du CMH-I conduit souvent en l'absence de clivage le long de la chaîne peptidique et donc à des séquences candidates équiprobables. Il faut donc tenir compte des k premières solutions ($k > 5$) à moins que la première séquence *de novo* proposée ait un score supérieur à 95. L'utilisation d'un filtre basé sur le programme *Blast* permet dans la plupart des cas d'identifier la bonne séquence *de novo* parmi les séquences candidates. On montre dans le prochain chapitre une application de l'utilisation de programme Peaks pour la recherche d'antigènes mineures d'histocompatibilité.

CHAPITRE 5

LA RECHERCHE D'ANTIGÈNES MINEURS D'HISTOCOMPATIBILITÉ

La stratégie thérapeutique indiquée dans le traitement de la leucémie consiste à détruire les cellules tumorales, le système sanguin ainsi que le système immunitaire du patient et remplacer ce dernier par celui d'un donneur compatible HLA. Les cellules néoplastiques du patient sont la cible du système immunitaire nouvellement transplanté. A lieu alors ce qu'on appelle la réaction du greffon contre la tumeur (GVL). Celle-ci est bénéfique car elle vise les cellules indésirables. Malheureusement il n'est pas rare qu'une réaction du greffon contre l'hôte (GVH) ait lieu, même en cas de compatibilité HLA parfaite. Il s'avère en effet que des locus situés en dehors du CMH-I peuvent en être la cause. Il s'agit des antigènes mineurs d'histocompatibilité (AgMH) qui sont des peptides de protéines cellulaires polymorphes fixés aux molécules du CMH-I. Même sein d'une fratrie, il existe des antigènes mineurs entre deux individus, à moins qu'il ne s'agisse de jumeaux monozygotes. Un grand nombre de protéines sont différentes à cause du polymorphisme génétique. Il s'agit des SNPs (Single Nucleotide Polymorphism) qui correspondent à des variations stables d'une seule paire de bases du génome présentes dans au moins 1% de la population. Ils constituent la source génétique principale de la variabilité phénotypique qui distingue les individus les uns des autres. Il y a plusieurs millions de positions de nucléotide dans le génome humain pour lesquelles il peut y avoir une variation (toutes les 100 à 300 bases dans un génome qui compte 3 milliards de base). À la fin de l'année 2010, la base de données dbSNP (dbSNP 132) compte plus de 12 millions de SNPs identifiés pour l'humain [2]. La plupart des SNPs n'implique pas de modifications fonctionnelles. Les SNPs peuvent se trouver dans des

régions codantes (exons) ou des régions non-codantes (introns). Les SNPs dans les régions codantes peuvent conduire à une modification dans la chaîne peptidique de la protéine (SNP non-synonyme) ou ne pas affecter cette dernière (SNP synonyme). La figure 5.1 illustre les deux types de SNP dans les régions codantes.

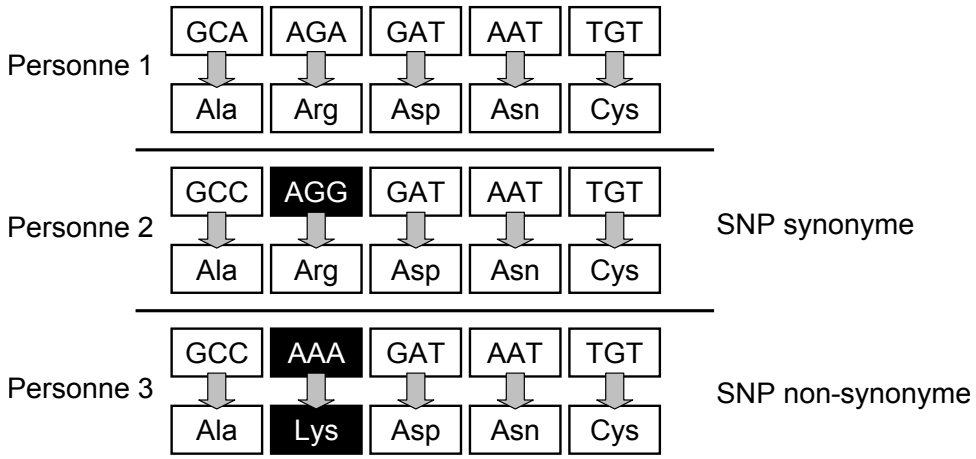


Figure 5.1 – SNPs synonymes et SNPs non-synonymes. La variation de la séquence ADN peut ne conduire à aucun changement de la séquence peptidique (personne 2). On parle alors de SNPs synonymes. Dans le cas contraire, on parle de SNPs non-synonymes (personne 3).

Les AgMH résultent de la présence de SNPs non-synonymes. La découverte d'AgMHs potentiellement responsables de la réaction du greffon contre l'hôte pourrait permettre d'induire un mécanisme de tolérance envers ceux-ci. On se propose donc de développer une chaîne de traitement de recherche de AgMHs. Le séquençage *de novo* est tout indiqué pour l'identification de mutations et notamment de SNPs. Comme on l'a vu dans les chapitres précédents, le mode de fragmentation HCD offrant une grande précision en masse et conduisant à des spectres plus informatifs laisse croire à la faisabilité d'une stratégie basée sur le séquençage *de novo* pour la découverte de AgMH.

5.1 L'approche expérimentale

L'objectif est d'identifier des AgMHs en comparant le répertoire de peptides du CMH-I de deux individus HLA-identiques et de même sexe (car il pourrait y avoir des peptides différents provenant de protéines codées sur le chromosome Y si nous avions deux personnes de sexe différent). Pour ce faire, on isole des cellules B du sang de deux donneurs d'une même fratrie (désignés par M et R). Les cellules sont transformées *in vitro* avec le virus Epstein-Barr. Ces dernières, en suspension, présentent plusieurs avantages pour leur culture et l'extraction des peptides. Un élément intéressant notamment est qu'elles expriment davantage de complexes d'histocompatibilité à la surface cellulaire. Les peptides sont élués des cellules vivantes avec une solution acide douce [24]. Le principe est d'éviter la lyse des cellules afin que les préparations peptidiques contiennent le moins possible de contaminants. De plus, ces peptides sont présentés en un petit nombre de copies à la surface des cellules. Il importe donc d'avoir un bon rendement d'extraction. Ensuite, les extraits sont fractionnés par chromatographie en phase liquide à haute performance (CLHP ou plus fréquemment HPLC) avant d'être analysés par le spectromètre LTQ-Orbitrap Velos 5.3. On obtient pour chacun des individus M et R respectivement 42234 et 30915 spectres MS/MS. Le chaîne de traitement de recherche d'AgMHs prend en entrée les deux ensembles de spectres. Les cellules de M et R expriment les protéines HLA-A*0301, HLA-A*2902, HLA-B*0801, HLA-B*4403, HLA-C*0701 et HLA-C*1601 (identifiées par une technique de séquençage de HLA à haut résolution à l'hôpital Maisonneuve de Montréal par Marie-Christine Meunier, responsable du laboratoire de typage HLA).

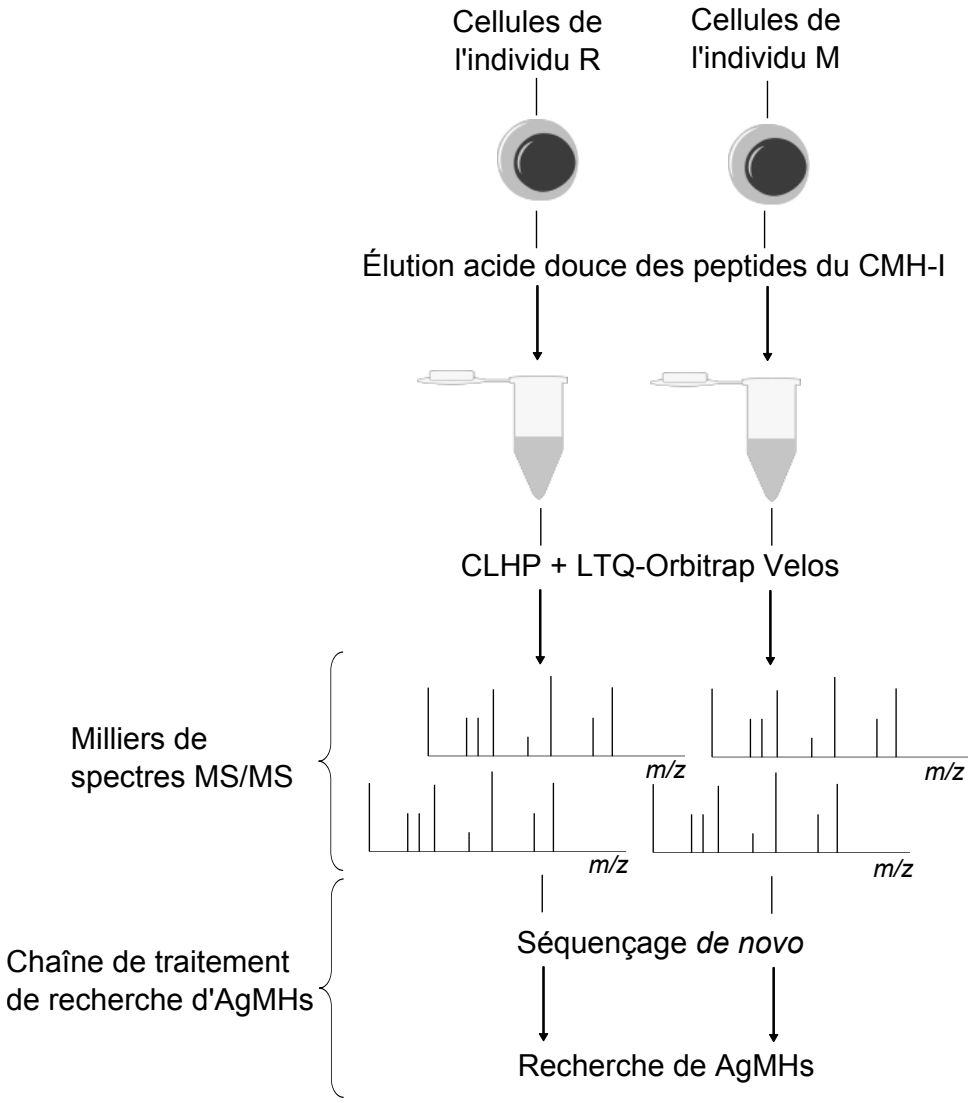


Figure 5.2 – L'approche expérimentale pour la recherche d'AgMHs.

5.2 La chaîne de traitement

La chaîne de traitement compte plusieurs étapes. La première étape consiste à procéder à la recherche MASCOT à partir des spectres de l'individu M d'une part et de l'individu R d'autre part. On utilise la base de données IPI de l'humain. Les identifications sont retenues si le score associé est supérieur ou égal à 25. On met de côté provisoirement

ces peptides identifiés qui seront utilisés dans une étape ultérieure. Dans une deuxième étape, on filtre les spectres non assignés par MASCOT, ou ayant un score inférieur à 25, pour ne conserver que ceux qui sont suffisamment informatifs. Pour ce faire, on utilise le programme MASCOT Filter qui a été développé (par moi-même) spécifiquement pour cette chaîne de traitement. Les spectres non assignés par MASCOT correspondent potentiellement à des antigènes mineurs d'histocompatibilité. En effet, les peptides porteurs de mutations ne peuvent être identifiés à partir de la base de données IPI de l'humain qui ne comporte pas les SNPs possibles. Par conséquent, dans une troisième étape, on utilise le programme de séquençage *de novo* PEAKS pour séquencer les spectres non assignés qui ont été retenus par le filtre MASCOT Filter. On conserve les k premières séquences *de novo* ($k = 5$). On se retrouve avec un ensemble de séquences *de novo* pour l'individu M et un autre pour l'individu R. L'idée est de comparer les séquences *de novo* de l'individu M avec les séquences trouvées par MASCOT pour l'individu R et *vice versa*. De cette manière, on peut espérer trouver des peptides porteurs de mutations provenant d'un individu pour lesquels l'autre individu a le type non muté. L'annexe VIII montre une vue schématique de la chaîne de traitement.

5.2.1 La recherche MASCOT

La recherche MASCOT a pour objet d'identifier les peptides non mutés. On n'utilise la base de donnée IPI de l'humain (version 3.54) qui contient 69169 séquences de protéines. La tolérance sur le précurseur est fixée à 10 ppm alors que celle sur les fragments est égale à 0,02 Da. Les modifications chimiques variables sont l'oxydation de la méthionine et la déamination pour les acides aminés N et Q. Ces modifications doivent être prises en compte car elles peuvent avoir lieu naturellement ou être induites par l'isolation des peptides. Le seuil du score MASCOT est fixé à 25. On identifie un total de 1601

peptides différents pour l'individu M et 1427 pour l'individu R. Un total de 994 peptides sont partagés par M et R. Un ensemble de 668 peptides est exclusif à M et un de 471 peptides l'est à R. On détermine pour chacun des peptides la protéine du HLA à laquelle il a la plus grande probabilité de se lier. Pour ce faire, on utilise un script faisant appel au programme NetMHCpan (version 2.0) [56].

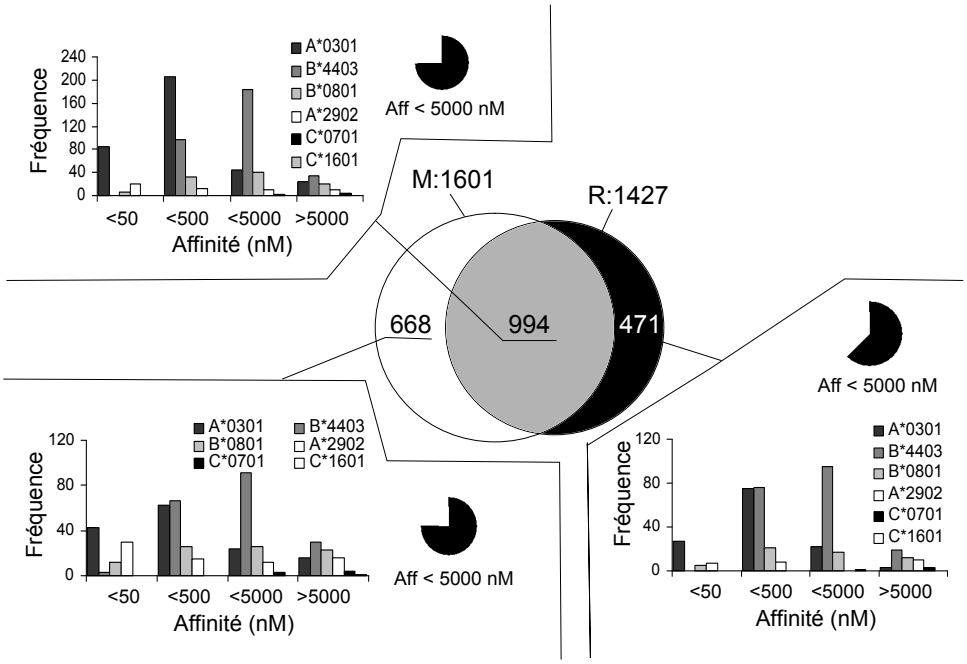


Figure 5.3 – Statistiques relatives à la recherche MASCOT sur les spectres MS/MS pour M et R. Il y a 994 peptides identifiés partagés par les individus M et R. Un total de 668 peptides sont exclusifs à M alors que 471 le sont à R. Pour chacun des sous-ensembles, on a la distribution du nombre de peptides associés à chaque allèle du HLA en fonction de l'affinité (Plus la valeur en nM est faible, plus l'affinité est élevée). Très peu de peptides sont associés aux protéines de la sous-classe HLA-C. Les trois secteurs en noir indique la proportion de peptides ayant une affinité forte, moyenne ou faible.

La taille de la plupart des peptides du CMH-I est comprise entre 8 et 11 acides aminés. Cependant, il est connu que des peptides de taille non consensuelle peuvent se lier aux molécules du CMH-I [9]. Parmi les 994 peptides partagés par M et R, il y a 157 d'entre eux qui ont une taille inférieure ou supérieure à la taille consensuelle. L'analyse de la composition de ces peptides nous révèle que 34 d'entre eux ont un motif qui cor-

respond à l'allèle HLA-B*4403. En effet, 17 ont un acide aminé Glu (E) en position 2 (P2) et un acide aminé Tyr en position C-terminale (PC). Il a été montré qu'un sous-ensemble d'acides aminés (Leu, Ala, Arg, Lys, His et Phe) peut être toléré en PC [19]. Le tableau 5.I donne le nombre de peptides pour les différentes longueurs observées. Il y a une large proportion de peptides de 12 résidus. Il a été démontré que la partie centrale d'un peptide de 14 résidus se liant à la molécule HLA-B*3501 forme un bombement en dehors du sillon [62]. En effet, le sillon dans lequel se le peptide est fermé à chacune des extrémités. Il est plausible qu'il y ait également un bombement entre les résidus situés en P2 et PC. De plus, on calcule que la probabilité d'avoir un peptide avec un E en P2, et un Y en PC est inférieure à 0,003. Si on tolère un L, A, K, R, H ou F en PC, la probabilité est inférieure à 0,018. Autrement dit, il est loin d'être impossible qu'au moins une partie de ces peptides soient réellement des peptides du CMH-I associé à l'allèle HLA-B*4403. Nous pourrions faire la même analyse avec les autres allèles. Ceci illustre l'importance d'utiliser les programmes de prédiction avec précaution. Leur usage comme filtre est à proscrire dans certaines circonstances.

Tableau 5.I – Proportion des peptides de taille non consensuelle ayant le motif correspondant à l'allèle HLA-B*4403

nombre de résidus	nombre de peptides
6	1
7	3
12	19
13	3
14	2
15	2
16	2
17	1
21	1

5.2.2 Le séquençage *de novo*

Un total de 37931 spectres pour M et un total de 28500 spectres pour R n'ont pas été assignés. Ces spectres sont filtrés, et analysés par le programme PEAKS pour identifier des peptides porteurs de mutations.

5.2.2.1 Le programme MASCOT Filter

Une proportion des spectres non assignés est de médiocre qualité ou peu informative. Il convient donc de filtrer ces spectres. Le séquençage *de novo* de plusieurs milliers de spectres nécessite de longues heures de calcul et beaucoup de mémoire vive (jusqu'à 1,5 Go avec PEAKS). Il est donc préférable d'éviter de soumettre une grande quantité de spectres de mauvaise qualité ne donnant lieu à aucun séquençage avec une confiance raisonnable. À partir d'un fichier MGF contenant les informations relatives à un ensemble de spectres [1], le programme MASCOT Filter génère deux nouveaux fichiers MGF. L'un contient les spectres MS/MS retenus, répondant aux critères spécifiés : la tolérance de masse pour les fragments, le seuil de bruit et le ratio minimum de résidus. Et l'autre contient les spectres rejetés. Ce dernier fichier peut être utile pour déterminer le taux de spectres rejetés à tort. Le programme prend en entrée 3 paramètres. Le premier correspond à la tolérance de masse pour les fragments. Celle-ci dépend de l'instrument utilisé. Pour le LTQ-Orbitrap Velos que nous utilisons ici, nous pouvons considérer une tolérance pour les fragments de 0,02 Da. Le deuxième paramètre correspond au seuil sur le bruit. Il correspond au pourcentage de la plus grande intensité. Le troisième paramètre est le ratio minimum de résidus (R). Le rejet d'un spectre se base sur ce critère. Un spectre est retenu s'il contient un nombre d'espacements entre pics correspondant à la masse d'un acide aminé ($N_{esp.}$) supérieur à un nombre minimum (N_{min}). Le minimum

d'espacements dépend de la masse du précurseur (M). Plus la masse est importante, plus grand sera le minimum d'espacements à atteindre. Il est évident que pour un peptide de 4 acides aminés, on s'attend à avoir moins de fragments que pour un peptide 20 acides aminés. On définit la valeur de N_{min} par la formule 5.1.

$$N_{esp.} \geq N_{min} = R.\frac{M}{K} \tag{5.1}$$

avec

$$K = \sum_{a=1}^{20} m_a.f_a \simeq 130 \tag{5.2}$$

où m_a est la masse de l'acide aminé a appartenant à l'ensemble des 20 acides aminés et f_a est l'abondance relative de ce même acide aminé [3]. Le tableau 5.II montre, pour différentes valeurs de R et de masses M, le nombre minimal d'espacements correspondant à un acide aminé à détecter pour que le spectre soit retenu.

Tableau 5.II – Nombre minimum d'espacements correspondant à la masse d'un acide aminé requis pour une masse de peptide M et un ratio R.

M \ R	0,2	0,3	0,4	0,5	1
500	1	2	2	2	4
600	1	2	2	3	5
700	2	2	3	3	6
800	2	2	3	4	7
900	2	3	3	4	7
1000	2	3	4	4	8
1100	2	3	4	5	9
1200	2	3	4	5	10
1300	2	3	4	5	10
1400	3	4	5	6	11

Le programme MASCOT Filter a été testé sur un ensemble de 2219 spectres. Chacun

des spectres a été séquencé par PEAKS. On a vérifié que les spectres exclus par MASCOT Filter ne sont effectivement pas séquencés par PEAKS ou qu'ils conduisent à des scores de confiance très bas. On s'est également assuré que MASCOT Filter n'exclut pas ou très peu de spectres conduisant à un séquençage *de novo* présentant de bons scores de confiance. Pour les analyses subséquentes, on a retenu pour R une valeur de 0,3 et un seuil sur le bruit de 0%. Comme le montre la figure 5.4, cette valeur permet d'éliminer près de 30% des spectres conduisant au mieux à des séquences pour lesquelles le score de confiance PEAKS est de 5. Alors qu'aucun spectre conduisant à un score de confiance supérieur à 45 n'est rejeté. Un faible proportion de spectres pour lesquels le score PEAKS est compris entre 5 et 45 est rejetée (< 5%). Pour la recherche de SNPs, il est préférable d'opter pour une valeur de R conservatrice. D'autres étapes dans la chaîne de traitement permettront d'identifier les séquences candidates les plus probables, notamment la prédiction de l'affinité de liaison avec la molécule du CMH-I, la probabilité de mutation et la vérification manuelle du spectre. Le filtrage des spectres donne 29730 spectres pour M et 22338 spectres pour R.

5.2.2.2 Le programme PEAKS

La deuxième étape consiste à séquencer les spectres filtrés par MASCOT Filter. Pour ce faire, on utilise le programme PEAKS. La tolérance sur la masse du précurseur est fixée 10 pmm et celle sur la masse des fragments est à 0,02 Da. Aucune enzyme n'est sélectionnée. Les modifications chimiques variables sont la déamination et l'oxydation. Le nombre maximum de modifications variables est égale à 3. On obtient respectivement 148650 et 111690 séquences *de novo* pour M et R.

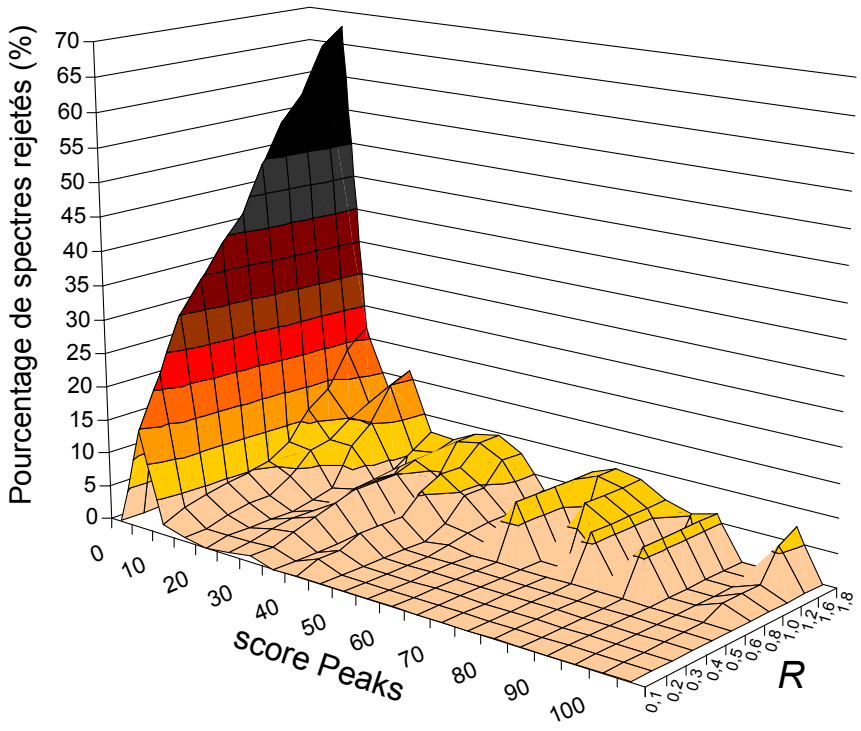


Figure 5.4 – Nombre de spectres rejetés en fonction du score PEAKS et de R

5.2.3 La recherche d'antigènes mineurs d'histocompatibilité

5.2.3.1 Le programme SNPdiscoverer

Le programme SNPdiscoverer, développé par moi-même, prend en entrée deux listes de séquences. Il compare chacune des séquences de la première liste avec chacune des séquences de la deuxième liste. Autrement dit, si la première liste contient n séquences et la deuxième contient m séquences, le programme fait $n \times m$ comparaisons. Le nombre de comparaisons peut donc être très important et nécessiter des heures de calcul. Le programme est basé sur l'algorithme de Needleman-Wunsch pour identifier les mutations dans les séquences peptidiques. Pour réduire le temps de calcul, on ne procède à l'alignement de séquences que pour celles ayant une composition en acides aminés proches.

Le temps de calcul pour la comparaison de composition en acides aminés est plus court que celui nécessaire à l'alignement de séquences. Elle est en linéaire, de complexité en $O(p)$ alors que l'algorithme de Needleman-Wunsch a une complexité en $O(n*m)$ où n et m sont les longueurs de deux séquences. La figure 5.5 montre l'organigramme du programme SNPdiscoverer. Il compare chacun des séquences *de novo* de M avec chacune des séquences de R. Si deux séquences sont similaires en composition, elles sont alignées et l'acide aminé qui diffère est identifié. On désigne par X l'acide aminé chez M et Y celui chez R. On recherche la protéine source possiblement associée à chacune des deux séquences à l'aide du programme Blast. Si elles sont identifiées comme appartenant à la même protéine et à la même position dans cette dernière, on considère que ce couple peut révéler l'existence d'un AgMH. La grande majorité des couples de séquences similaires sont des faux positifs à cause du séquençage *de novo* qui retourne pour un même spectre des séquences similaires. Pour identifier les couples qui ont le plus de chance de résulter réellement d'un SNP, on calcule 3 valeurs. Pour les deux séquences, on détermine la meilleure affinité avec l'une des molécules HLA associés à M et R à l'aide du programme netMHCpan. On calcule également un score de substitution $\mathfrak{S}$ qui reflète la plausibilité de substitution de l'acide aminé X par Y ou Y par X ($\mathfrak{S}(X,Y) = \mathfrak{S}(Y,X)$). Et enfin, on détermine la similarité des spectres associés à chacune des séquences. Les couples les plus prometteurs auront un score de substitution élevé, une similarité de spectres faible et pour chacune des séquences une affinité élevée avec l'une des molécules du HLA.

5.2.3.1.1 La prédiction d'affinité *A priori*, on peut penser qu'un score PEAKS élevé, associé à un peptide, nous indique que le séquençage *de novo* a de forte chance d'être bon et que nous avons réellement affaire à un peptide du CMH-I, si l'on fait abstraction

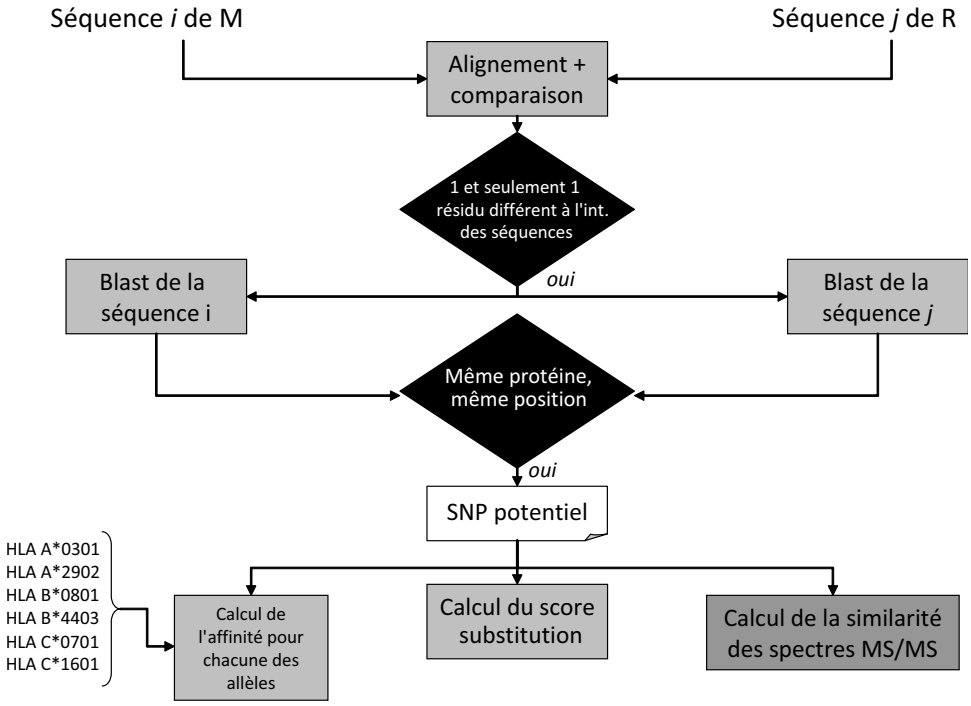


Figure 5.5 – Organigramme du programme SNPdiscoverer

des contaminants. Si tel est le cas, l'affinité prédite pour ce peptide devrait être bonne en considérant bien entendu que le prédicteur ne commet pas d'erreur. Cependant, la corrélation de Spearman entre le score PEAKS et l'affinité pour un ensemble de 3000 séquences est égale à 0,004. Il n'y a donc aucune corrélation. Il y a des séquences associés à des hauts scores PEAKS ayant une affinité prédite faible et *vice versa*. L'information sur l'affinité est donc complémentaire au score PEAKS.

5.2.3.1.2 Le score de substitution Toutes les substitutions d'acides aminés n'ont pas la même probabilité de se produire. Les acides aminés sont codés par des codons constitués de 3 nucléotides. Les substitutions d'acides aminés qui impliquent 2 ou 3 mutations de nucléotide ont une probabilité quasiment nulle d'être observées. En revanche, celles qui ne nécessitent qu'un changement de nucléotide sont beaucoup plus probables.

Le score de substitution $\mathfrak{S}$ se base donc sur le nombre de changements de nuclétoides nécessaires pour la substitution d'un acide aminé par un autre. Soient C_a et C_b les ensembles des triplets de nucléotides codant pour les acides aminés a et b. Par exemple, pour l'acide aminé R, on a $C_R = \{CGU, CGC, CGA, CGG, AGA, AGG\}$. Pour 2 codons c_1 et c_2, on définit la fonction δ comme suit :

$$\delta(c_1, c_2) = \begin{cases} 1 \text{ si seul un nucléotide diffère} \\ 0 \text{ sinon} \end{cases} \tag{5.3}$$

$\mathfrak{S}(a,b)$ se calcule par la formule suivante :

$$\mathfrak{S}(a,b) = \sqrt{\sum_{c_i \in C_a} \sum_{c_j \in C_b} \delta(c_i, c_j)} \tag{5.4}$$

Le score donne une idée de la plausibilité de la substitution. Elle ne reflète en rien la fréquence d'observations des substitutions dans le protéome. Certaines substitutions ont peu d'impacts sur la structure tertiaire de la protéine alors que d'autres provoquent des changements radicaux qui peuvent conduire à une protéine disfonctionnelle. On pourrait donc tenir compte de la nature des acides aminés (acide, basique, aromatique, polaire, etc.) et de la différence en masse. En effet, les substitutions au sein d'un même groupe d'acides aminés partageant des propriétés physiochimiques sont en générales plus conservatrices. Les scores de substitution peuvent être modifiés dans un fichier que le programme SNPdiscoverer lit. Il est donc aisé de procéder à des changements sans modifier le code source du programme. Dans le programme actuel, on ne connait pas les codons réellement associés à chacun des acides aminés. Avec la connaissance de la séquence nucléotidique codant pour la protéine source, on pourrait raffiner le score de substitution basé sur la connaissance du codon réellement associé à l'acide aminé muté.

Pour ce faire, il faudrait modifier le programme pour récuperer la séquence nucléotidique associé au peptide. Cette modification pourrait faire l'objet d'une nouvelle version du programme SNPDiscoverer.

5.2.3.1.3 La similarité spectrale La similarité spectrale nous indique à quel point deux séquences peptidiques conduisent à des spectres semblables. Si les spectres sont semblables, la probabilité que le couple de séquences *de novo* similaires résulte d'un séquençage ambigü est élevée. Autrement dit, on évalue si on est en présence de peptides homéométriques ou pas. En revanche, si la similarité spectrale est faible, il a plus de chance qu'il s'agisse effectivement d'une mutation.

5.2.3.2 Les résultats

La comparaison des séquences retournées par MASCOT pour l'individu M avec celles retournées par PEAKS pour R donne un total de 7312 couples de séquences similaires. Si l'on supprime les doublons, dus à des spectres donnant des séquences *de novo* identiques, on obtient un total de 233 couples de séquences similaires. La même comparaison mais avec les séquences retournées par MASCOT pour R versus les séquences *de novo* pour M donne un total de 5036 couples de séquences similaires. La suppression des doublons nous donne un total de 272 couples de séquences similaires. En annexe IX, on a l'ensemble des résultats pour chacune des substitutions d'acides aminés pour lesquelles la score de substitution est supérieure à 0. En effet, un certain nombre de substitutions trouvées sont improbables car elles impliquent des acides aminés distants (qui nécessitent la mutation de plus d'un nucléotide). On note que parmi les 233 coupes de séquences similaires, il y a 118 qui concernent la substitution d'un acide aminé E pour D et 59 qui concernent une substitution d'un acide aminé D pour N. La

proportion est la même pour l'autre ensemble. Il conviendrait de vérifier si cette proportion élevée n'est pas attribuable à un artéfact. Si l'on exclut ces substitutions, il ne reste que quelques dizaines de couples de séquences. Parmi ceux-ci, la majorité a une affinité prédite faible avec la molécule du CMH-I. Après vérification manuelle des couples de séquences pour lesquelles l'affinité est inférieure à 500 aucune ne peut correspondre à un antigène mineur. Ce résultat illustre la difficulté de trouver des AgMHs.

5.3 Conclusion

L'analyse présentée ici s'est limitée à la comparaison des séquences retournées par MASCOT pour l'un des individus avec celles retournées par PEAKS pour l'autre individu et *vice versa*. Une première tentative avait été faite en comparant l'ensemble de toutes les séquences *de novo* trouvées à partir de tous les spectres des deux individus. Celle-ci avait donné un grand nombre de couples de substitution. Il conviendrait de reprendre les données et de les analyser après avoir éliminé les couples candidats les plus improbables. D'autre part, le programme actuel utilise l'algorithme *consensus* de prédiction d'affinité. La dernière version de netMHCpan n'a pas été intégrée à cause du fait qu'aucune version n'est disponible pour la machine utilisée qui a une architecture x86-64 bits. Cette dernière version intègre un plus grand nombre d'allèles et de plus grande taille [45]. De plus, il faudrait modifier le programme de telle sorte qu'il tienne compte des régions flanquantes du peptide de type sauvage. En effet, la mutation qui conduit à la substitution d'un acide aminé peut conduire à un déplacement du site de clivage du peptide. Bien que les résultats préliminaires obtenus avec cette chaîne de traitement ne soient pas probants pour l'instant, d'autres investigations devraient permettre d'identifier les ajustements nécessaires. Une méthode alternative (non mise en oeuvre) à l'utilisation

du séquençage *de novo* est d'identifier les peptides du type sauvage chez chacun des individus, séparément, à l'aide de MASCOT sur la base de données de l'humain. Ensuite, on génère à partir de ces séquences toutes les séquences possibles portant une et une seule mutation à chacune des positions. On obtient ainsi pour une séquence de n résidus $(n \times 19) + 1$ séquences. L'ensemble des séquences ainsi générées constitue une base de données avec laquelle nous procédons à une nouvelle recherche MASCOT. Nous pouvons aussi étendre les séquences générées aux régions flanquantes. Les résultats trouvés peuvent permettre éventuellement de corroborer ceux trouvés par PEAKS après avoir fait les amendements nécessaires à la chaîne de traitement. Cependant, cette méthode n'est valable que pour les peptides qui sont trouvés à la surface de la cellule malgré la substitution d'un acide aminé. En effet, il n'est pas exclu que la substitution d'un acide aminé, surtout s'il est situé à une position d'ancrage, conduise à la présentation d'un peptide qui n'aurait pas lieu sans celle-ci.

CHAPITRE 6

L'APPLICATION WEB MHCDB

La technologie web est devenue la solution de choix pour donner accès à des données informatisées. C'est pourquoi il a été décidé de mettre en place une base de données accessible par intranet. Dans un premier temps, elle ne contient que les peptides synthétiques dont nous disposons. Néanmoins, sa vocation est de s'enrichir d'un nombre croissant de peptides du CMH-I. Elle est accessible à l'adresse suivante : `mhcdb_prod.thibault.iric.ca`. Elle permet de consulter pour chacun des peptides du CMH-I le spectre acquis comparé au spectre théorique. De plus, celle-ci a vocation à contenir un nombre croissant de spectres de peptides du CMH de classe I issus notamment des études réalisées au sein du laboratoire du Dr Pierre Thibault. Elle a pour objet notamment de répondre au besoin d'une masse critique de données souvent nécessaire au développement d'outils d'analyse.

6.0.1 L'interface utilisateur

L'interface utilisateur se veut être la plus simple possible et intuitive que possible. La page d'accueil qui contient le formulaire de recherche est largement inspiré du site Immune Epitope Database (www.iedb.org). Cette base de données stocke des informations en lien avec celle qui se trouvent dans la base de donnée MHCDB, il est donc apparu logique de conserver une certaine consistance quant à l'aspect visuel. La figure 6.1 montre le formulaire de recherche. Le champ `Linear sequence` permet de spécifier un motif de peptide et de chercher des peptides partageant ce motif de séquence. Un exemple de résultat de recherche se trouve en annexe I. Le tableau 6.I liste des exemples de recherche

par motif. Le champ `Project` permet de limiter la recherche à un projet particulier. La section `Side chain charge` permet de filtrer la recherche par rapport à la charge des peptides. Ensuite, il est possible de spécifier l'organisme source, l'organisme hôte et l'allèle. L'utilisateur peut préciser quels sont les champs qu'il veut afficher ; par défaut ils le sont tous.

Tableau 6.I – Exemples de recherche de peptides par motif

Motif	Résultat de la recherche
`*`	tous les peptides sans distinction
`... .....`	les peptides de 8 résidus
`... ....L`	peptides de 8 résidus se terminant par L
`.*L`	tous les peptides se terminant par L
`L.*`	tous les peptides se commençant par L
`.*K|.*R`	les peptides se terminant par K ou R
`K.*|F.*`	les peptides commençant par K ou F
`... R...`	les peptides de 7 résidus ayant un R en 4ème position

La figure 6.2 montre la page relative au peptide YFISIYSRPK. On peut visualiser le spectre réquisitionné ainsi que les ions identifiés. Le tableau supérieur contient les ions théoriques, et le tableau inférieur contient les pics réels identifiés. Les pics théoriques associés à des pics observés sont surlignés en rouge.

6.1 L'implémentation

L'application web a été développée dans le langage de programmation PHP 5. Il s'agit d'un langage de script libre. L'application a été développée à l'aide de *symfony 2.0* basé sur le paradigme Modèle-Vue-Contrôleur (MVC). Celui-ci facilite et accélère le développement d'applications Internet. Il répond au standard de développement les plus élevés. De plus, il permet une maintenance et une versatilité accrues pour des améliorations ultérieures ou des ajouts de fonctionnalités. MHCDB utilise le système de ges-

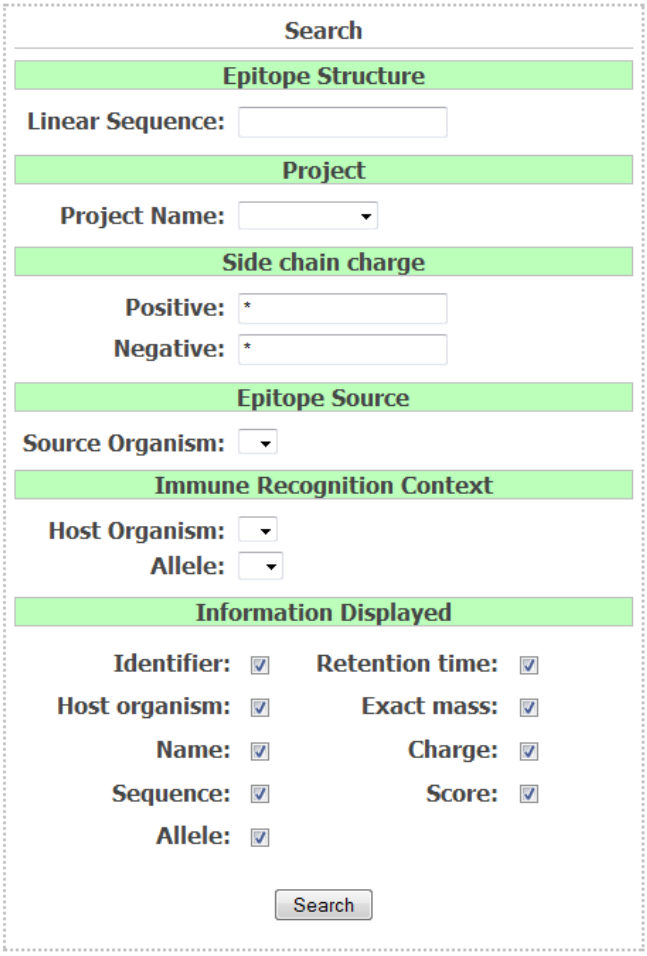


Figure 6.1 – Le formulaire de recherche des peptides de CMH-I

tion de base de données MySQL 5.1, l'un des plus utilisés au monde pour toutes sortes d'applications, y compris les plus complexes et volumineuses. On utilise une interface objet-relationnelle (ORM pour l'anglais object-relational mapping), appelé propel, qui permet de manipuler la base de données à travers la programmation objet. La figure 6.3 montre l'architecture de l'application. La base de données contient les spectres expérimentaux et *in silico* des peptides du CMH-I. L'application proprement dite est composée des éléments suivants : modèle, vue, contrôleur. Le modèle fait l'interface avec la base de données. La vue gère l'affichage des informations à l'écran. Tandis que le contrôleur

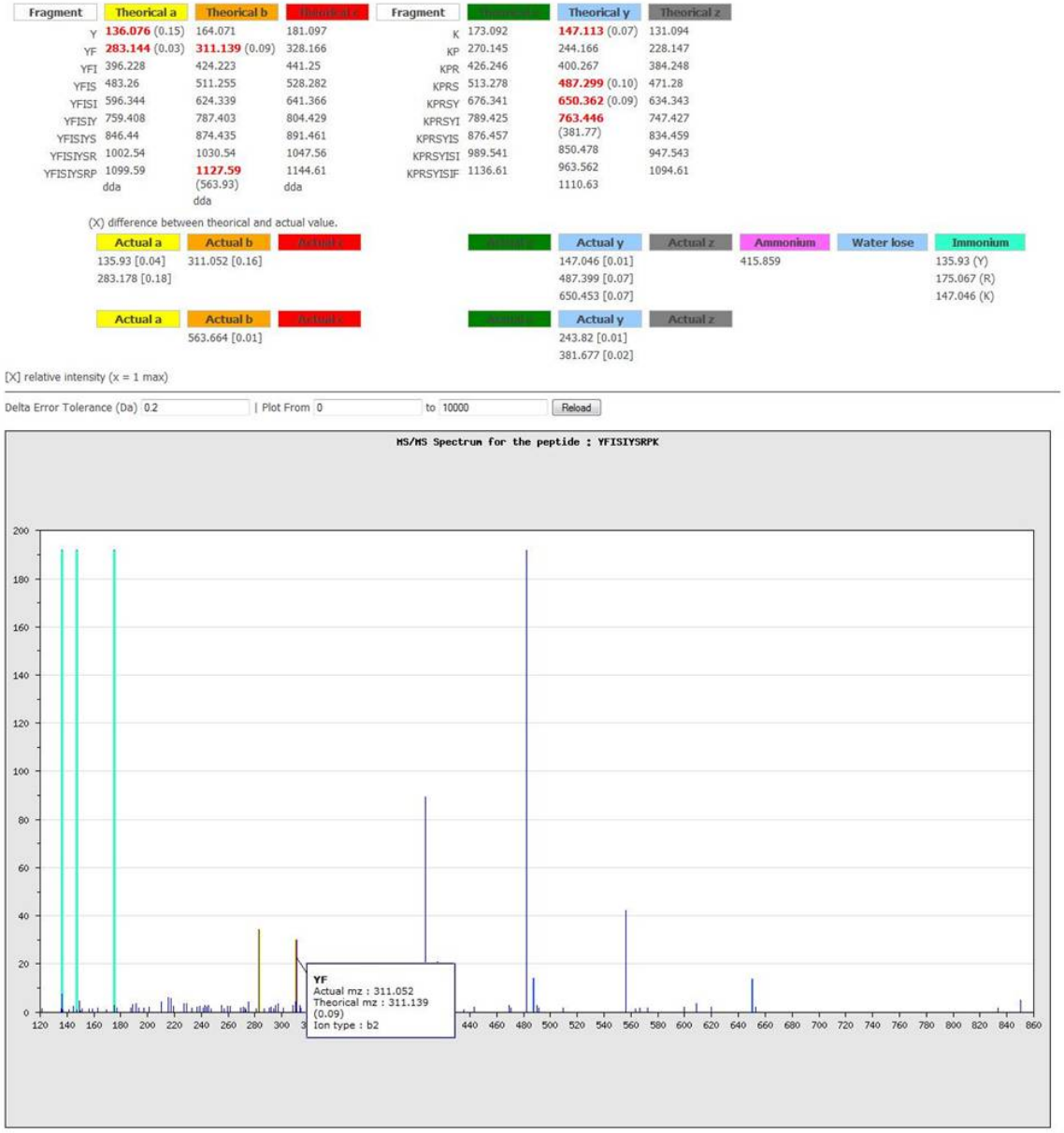


Figure 6.2 – La page d'information du peptide YFISIYSRPK

gère les requêtes de l'utilisateur. L'application est conçue pour permettre d'intégrer un service web pour l'échange des données avec d'autres applications. Des modules pourront être connectés à l'application. Par exemple, un outil d'analyse d'ontologie de gènes

(Gene ontology) qui a été développé mais dont ne parlera pas ici.

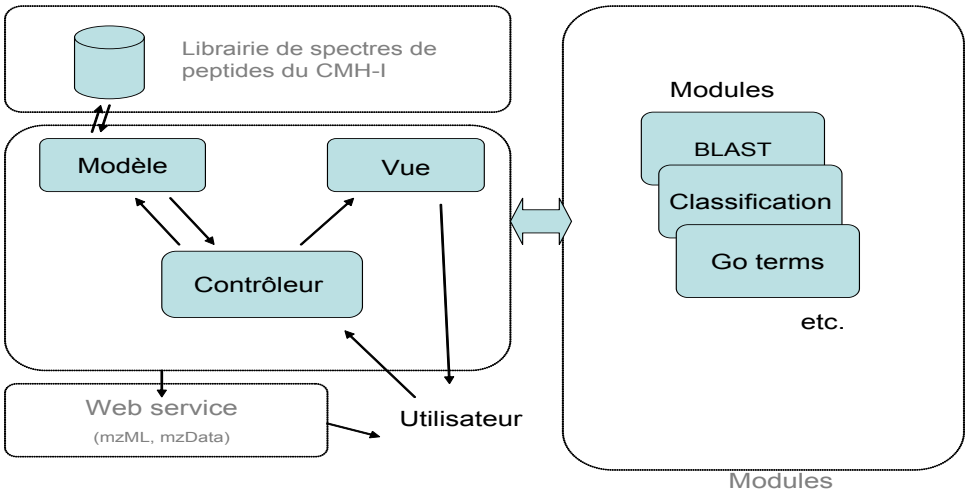


Figure 6.3 – L'architecture générale de l'application MHCDB. L'application utilise le paradigme MVC. Le modèle fait le lien avec la base de donnée, la vue s'occupe de l'affichage des pages web et enfin le contrôleur gère les requêtes de l'utilisateur.

6.1.1 La base de données MySQL

La figure 6.4 montre le diagramme entité-relation de la base de données consacrée au stockage des informations relatives aux peptides du CMH de classe I et des spectres associés. Par soucis de simplicité, les tables liées à la gestion des droits d'utilisateur ne sont pas représentées. La table `project`, comme son nom l'indique, stocke les informations relatives aux projets. Un projet peut être composé de plusieurs analyses dont les informations sont stockées dans la table `analysis`. Une analyse comporte un certain nombre de peptides (`Peptide`) pour lesquels sont associés un spectre constitué de plusieurs pics (`Peak`). Certains pics sont associés à un type d'ion (`ion_type`) et d'autres, non identifiés ne le sont pas. Un pic observé, s'il est identifé, est associé à son pic théorique calculé à partir de la séquence connue du peptide. À chaque peptide sont associés un organisme source identifié par la propriété `source_organism_id`)

et un organisme hôte identifié par `host_organism_id`. On a par exemple pour le peptide **FADINGKLY** comme organisme source le virus du **SARS** et comme l'organisme hôte **l'homme**. Le SARS n'est pas un organisme à part entière mais la base de donnée ne fait pas la distinction entre virus, bactérie, parasite ou mammifère pour ce qui est de l'information relative à l'organisme source. À un peptide peuvent être associés plusieurs scores : Mascot, PEAKS, etc. Ces informations sont stockées dans les tables `score_peptide`, `software` et `score`.

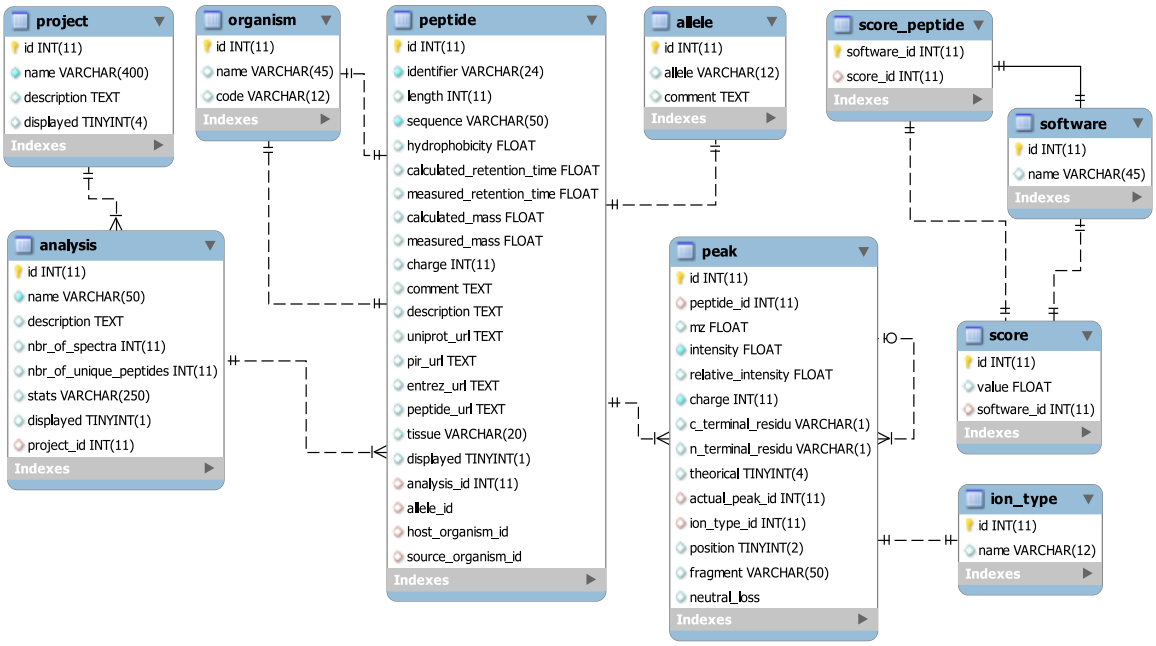


Figure 6.4 – Diagramme entité-association de la base de données de MHCDB.

CHAPITRE 7

DISCUSSION ET CONCLUSION

On a eu l'occasion, notamment dans l'introduction, d'expliquer à quel point les peptides du CMH de classe I revêtent d'une grande importance et ce pour plusieurs raisons. De nombreuses études reposent sur la caractérisation de ces peptides. L'étude de l'impact de l'immunoprotéasome sur la population des peptides du CMH de classe I [18] menée conjointement par les laboratoires du Dr Pierre Thibault et du Dr Claude Perreault est un exemple parmi tant d'autres. Un pré-requis incontournable pour les analyses à grande échelle est l'identification ou le séquençage de ces peptides à partir de spectres de masse en tandem. Il est connu que ces peptides fragmentent moins bien et qu'en conséquence les spectres résultants sont de moins bonne qualité.

La présente étude, basée sur une librairie de plusieurs centaines de peptides synthétiques du CMH de classe I et du programme StatPeaks, a permis de quantifier l'informativité des spectres résultants dans deux modes de fragmentation différents. En mode CID, environ 50% spectres associés aux peptides du CMH de classe I sont complets ou presque complets (1 clivage manquant) alors que cette proportion est de 97% avec les peptides trypsiques. Cette différence explique à elle seule pourquoi les peptides auxquels on s'intéresse représentent un défi en spectrométrie de masse. Par ailleurs, on montre que le profil de fragmentation est différent. À la lumière de cette observation, on peut comprendre pourquoi certains programmes de séquençage *de novo* conçus spécifiquement pour les peptides trypsiques montrent des résultats passablement mauvais avec les peptides du CMH de classe I. Néanmoins, on fait la démonstration que la fragmentation en mode HCD avec le spectromètre de masse LTQ-Orbitrap Velos permet d'obtenir près de

75% de spectres complets ou presque complets. En outre, les spectres HCD contiennent des ions fragments internes et des ions immoniuns en nombre beaucoup plus important qu'en mode CID. À cela s'ajoute le gain en précision de masse. Tout ceci pris ensemble permet de compenser les propriétés intrinsèques des peptides du CMH de classe I défavorables à la fragmentation. On peut donc dire que la fragmentation en mode HCD doit constituer un des éléments clefs d'une nouvelle approche d'identification ou de séquençage des peptides du CMH de classe I.

Dans notre présente étude, nous ne disposions que de peptides associés aux allèles HLA-A*01, HLA-A*02 et HLA-A*03 pour l'humain. Il serait intéressant de mener la même étude pour des peptides associés à des allèles différentes pour lesquels les motifs d'ancrages ne sont pas semblables. En effet, la qualité de la fragmentation est dépendante de la composition des peptides. De fait, elle l'est aussi du motif d'ancrage qui influe sur la composition des peptides. Le programme StatPeaks permet de procéder à des analyses sur n'importe quel type de peptides. Cependant, le facteur limitant pour cette étude est la disponibilité de peptides du CMH de classe I avérés ou prédits. Au chapitre 5 relatif à la recherche d'antigènes mineurs d'histocompatibilité, on identifie une plus grande proportion de peptides associés aux allèles HLA-A*0301 et HLA-B*4403 qu'aux autres. Ceci peut être attribuable, entre autres, à la meilleure fragmentation des peptides associés à ces allèles. Néanmoins, il conviendrait de mener une analyse approfondie pour confirmer ou infirmer cette hypothèse. D'autres facteurs telle que la prédiction d'affinité notamment peut aussi introduire un biais. En effet, des peptides peuvent avoir été classés à tort parmi les peptides associés à ces allèles.

De nombreux algorithmes de séquençage *de novo* sont proposés et régulièrement une nouvelle approche plus ou moins inspirée des précédentes voit le jour. Aucune d'entre elles ne fournit des résultats aussi fiables qu'avec les programmes utilisant les bases

de données. Cependant, il y a des cas où le séquençage *de novo* s'avère très utile. La recherche d'antigène mineurs d'histocompabilité est un exemple. PEAKS montre sans équivoque des performances nettement meilleures que celles de 4 de ses compétiteurs. Le programme PEAKS est moins affecté, de part son fonctionnement, par l'absence de clivages et compense ceux-ci par les informations supplémentaires fournies par la fragmentation en mode HCD. Il faut noter que notre analyse ne tient pas compte de la diversité des motifs d'ancrage associés aux peptides du CMH-I. Il faudrait reproduire cette analyse pour chacun des différents super-types de molécules HLA. Par exemple, les peptides se liant aux molécules de type HLA-A31 se terminant majoritairement par un acide aminé Lys ou Arg ne présentent, à n'en point douter, un profil de fragmentation sensiblement différent des peptides associés aux molécules de type HLA-B27 pour lesquels un acide aminé Arg constitue un résidu d'ancrage en position 2 [42]. Il serait aussi pertinent d'adapter le programme pepNovo, qui montre des performances supérieures à PEAKS en mode CID, à la fragmentation en mode HCD avec les peptides du CMH de classe I et ce pour chacun des 9 super-types. Pour ce faire, il faudrait néanmoins disposer de suffisamment de spectres pour chaque super-type afin de constituer des ensembles d'apprentissage et de validation de tailles suffisantes. Il serait envisageable d'amender le calcul du score pour tenir compte des fragments internes et des ions immoniums. Aucun programme de séquençage *de novo* ne tient compte de la nature spécifique des peptides du CMH de classe I. Le programme PEAKS peut fournir pour un spectre expérimental donné un certain nombre de séquences candidates dont plusieurs sont équiprobables. Il arrive que ce nombre comporte deux chiffres. La connaissance de l'allèle associé aux peptides que l'on cherche à séquencer peut permettre de discriminer entre différentes séquences candidates. Autrement dit, on peut compenser le fait que la composition des peptides du CMH de classe I soit défavorable à la fragmentation par la connaissance des

motifs d'ancrage qui influent sur cette composition. Une solution relativement simple et qui permet de bénéficier des performances du programme PEAKS serait d'adjoindre un post-traitement filtrant les candidats proposées en tenant compte des motifs des allèles connus pour les peptides recherchés. Ce qui semble être un désavantage peut (peut-être) devenir un avantage. En somme, il y a place à l'amélioration des algorithmes de séquençage *de novo* pour les peptides du CMH de classe I car peu d'initiatives ont été menées dans ce domaine.

L'identification des peptides du CMH de classe I par spectrométrie de masse, et *a fortiori* le séquençage *de novo*, nécessite l'optimisation de chacune des étapes allant de l'extraction des peptides à l'analyse bioinformatique des données recueillies en passant par la séparation par chromatographie. Notre laboratoire s'est récemment penché sur l'optimisation de l'extraction des peptides du CMH de classe I à l'aide d'un tampon acide citrate-phosphate qui libère les peptides par la dénaturation des complexes CMH-I et l'utilisation d'un système nanoLC-MS (nano Liquid Chromatography Mass Spectrometry) pour la séparation permettant la détection de peptides CMH-I à partir d'une faible quantité de cellules (inférieure à 5 millions de cellules) [24]. La présente étude visait à déterminer la meilleure stratégie bioinformatique pour l'analyse des données spectrales afin de compléter la chaîne de traitement de séquençage et d'identification des peptides du CMH de classe I. Le séquençage *de novo* nécessite d'une part la meilleure qualité de données possible et un algorithme de traitement des données performant. Le protocole expérimental mis en place dans notre laboratoire et l'utilisation du spectrométre LTQ-Orbitrap Velos permet de générer des données d'une qualité inégalée à ce jour et particulièrement adaptées au séquençage *de novo*. La présente étude a permis d'identifier le programme PEAKS comme le plus performant pour les peptides du CMH de classe I. Comme expliqué au chapitre 5, un projet collaboratif des laboratoires du Dr Pierre

Thibault et Dr Claude Perreault, mettant à profit la chaîne de traitement ainsi constituée, a pour objectif d'identifier des antigènes mineurs d'histocompatibilité (AgMH) potentiellement responsables de la maladie contre l'hôte à la suite d'une greffe dans le cadre du traitement de la leucémie. Toute molécule peut générer des AgMH si celle-ci contient une région polymorphique qui passe à travers le système d'apprêtement des antigènes et correspond au motif d'une molécule du CMH de classe I. Par conséquent, toute variation génomique, en particulier les SNPs (Single Nucleotide Polymorphisms), peut potentiellement produire des antigènes mineurs d'histocompatibilité. En raison de l'absence d'une approche à haut débit pour la découverte d'AgMH, pas plus de 40 AgMH chez l'homme ont été identifiés à ce jour [27]. L'objectif du projet collaboratif dans lequel s'intègre cette présente étude vise à évaluer pour la première fois l'impact du polymorphisme génétique sur le répertoire des peptides du CMH de classe I et l'identification d'antigènes mineurs d'histocompatibilité par spectrométrie de masse. L'approche exposée dans le chapitre 5 utilisant le programme SNPDiscoverer a permis de constituer une liste des peptides du CMH de classe I qui pourraient s'avérer être des antigènes mineurs d'histocompatibilité. Les travaux sont toujours en cours. Une validation immunologique des antigènes mineurs d'histocompatibilité potentiels devra être menée pour confirmer leur existence. La découverte d'antigènes mineurs d'histocompatibilité par la démarche proposée ici constituerait une première. Le projet fera l'objet d'une publication en 2011 ou 2012.

BIBLIOGRAPHIE

[1] Mascot - data file format, 12 2010. URL http://www.matrixscience.com/help/data_file_help.html.

[2] Ncbi single nucleotide polymorphism release announcement (11/09/2010), 12 2010. URL http://www.ncbi.nlm.nih.gov/SNP/snp_summary.cgi.

[3] Swiss-prot database, 12 2010. URL http://ca.expasy.org/sprot/.

[4] Zeev B Alfassi. On the normalization of a mass spectrum for comparison of two spectra. *J Am Soc Mass Spectrom*, 15(3):385–7, Mar 2004.

[5] Frank AM. De novo peptide sequencing and identification with precision mass spectrometry. *J Proteome Res*, 6(1):114–23, Jan 2007.

[6] W Baumeister, J Walz, F Zühl et E Seemüller. The proteasome : paradigm of a self-compartmentalizing protease. *Cell*, 92(3):367–80, Feb 1998.

[7] Arnaud G Blouin, David R Greenwood, Ramesh R Chavan, Michael N Pearson, Gerard R G Clover, Robin M MacDiarmid et Daniel Cohen. A generic method to identify plant viruses by high-resolution tandem mass spectrometry of their coat proteins. *J Virol Methods*, 163(1):49–56, Jan 2010.

[8] Anne Burgevin, Loredana Saveanu, Yohan Kim, Emilie Barilleau, Maya Kotturi, Alessandro Sette, Peter van Endert et Bjoern Peters. A detailed analysis of the murine tap transporter substrate specificity. *PLoS One*, 3(6):e2402, 2008.

[9] Jacqueline M. Burrows, Melissa J. Bell, Rebekah Brennan, John J. Miles, Rajiv Khanna et Scott R. Burrows. Preferential binding of unusually long peptides to mhc class i and its influence on the selection of target peptides for t cell recognition. *Molecular Immunology*, 45(6):1818 – 1824, 2008. ISSN 0161-5890.

[10] N B Cech et C G Enke. Practical implications of some recent studies in electrospray ionization fundamentals. *Mass Spectrom Rev*, 20(6):362–87, 2001.

[11] Hao Chi. sdnjch@gmail.com. 2010-08-31.

[12] Hao Chi, Rui-Xiang Sun, Bing Yang, Chun-Qing Song, Le-Heng Wang, Chao Liu, Yan Fu, Zuo-Fei Yuan, Hai-Peng Wang, Si-Min He et Meng-Qiu Dong. pnovo : de novo peptide sequencing and identification using hcd spectra. *J Proteome Res*, 9 (5):2713–24, May 2010.

[13] A Ciechanover. The ubiquitin-proteasome pathway : on protein death and cell life. *EMBO J*, 17(24):7151–60, Dec 1998.

[14] Kevin R Coombes, Spiridon Tsavachidis, Jeffrey S Morris, Keith A Baggerly, Mien-Chie Hung et Henry M Kuerer. Improved peak detection and quantification of mass spectrometry data acquired from surface-enhanced laser desorption and ionization by denoising spectra with the undecimated discrete wavelet transform. *Proteomics*, 5(16):4107–17, Nov 2005.

[15] Robertson Craig et Ronald C Beavis. A method for reducing the time required to match protein sequences with tandem mass spectra. *Rapid Commun Mass Spectrom*, 17(20):2310–6, 2003.

[16] P Cresswell, N Bangia, T Dick et G Diedrich. The nature of the mhc class i peptide loading complex. *Immunol Rev*, 172:21–8, Dec 1999.

[17] J Cui, L Y Han, H H Lin, H L Zhang, Z Q Tang, C J Zheng, Z W Cao et Y Z Chen. Prediction of mhc-binding peptides of flexible lengths from sequence-derived structural and physicochemical properties. *Mol Immunol*, 44(5):866–77, Feb 2007.

[18] Danielle de Verteuil, Tara L Muratore-Schroeder, Diana P Granados, Marie-Hélène Fortier, Marie-Pierre Hardy, Alexandre Bramoullé, Etienne Caron, Krystel Vincent, Sylvie Mader, Sébastien Lemieux, Pierre Thibault et Claude Perreault. Deletion of immunoproteasome subunits imprints on the transcriptome and has a broad impact on peptides presented by major histocompatibility complex i molecules. *Mol Cell Proteomics*, 9(9):2034–47, Sep 2010.

[19] M DiBrino, K C Parker, D H Margulies, J Shiloach, R V Turner, W E Biddison et J E Coligan. Identification of the peptide binding motif for hla-b44, one of the most common hla-b alleles in the caucasian population. *Biochemistry*, 34(32):10130–8, Aug 1995.

[20] A R Dongré, A Somogyi et V H Wysocki. Surface-induced dissociation : an effective tool to probe structure, energetics and fragmentation mechanisms of protonated peptides. *J Mass Spectrom*, 31(4):339–50, Apr 1996.

[21] EMBL-EBI. Imgt/hla database, Dec 2010. URL `http://www.ebi.ac.uk/imgt/hla/blast.html`.

[22] Jimmy K. Eng, Ashley L. McCormack et John R. Yates III. An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein

database. *Journal of the American Society for Mass Spectrometry*, 5(11):976 – 989, 1994. ISSN 1044-0305.

[23] Bernd Fischer, Volker Roth, Franz Roos, Jonas Grossmann, Sacha Baginsky, Peter Widmayer, Wilhelm Gruissem et Joachim M Buhmann. Novohmm : a hidden markov model for de novo peptide sequencing. *Anal Chem*, 77(22):7265–73, Nov 2005.

[24] Marie-Hélène Fortier, Etienne Caron, Marie-Pierre Hardy, Grégory Voisin, Sébastien Lemieux, Claude Perreault et Pierre Thibault. The mhc class i peptide repertoire is molded by the transcriptome. *J Exp Med*, 205(3):595–610, Mar 2008.

[25] Ari Frank et Pavel Pevzner. Pepnovo : de novo peptide sequencing via probabilistic network modeling. *Anal Chem*, 77(4):964–73, Feb 2005.

[26] Georgeen Gaza-Bulseco, Biqin Li, Ashley Bulseco et Hong Cheng Liu. Method to differentiate asn deamidation that occurred prior to and during sample preparation of a monoclonal antibody. *Anal Chem*, 80(24):9491–8, Dec 2008.

[27] Els Goulmy. Minor histocompatibility antigens : from transplantation problems to therapy of cancer. *Hum Immunol*, 67(6):433–8, Jun 2006.

[28] Gianna Elena Hammer, Federico Gonzalez, Marine Champsaur, Dragana Cado et Nilabh Shastri. The aminopeptidase eraap shapes the peptide repertoire displayed by major histocompatibility complex class i molecules. *Nat Immunol*, 7(1):103–12, Jan 2006.

[29] Alex G Harrison. Fragmentation reactions of protonated peptides containing glutamine or glutamic acid. *J Mass Spectrom*, 38(2):174–87, Feb 2003.

[30] Alex G Harrison. To b or not to b : the ongoing saga of peptide b ions. *Mass Spectrom Rev*, 28(4):640–54, 2009.

[31] Ben Herbert, Femia Hopwood, David Oxley, John McCarthy, Matt Laver, J Grinyer, A Goodall, Keith Williams, Annalisa Castagna et Pier Giorgio Righetti. Beta-elimination : an unexpected artefact in proteome analysis. *Proteomics*, 3(6):826–31, Jun 2003.

[32] Edmond de Hoffmann et Vincent Stroobant. *Mass spectrometry : principles and applications*. J. Wiley, Chichester, West Sussex, England, 3rd ed édition, 2007. ISBN 9780470033104.

[33] Qizhi Hu, Robert J Noll, Hongyan Li, Alexander Makarov, Mark Hardman et R Graham Cooks. The orbitrap : a new mass spectrometer. *J Mass Spectrom*, 40(4):430–43, Apr 2005.

[34] Eugene Kapp et Frédéric Schütz. Overview of tandem mass spectrometry (ms/ms) database search algorithms. *Curr Protoc Protein Sci*, Chapter 25:Unit25.2, Aug 2007.

[35] Masahiro Kawahara, Ian A York, Arron Hearn, Diego Farfan et Kenneth L Rock. Analysis of the role of tripeptidyl peptidase ii in mhc class i antigen presentation in vivo. *J Immunol*, 183(10):6069–77, Nov 2009.

[36] P Kearney et P Thibault. Bioinformatics meets proteomics–bridging the gap between mass spectrometry data analysis and cell biology. *J Bioinform Comput Biol*, 1(1):183–200, Apr 2003.

[37] Jainab Khatun, Kevin Ramkissoon et Morgan C Giddings. Fragmentation characteristics of collision-induced dissociation in maldi tof/tof mass spectrometry. *Anal Chem*, 79(8):3032–40, Apr 2007.

[38] A F Kisselev, T N Akopian, K M Woo et A L Goldberg. The sizes of peptides generated from protein by mammalian 26 and 20 s proteasomes. implications for understanding the degradative mechanism and antigen presentation. *J Biol Chem*, 274(6):3363–71, Feb 1999.

[39] Peter M Kloetzel. Generation of major histocompatibility complex class i antigens : functional interplay between proteasomes and tppii. *Nat Immunol*, 5(7):661–9, Jul 2004.

[40] AI Lamond. Molecular biology of the cell, 4th edition. *Nature*, 417(6887):383–383, mai 2002.

[41] J A Loo, C G Edmonds et R D Smith. Tandem mass spectrometry of very large molecules. 2. dissociation of multiply charged proline-containing proteins from electrospray ionization. *Anal Chem*, 65(4):425–38, Feb 1993.

[42] Ole Lund. *Immunological bioinformatics*. MIT Press, Cambridge, Mass., 2005. ISBN 0262122804 (alk. paper).

[43] Ole Lund, Morten Nielsen, Can Kesmir, Anders Gorm Petersen, Claus Lundegaard, Peder Worning, Christina Sylvester-Hvid, Kasper Lamberth, Gustav Røder, Sune Justesen, Søren Buus et Søren Brunak. Definition of supertypes for hla molecules using clustering of specificity matrices. *Immunogenetics*, 55(12):797–810, Mar 2004.

[44] Claus Lundegaard, Kasper Lamberth, Mikkel Harndahl, Søren Buus, Ole Lund et Morten Nielsen. Netmhc-3.0 : accurate web accessible predictions of human, mouse and monkey mhc class i affinities for peptides of length 8-11. *Nucleic Acids Res*, 36(Web Server issue):W509–12, Jul 2008.

[45] Claus Lundegaard, Ole Lund, Søren Buus et Morten Nielsen. Major histocompatibility complex class i binding predictions as a tool in epitope discovery. *Immunology*, 130(3):309–18, Jul 2010.

[46] Claus Lundegaard, Ole Lund et Morten Nielsen. Prediction of epitopes using neural network based methods. *J Immunol Methods*, Oct 2010.

[47] Bin Ma, Kaizhong Zhang, Christopher Hendrie, Chengzhi Liang, Ming Li, Amanda Doherty-Kirby et Gilles Lajoie. Peaks : powerful software for peptide de novo sequencing by tandem mass spectrometry. *Rapid Commun Mass Spectrom*, 17 (20):2337–42, 2003.

[48] Makarov. Electrostatic axially harmonic orbital trapping : a high-performance technique of mass analysis. *Anal Chem*, 72(6):1156–62, Mar 2000.

[49] R Martino, M D Caballero, C Canals, J A Simón, C Solano, A Urbano-Ispízua, J Bargay, C Rayón, A Léon, J Sarrá, J Odriozola, J G Conde, J Sierra, J San Miguel, ALLOPBSCT Subcommittee of the Spanish Group for Haematopoietic Transplantation (GETH) et Group GEL-TAMO. Allogeneic peripheral blood stem cell transplantation with reduced-intensity conditioning : results of a prospective multicentre study. *Br J Haematol*, 115(3):653–9, Dec 2001.

[50] Marie-Christine Meunier, Jean-Sébastien Delisle, Julie Bergeron, Vincent Rineau,

Chantal Baron et Claude Perreault. T cells targeted against a single minor histocompatibility antigen can cure solid tumors. *Nat Med*, 11(11):1222–9, Nov 2005.

[51] Birgit Meusser, Christian Hirsch, Ernst Jarosch et Thomas Sommer. Erad : the long road to destruction. *Nat Cell Biol*, 7(8):766–72, Aug 2005.

[52] F Momburg, J J Neefjes et G J Hämmerling. Peptide selection by mhc-encoded tap transporters. *Curr Opin Immunol*, 6(1):32–7, Feb 1994.

[53] Mora, Van Berkel GJ, Enke, Cole, Martinez-Sanchez et Fenn. Electrochemical processes in electrospray ionization mass spectrometry. *J Mass Spectrom*, 35(8): 939–52, Aug 2000.

[54] S B Needleman et C D Wunsch. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J Mol Biol*, 48(3):443–53, Mar 1970.

[55] Alexey I Nesvizhskii. A survey of computational methods and error rate estimation procedures for peptide and protein identification in shotgun proteomics. *J Proteomics*, 73(11):2092–123, Oct 2010.

[56] Morten Nielsen, Claus Lundegaard, Thomas Blicher, Kasper Lamberth, Mikkel Harndahl, Sune Justesen, Gustav Røder, Bjoern Peters, Alessandro Sette, Ole Lund et Søren Buus. Netmhcpan, a method for quantitative predictions of peptide binding to any hla-a and -b locus protein of known sequence. *PLoS One*, 2(8):e796, 2007.

[57] Jesper V Olsen, Jae C Schwartz, Jens Griep-Raming, Michael L Nielsen, Eugen Damoc, Eduard Denisov, Oliver Lange, Philip Remes, Dennis Taylor, Maurizio Splendore, Eloy R Wouters, Michael Senko, Alexander Makarov, Matthias Mann

et Stevan Horning. A dual pressure linear ion trap orbitrap instrument with very high sequencing speed. *Mol Cell Proteomics*, 8(12):2759–69, Dec 2009.

[58] Chongle Pan, Byung H Park, William H McDonald, Patricia A Carey, Jillian F Banfield, Nathan C VerBerkmoes, Robert L Hettich et Nagiza F Samatova. A high-throughput de novo sequencing approach for shotgun proteomics using high-resolution tandem mass spectrometry. *BMC Bioinformatics*, 11:118, 2010.

[59] K C Parker, M A Bednarek et J E Coligan. Scheme for ranking potential hla-a2 binding peptides based on independent binding of individual peptide side-chains. *J Immunol*, 152(1):163–75, Jan 1994.

[60] D N Perkins, D J Pappin, D M Creasy et J S Cottrell. Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis*, 20(18):3551–67, Dec 1999.

[61] Sergey Pevtsov, Irina Fedulova, Hamid Mirzaei, Charles Buck et Xiang Zhang. Performance evaluation of existing de novo sequencing algorithms. *J Proteome Res*, 5(11):3018–28, Nov 2006.

[62] Michael Probst-Kepper, Hans-Jürgen Hecht, Hanne Herrmann, Viktoria Janke, Frank Ocklenburg, Jürgen Klempnauer, Benoit J van den Eynde et Siegfried Weiss. Conformational restraints and flexibility of 14-meric peptides in complex with hla-b*3501. *J Immunol*, 173(9):5610–6, Nov 2004.

[63] Alberto Pugliese. Peptide-based treatment for autoimmune diseases : learning how to handle a double-edged sword. *J Clin Invest*, 111(9):1280–2, May 2003.

[64] Anthony W Purcell, James McCluskey et Jamie Rossjohn. More than one reason to

rethink the use of peptides in vaccine design. *Nat Rev Drug Discov*, 6(5):404–14, May 2007.

[65] H Rammensee, J Bachmann, N P Emmerich, O A Bachor et S Stevanović. Syfpeithi : database for mhc ligands and peptide motifs. *Immunogenetics*, 50(3-4):213–9, Nov 1999.

[66] Eric Reits, Joost Neijssen, Carla Herberts, Willemien Benckhuijsen, Lennert Janssen, Jan Wouter Drijfhout et Jacques Neefjes. A major role for tppii in trimming proteasomal degradation products for mhc class i antigen presentation. *Immunity*, 20(4):495–506, Apr 2004.

[67] Kenneth L Rock, Ian A York et Alfred L Goldberg. Post-proteasomal antigen processing for major histocompatibility complex class i presentation. *Nat Immunol*, 5(7):670–7, Jul 2004.

[68] Kirsten Roomp, Iris Antes et Thomas Lengauer. Predicting mhc class i epitopes in large datasets. *BMC Bioinformatics*, 11:90, 2010.

[69] Tomo Saric, Claudia I Graef et Alfred L Goldberg. Pathway for degradation of peptides generated by proteasomes : a key role for thimet oligopeptidase and other metallopeptidases. *J Biol Chem*, 279(45):46723–32, Nov 2004.

[70] A Sasaki. Host-parasite coevolution in a multilocus gene-for-gene system. *Proc Biol Sci*, 267(1458):2183–8, Nov 2000.

[71] Glen A Satten, Somnath Datta, Hercules Moura, Adrian R Woolfitt, Maria da G Carvalho, George M Carlone, Barun K De, Antonis Pavlopoulos et John R Barr.

Standardization and denoising algorithms for mass spectra to classify whole-organism bacterial specimens. *Bioinformatics*, 20(17):3128–36, Nov 2004.

[72] Loredana Saveanu, Oliver Carroll, Vivian Lindo, Margarita Del Val, Daniel Lopez, Yves Lepelletier, Fiona Greer, Lutz Schomburg, Doriana Fruci, Gabriele Niedermann et Peter M van Endert. Concerted peptide trimming by human erap1 and erap2 aminopeptidase complexes in the endoplasmic reticulum. *Nat Immunol*, 6 (7):689–97, Jul 2005.

[73] Nilabh Shastri, Susan Schwab et Thomas Serwold. Producing nature's gene-chips : the generation of peptides for display by mhc class i molecules. *Annu Rev Immunol*, 20:463–93, 2002.

[74] C L Silva, F C Portaro, V L Bonato, A C de Camargo et E S Ferro. Thimet oligopeptidase (ec 3.4.24.15), a novel protein on the route of mhc class i antigen presentation. *Biochem Biophys Res Commun*, 255(3):591–5, Feb 1999.

[75] Harpreet Singh-Jasuja, Niels P N Emmerich et Hans-Georg Rammensee. The tübingen approach : identification, selection, and validation of tumor-associated hla peptides for cancer therapy. *Cancer Immunol Immunother*, 53(3):187–95, Mar 2004.

[76] Hanno Steen et Matthias Mann. The abc's (and xyz's) of peptide sequencing. *Nat Rev Mol Cell Biol*, 5(9):699–711, Sep 2004.

[77] David L Tabb, Lori L Smith, Linda A Breci, Vicki H Wysocki, Dayin Lin et John R Yates, 3rd. Statistical characterization of ion trap tandem mass spectra from doubly charged tryptic peptides. *Anal Chem*, 75(5):1155–63, Mar 2003.

[78] J A Taylor et R S Johnson. Implementation and uses of automated de novo peptide sequencing by tandem mass spectrometry. *Anal Chem*, 73(11):2594–604, Jun 2001.

[79] Matthias Trost, Luc English, Sébastien Lemieux, Mathieu Courcelles, Michel Desjardins et Pierre Thibault. The phagosomal proteome in interferon-gamma-activated macrophages. *Immunity*, 30(1):143–54, Jan 2009.

[80] Ramunas M Vabulas et F Ulrich Hartl. Protein synthesis upon acute nutrient restriction relies on proteasome function. *Science*, 310(5756):1960–3, Dec 2005.

[81] Peter M van Endert, Loredana Saveanu, Eric W Hewitt et Paul Lehner. Powering the peptide pump : Tap crosstalk with energetic nucleotides. *Trends Biochem Sci*, 27(9):454–61, Sep 2002.

[82] L Van Kaer, P G Ashton-Rickardt, H L Ploegh et S Tonegawa. Tap1 mutant mice are deficient in antigen presentation, surface class i molecules, and cd4-8+ t cells. *Cell*, 71(7):1205–14, Dec 1992.

[83] Randi Vita, Laura Zarebski, Jason A Greenbaum, Hussein Emami, Ilka Hoof, Nima Salimi, Rohini Damle, Alessandro Sette et Bjoern Peters. The immune epitope database 2.0. *Nucleic Acids Res*, 38(Database issue):D854–62, Jan 2010.

[84] M G von Herrath, B Coon, H Lewicki, H Mazarguil, J E Gairin et M B Oldstone. In vivo treatment with a mhc class i-restricted blocking peptide can prevent virus-induced autoimmune diabetes. *J Immunol*, 161(9):5087–96, Nov 1998.

[85] J W Wells, K Choy, C M Lloyd et A Noble. Suppression of allergic airway inflammation and ige responses by a class i restricted allergen peptide vaccine. *Mucosal Immunol*, 2(1):54–62, Jan 2009.

[86] Nicholas S Wilson et Jose A Villadangos. Regulation of antigen presentation and cross-presentation in the dendritic cell network : facts, hypothesis, and immunological implications. *Adv Immunol*, 86:241–305, 2005.

[87] J W Yewdell et J R Bennink. Immunodominance in major histocompatibility complex class i-restricted t lymphocyte responses. *Annu Rev Immunol*, 17:51–88, 1999.

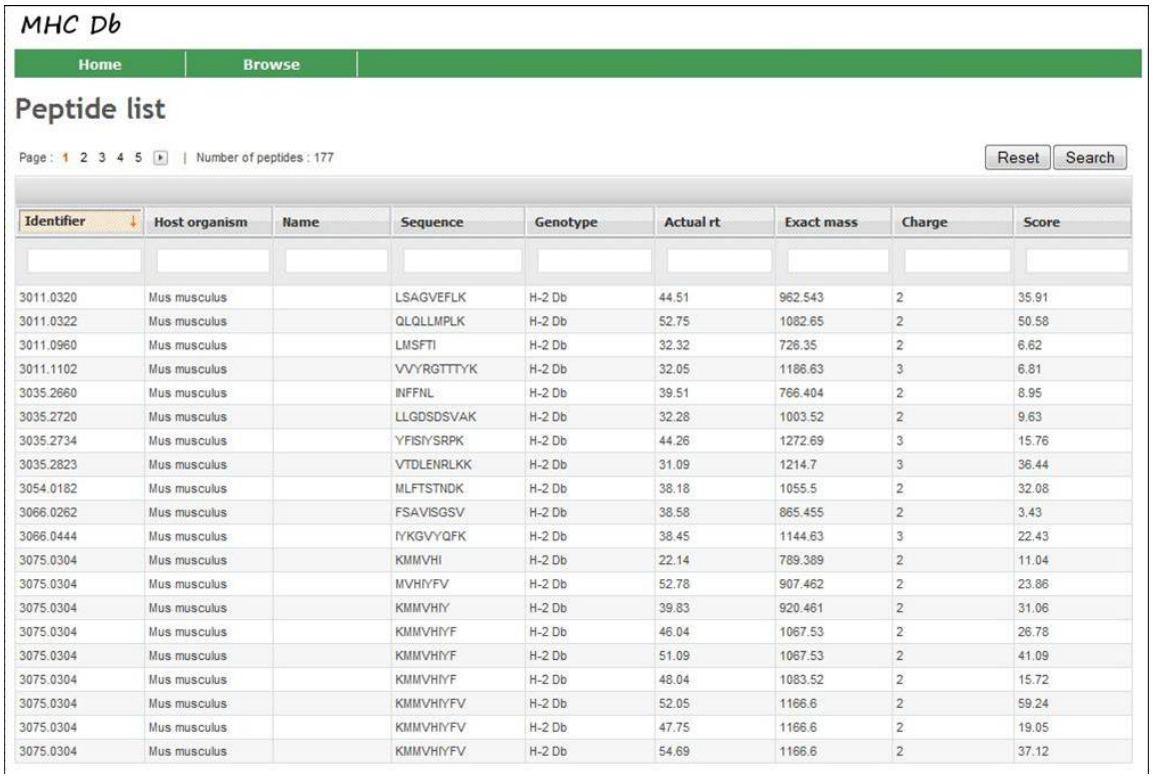
[88] Jonathan Yewdell. To drip or not to drip : generating peptide ligands for mhc class i molecules from biosynthesized proteins. *Mol Immunol*, 39(3-4):139–46, Oct 2002.

[89] Chenhong Zhang, Mikelis G Bickis, Fang-Xiang Wu et Anthony J Kusalik. Optimally-connected hidden markov models for predicting mhc-binding peptides. *J Bioinform Comput Biol*, 4(5):959–80, Oct 2006.

Annexe I

Copie d'écran d'une recherche de peptides avec MHCDB

En haut de chacune des colonnes, il est possible de spécifier une valeur de filtrage. Il suffit de cliquer sur la ligne correspondant à un peptide pour obtenir de plus amples informations.



MHC Db

Home | Browse

Peptide list

Page : 1 2 3 4 5 | Number of peptides : 177

Reset Search

Identifier	Host organism	Name	Sequence	Genotype	Actual rt	Exact mass	Charge	Score
3011.0320	Mus musculus		LSAGVEFLK	H-2 Db	44.51	962.543	2	35.91
3011.0322	Mus musculus		QLQLLMPLK	H-2 Db	52.75	1082.65	2	50.58
3011.0960	Mus musculus		LMSFTI	H-2 Db	32.32	726.35	2	6.62
3011.1102	Mus musculus		VVYRGTTTYK	H-2 Db	32.05	1186.63	3	6.81
3035.2660	Mus musculus		INFFNL	H-2 Db	39.51	766.404	2	8.95
3035.2720	Mus musculus		LLGDSDSVAK	H-2 Db	32.28	1003.52	2	9.63
3035.2734	Mus musculus		YFISIYSRPK	H-2 Db	44.26	1272.69	3	15.76
3035.2823	Mus musculus		VTDLENRLKK	H-2 Db	31.09	1214.7	3	36.44
3054.0182	Mus musculus		MLFTSTNDK	H-2 Db	38.18	1055.5	2	32.08
3066.0262	Mus musculus		FSAVISGSV	H-2 Db	38.58	865.455	2	3.43
3066.0444	Mus musculus		IYKGVYQFK	H-2 Db	38.45	1144.63	3	22.43
3075.0304	Mus musculus		KMMVHI	H-2 Db	22.14	789.389	2	11.04
3075.0304	Mus musculus		MVHIYFV	H-2 Db	52.78	907.462	2	23.86
3075.0304	Mus musculus		KMMVHIY	H-2 Db	39.83	920.461	2	31.06
3075.0304	Mus musculus		KMMVHIYF	H-2 Db	46.04	1067.53	2	26.78
3075.0304	Mus musculus		KMMVHIYF	H-2 Db	51.09	1067.53	2	41.09
3075.0304	Mus musculus		KMMVHIYF	H-2 Db	48.04	1083.52	2	15.72
3075.0304	Mus musculus		KMMVHIYFV	H-2 Db	52.05	1166.6	2	59.24
3075.0304	Mus musculus		KMMVHIYFV	H-2 Db	47.75	1166.6	2	19.05
3075.0304	Mus musculus		KMMVHIYFV	H-2 Db	54.69	1166.6	2	37.12

Annexe II

Proportion des fragments observés par rapport aux fragments théoriques

Tableau II.I – Rapport du nombre de fragments observés de chaque type par rapport au nombre de fragments théoriques du même type.

		Peptides trypsiques		Peptides du CMH-I	
Type	Charge	Proportion moyenne	IC95	Proportion moyenne	IC95
y	1	80,7	+/-1,7	58,0	+/-3,4
y	2	24,6	+/-5,2	16,4	+/-8,7
y-H2O	1	32,7	+/-5,0	24,5	+/-7,1
y-H2O	2	8,6	+/-10,4	5,3	+/-13,0
y-NH3	1	21,6	+/-6,4	12,9	+/-10,2
y-NH3	2	20,3	+/-6,1	12,5	+/-9,6
b	1	59,2	+/-3,2	52,2	+/-3,6
b	2	6,0	+/-12,8	10,7	+/-7,9
b-H2O	1	45,3	+/-4,5	26,6	+/-6,5
b-H2O	2	5,8	+/-11,7	5,9	+/-11,5
b-NH3	1	20,1	+/-9,0	16,0	+/-10,6
b-NH3	2	11,4	+/-8,1	12,5	+/-8,8
a	1	25,7	+/-5,8	13,8	+/-7,8
a	2	9,2	+/-9,7	9,5	+/-9,2
a-H2O	1	16,5	+/-7,7	6,5	+/-12,9
a-H2O	2	5,1	+/-13,1	3,2	+/-15,6
a-NH3	1	20,4	+/-6,4	8,5	+/-11,6
a-NH3	2	10,6	+/-9,4	9,1	+/-10,4
interne	1	41,8	+/-5,4	27,8	+/-7,3
interne	2	14,1	+/-9,9	8,5	+/-13,6
interne-H2O	1	32,8	+/-6,0	13,7	+/-10,3
interne-H2O	2	10,9	+/-11,3	6,0	+/-14,9
interne-NH3	1	27,8	+/-6,0	12,0	+/-11,9
interne-NH3	2	5,9	+/-13,5	2,6	+/-20,5

Annexe III

Profil de fragmentation

Tableau III.I – Profil de fragmentation CID des peptides HLA-A*01 de 9 résidus (n=55)

Type	Charge	Perte neutre	Position	$\bar{I}$	Écart-type	2,5 perc.	97,5 perc.
y	1		1	6.92	1.83	3.36	10.66
y	1		2	7.14	2.52	3.25	12.82
y	1		3	7.46	3.53	2.4	15.8
y	1		4	14.11	5.04	5.69	25.24
y	1		5	13.99	5.24	5.16	25.37
y	1		6	7.52	2.72	2.96	13.47
y	1		7	26.94	8.74	11.06	44.99
y	1		8	2.04	1.01	0.52	4.36
y	1	H2O	1	0.04	0.04	0	0.14
y	1	H2O	2	1.18	0.63	0.13	2.59
y	1	H2O	3	0.89	0.46	0.21	1.97
y	1	H2O	4	2.27	1.26	0.53	5.24
y	1	H2O	5	2.13	1.88	0.11	6.97
y	1	H2O	6	2.46	2.91	0.05	10.29
y	1	H2O	7	0.91	0.54	0.17	2.21
y	1	H2O	8	1.75	1.81	0.05	6.3
y	1	NH3	1	1.46	0.52	0.56	2.59
y	1	NH3	2	0.35	0.25	0	0.93
y	1	NH3	3	0.51	0.42	0	1.51
y	1	NH3	4	1.33	1.41	0	5.1
y	1	NH3	5	2.5	3.09	0.14	10.65
y	1	NH3	6	0.71	0.85	0	2.99
y	1	NH3	7	0.52	0.38	0	1.4
b	1		2	2.98	1.08	1.11	5.28
b	1		3	1.67	0.79	0.48	3.51
b	1		4	4.41	1.84	1.55	8.64
b	1		5	7.71	3.34	2.31	15.12
b	1		6	15.04	5.88	4.91	27.98
b	1		7	23.21	6.73	11.48	37.61

b	1		8	34.25	9.3	16.75	53.13
b	1	H2O	2	0.48	0.26	0.04	1.05
b	1	H2O	3	1.4	1.11	0.11	4.07
b	1	H2O	4	3.03	2.63	0.24	9.69
b	1	H2O	5	2.42	1.67	0.39	6.51
b	1	H2O	6	4.15	3.08	0.86	12.29
b	1	H2O	7	2.43	1.44	0.57	5.95
b	1	H2O	8	3.55	2.97	0.65	11.48
b	1	NH3	1	0.02	0.04	0	0.13
b	1	NH3	2	0.2	0.22	0	0.76
b	1	NH3	3	0.15	0.19	0	0.66
b	1	NH3	4	0.69	0.4	0.04	1.57
b	1	NH3	5	1.74	1.67	0	5.98
b	1	NH3	6	0.72	0.46	0	1.76
b	1	NH3	7	0.53	0.34	0	1.3
b	1	NH3	8	0.87	0.53	0	2.09
a	1		2	2.39	1.45	0.4	5.73
a	1		3	0.08	0.07	0	0.25
a	1		4	0.58	0.34	0.04	1.34
a	1		5	0.83	0.45	0.08	1.86
a	1		6	0.48	0.39	0	1.41
a	1		7	0.06	0.07	0	0.21
a	1		8	0.15	0.17	0	0.61
a	1	H2O	2	0.3	0.22	0	0.8
a	1	H2O	3	0.38	0.37	0	1.33
a	1	H2O	4	0.77	0.66	0.06	2.49
a	1	H2O	5	0.69	0.52	0	1.85
a	1	H2O	6	0.26	0.19	0	0.68
a	1	H2O	7	0.13	0.17	0	0.6
a	1	H2O	8	0.11	0.11	0	0.4
a	1	NH3	2	0.4	0.51	0	1.77
a	1	NH3	3	0.6	0.69	0	2.3
a	1	NH3	4	0.88	1.06	0.02	3.82
a	1	NH3	5	1.22	1.32	0	4.5
a	1	NH3	6	0.12	0.1	0	0.36
a	1	NH3	7	0.04	0.07	0	0.23
a	1	NH3	8	1.78	3	0	10

Tableau III.II – Profil de fragmentation HCD des peptides HLA-A*01 de 9 résidus (n=51)

Type	Charge	Perte neutre	Position	$\bar{I}$	Écart-type	2,5 perc.	97,5 perc.
y	1		1	47.46	9.04	30.1	65.68
y	1		2	22.53	5.24	13	33.49
y	1		3	11.39	5.52	2.34	23.54
y	1		4	11.84	4.72	4.12	22.14
y	1		5	11.74	3.81	4.83	19.59
y	1		6	10.52	4.04	3.83	19.43
y	1		7	22.75	7.61	8.83	38.25
y	1		8	9.26	4.25	2.28	18.84
y	1	H2O	2	6.88	3.04	1.83	13.4
y	1	H2O	3	2.54	1.76	0.27	6.69
y	1	H2O	4	5.09	3.38	0.95	13.33
y	1	H2O	5	1.68	0.86	0.46	3.63
y	1	H2O	6	1.59	0.91	0.27	3.73
y	1	H2O	7	3.22	1.51	0.88	6.65
y	1	H2O	8	1.32	0.71	0.22	2.9
y	1	NH3	1	20.86	4.24	12.89	29.57
y	1	NH3	2	1.73	2.08	0.06	6.97
y	1	NH3	3	0.34	0.23	0	0.87
y	1	NH3	4	0.47	0.47	0	1.63
y	1	NH3	5	0.81	0.54	0.08	2.11
y	1	NH3	6	0.48	0.29	0	1.13
y	1	NH3	7	1.17	0.78	0.1	2.92
y	1	NH3	8	0.62	0.65	0	2.16
b	1		1	0.23	0.18	0	0.66
b	1		2	18.48	4.04	10.96	26.63
b	1		3	8.39	2.66	3.68	14.25
b	1		4	6.77	2.37	2.85	12.03
b	1		5	7.44	3.56	2.48	16.07
b	1		6	10.98	3.76	4.4	19.08
b	1		7	11.76	4.54	4.21	21.88
b	1		8	7.84	2.81	3.22	14.24
b	1	H2O	2	8.71	3.47	2.89	16.41
b	1	H2O	3	9.82	3.89	3.31	18.11

b	1	H2O	4	1.38	0.65	0.26	2.8
b	1	H2O	5	2.84	0.92	1.16	4.76
b	1	H2O	6	5	2.45	1.41	10.61
b	1	H2O	7	4.71	2.08	1.49	9.3
b	1	H2O	8	2.25	0.89	0.74	4.17
b	1	NH3	1	0.83	1.25	0	4.07
b	1	NH3	2	0.67	0.82	0	2.75
b	1	NH3	3	0.27	0.34	0	1.07
b	1	NH3	4	0.98	1.22	0	4.2
b	1	NH3	5	0.38	0.28	0	1.02
b	1	NH3	6	0.65	0.67	0	2.38
b	1	NH3	7	0.39	0.29	0	1.09
b	1	NH3	8	0.27	0.36	0	1.26
a	1		2	41.04	8.02	25.98	57.21
a	1		3	0.07	0.07	0	0.23
a	1		4	2.1	0.81	0.73	3.85
a	1		5	1.88	1	0.45	4.16
a	1		6	2.27	1.02	0.55	4.49
a	1		7	3.8	1.97	0.59	8.21
a	1		8	1.82	1.09	0.21	4.3
a	1	H2O	2	7.67	3.97	1.87	16.92
a	1	H2O	3	0.48	0.3	0	1.17
a	1	H2O	4	0.77	0.45	0.08	1.83
a	1	H2O	5	0.76	0.46	0.11	1.8
a	1	H2O	6	0.69	0.35	0.09	1.46
a	1	H2O	7	0.62	0.46	0	1.71
a	1	H2O	8	0.21	0.18	0	0.62
a	1	NH3	1	0.11	0.15	0	0.48
a	1	NH3	2	1.8	1.6	0	5.5
a	1	NH3	3	0.85	1.03	0	3.36
a	1	NH3	4	0.95	0.61	0.07	2.4
a	1	NH3	5	1.72	1.88	0.08	6.38
a	1	NH3	6	0.75	0.47	0	1.84
a	1	NH3	7	0.28	0.24	0	0.84
a	1	NH3	8	0.08	0.09	0	0.27

Annexe IV

Liste des combinaisons d'acides aminés possibles pour les masses comprises entre 114 u et 217 u

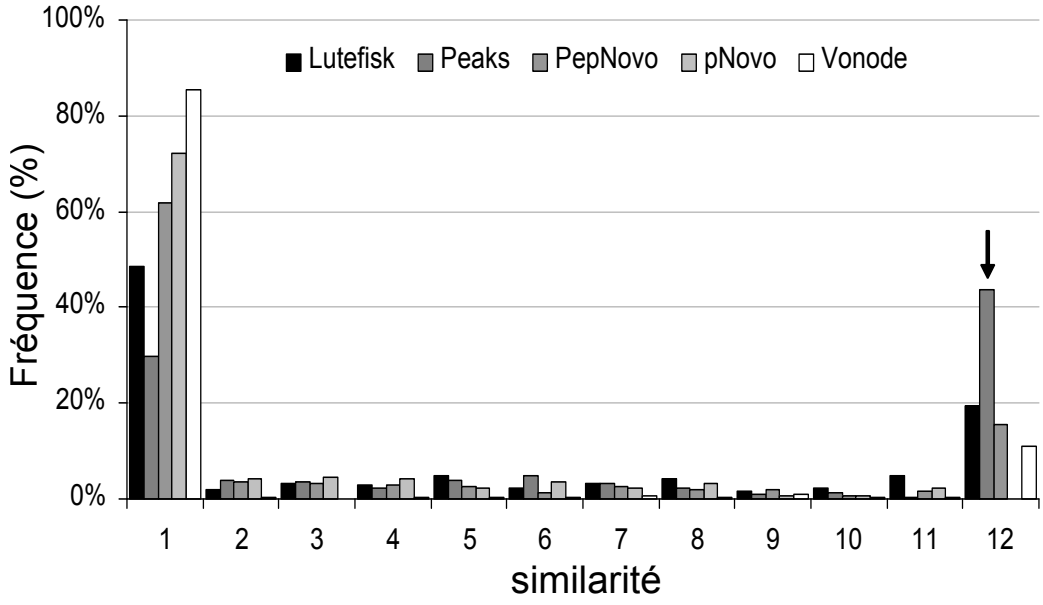
Masse nominale	Acides aminés	Différence	Masse nominale	Acides aminés	Différence
114	N	0,04293	218	A,F	0,10553
	G,G	0,04293		D,C	0,03613
115	D	0,02694		M,S	0,07251
128	Q	0,05858	220	G,Y	0,08479
	K	0,09496	224	H,S	0,09094
	A,G	0,05858	225	Q,P	0,11134
129	E	0,04259		K,P	0,14773
131	M	0,04048		A,G,P	0,11134
137	H	0,05891	226	E,P	0,09536
142	A,A	0,07423		I,I	0,16813
144	G,S	0,05349		I,L	0,16813
147	F	0,06841		L,L	0,16813
154	G,P	0,07423	227	A,R	0,13823
156	R	0,10111		N,I	0,12699
	G,V	0,08988		N,L	0,12699
158	A,S	0,06914		Q,V	0,12699
	G,T	0,06914		K,V	0,16338
160	C,G	0,03065		A,G,V	0,12699
163	Y	0,06333		G,G,I	0,12699
168	A,P	0,08988		G,G,L	0,12699
170	A,V	0,10553	228	N,N	0,08585
	G,I	0,10553		D,I	0,11101
	G,L	0,10553		D,L	0,11101
171	N,G	0,06439		E,V	0,11101
	G,G,G	0,06439		M,P	0,09325
172	A,T	0,08479		N,G,G	0,08586
	D,G	0,04841	229	N,D	0,06987
174	A,C	0,04630		Q,T	0,10626
	S,S	0,06406		K,T	0,14264

184	A,I	0,12118		A,A,S	0,10626
	A,L	0,12118		A,G,T	0,10626
	P,S	0,08479		D,G,G	0,06987
185	A,N	0,08004	230	D,D	0,05389
	Q,G	0,08004		E,T	0,09027
	G,K	0,11643		M,V	0,10890
	A,G,G	0,08004	231	C,Q	0,06776
186	W	0,07931		C,K	0,10415
	A,D	0,06406		A,C,G	0,06776
	E,G	0,06406		G,S,S	0,08552
	S,V	0,10044	232	C,E	0,05178
188	G,M	0,06195		M,T	0,08816
	S,T	0,07971	234	A,Y	0,10044
190	C,S	0,04121		C,M	0,04967
194	G,H	0,08038		H,P	0,11168
	P,P	0,10553		F,S	0,10044
196	P,V	0,12118	236	H,V	0,12733
198	P,T	0,10044	238	H,T	0,10659
	V,V	0,13683	239	A,A,P	0,12699
199	A,Q	0,09569	240	C,H	0,06810
	A,K	0,13208	241	Q,I	0,14264
	A,A,G	0,09569		Q,L	0,14264
200	A,E	0,07971		I,K	0,17903
	C,P	0,06195		L,K	0,17903
	I,S	0,11609		A,A,V	0,14264
	L,S	0,11609		A,G,I	0,14264
	T,V	0,11609		A,G,L	0,14264
201	N,S	0,07495		G,P,S	0,10626
	G,G,S	0,07496	242	N,Q	0,10151
202	A,M	0,07760		N,K	0,13789
	D,S	0,05897		E,I	0,12666
	C,V	0,07760		E,L	0,12666
	T,T	0,09536		A,N,G	0,10151
204	C,T	0,05686		Q,G,G	0,10151
	G,F	0,08988		G,G,K	0,13789
206	C,C	0,01837	243	R,S	0,13314
208	A,H	0,09603		N,E	0,08552

210	I,P	0,13683		D,Q	0,08552
	L,P	0,13683		D,K	0,12191
211	N,P	0,09569		G,W	0,10078
	G,G,P	0,09569		A,A,T	0,12191
212	D,P	0,07971		D,G,A	0,08552
	I,V	0,15248		E,G,G	0,08552
	L,V	0,15248		G,S,V	0,12191
213	R,G	0,12258	244	D,E	0,06954
	N,V	0,11134		I,M	0,12455
	A,A,A	0,11134		L,M	0,12455
	G,G,V	0,11134		F,P	0,12118
214	D,V	0,09536	245	N,M	0,08341
	I,T	0,13174		A,A,C	0,08341
	L,T	0,13174		A,S,S	0,10117
215	N,T	0,09061		G,G,M	0,08341
	Q,S	0,09061		G,S,T	0,10117
	K,S	0,12699	246	D,M	0,06743
	A,G,S	0,09061		F,V	0,13683
	G,G,T	0,09061	247	C,G,S	0,06268
216	D,T	0,07462	248	F,T	0,11609
	C,I	0,09325	250	C,F	0,07760
	C,L	0,09325		H,I	0,14298
	E,S	0,07462		H,L	0,14298
217	N,C	0,05211		S,Y	0,09536
	C,G,G	0,05211			

Annexe V

Comparaison des performances des différents programmes de séquençage *de novo* par la mesure de similarité avec la première séquence candidate uniquement



Annexe VI

Séquences *in silico* avec score > 99 erronées à cause de la non-prise en compte des modifications post-transcriptionnelles pertinentes

M* indique un acide aminé M oxydé.

La séquence réelle	La séquence *in silico*	Δ masse
HTSSM*RGVYY	HTSHPPGTGYY	16
EDQLLPFM*SD	TPSAELYCEPT	16
KFM*LNVSYL	KFPYGGVSYL	16
KAVYNFATM*	KAVYNFGDF	16
HTLM*SIVSS	SYPTHLVSS	16
KQM*YKTPTLK	KSMGYKTPTLK	16
KM*NPPKFSKV	QFNPPKFSKV	16
FM*SHVKSVTK	YMSHVKSVTK	16
LTSM*KYFVK	LTSFGAYFVK	16
FM*KIGAHPI	YMKLGAHPL	16

Annexe VII

Organismes pour lesquels l'ensemble des protéines a été utilisé pour la recherche MASCOT dans les conditions réelles

Tableau VII.I – Organismes pour les allèles H2-Db et H2-Kb

Mus musculus
Bordetella pertussis
Coxiella brunetti
Virus Guanarito
Virus Junin
Lassa
Virus de la chorioméningite lymphocytaire
Virus Machupo
Mouse (DQ8)
Herpesvirus 4 de Murid
Virus sabia
Virus Vaccinia
Virus Whitewater Arroyo

Tableau VII.II – Organismes pour les allèles HLA-A*01, HLA-A*03 et HLA-A*03

Dengue
Virus de l'hépatite B
Virus de l'hépatite C
Virus Guanarito
Homo sapiens
Infuenza
Virus Junin
Lassa
Virus de la chorioméningite lymphocytaire
Virus Machupo
Virus Sabia
SARS
Triticum aestivum
Virus Vaccinia
Virus variole (Inde)
Virus variole (Bengladesh)
Virus Whitewater Arroyo

Annexe VIII

Chaîne de traitement pour la recherche d'antigènes mineurs d'histocompatibilité

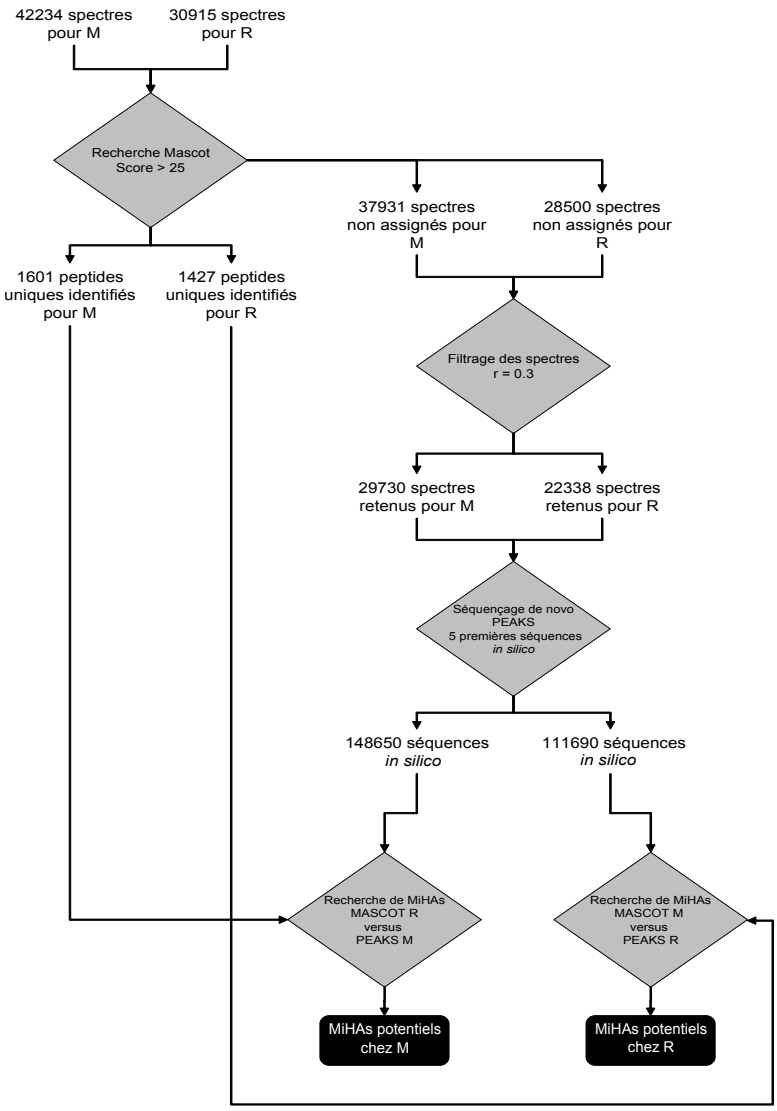


Figure VIII.1 – Schéma général de la chaîne de traitement pour la recherche de SNPs potentiels. Pour chacun des individus M et R, on a 3 réplicats composés de 5 échantillons. La chaîne de traitement se compose de 4 principales étapes : filtrage des spectres, séquençage *de novo*, filtrage des séquences *de novo* et enfin recherche de MiHAs

Annexe IX

Fréquence des substitutions d'acides aminés entre les séquences retournées par MASCOT et les séquences *de novo* retournées par PEAKS

Mascot pour M versus Peaks pour R		
Substitutions	Fréquence	Score substitution
E -> Q	118	1.4
D -> N	59	1.4
Q -> P	4	1.3
S -> N	3	2.6
I -> T	2	1.7
R -> K	2	2.6
Q -> E	2	1.4
V -> L	2	5.1
A -> S	1	1.9
R -> H	1	2.6
S -> G	1	1.5
E -> K	1	1.4
Y -> N	1	3.3

Tableau IX.I

Mascot pour R versus Peaks pour M		
Substitutions	Fréquence	Score substitution
E -> Q	137	1.4
D -> N	63	1.4
Q -> P	4	1.3
E -> D	3	3.8
Y -> N	3	3.3
V -> L	2	5.1
T -> S	2	5.1
I -> V	1	3.6
R -> P	1	1.6
L -> P	1	4.2
A -> S	1	1.9
S -> T	1	5.1
K -> Q	1	1.4
S -> P	1	2.1
Q -> E	1	1.4
E -> G	1	1.7
A -> P	1	4.3

Tableau IX.II