Extraction de connaissances à partir de données de protéomique de découverte haut-débit

Thomas Burger

Université Grenoble-Alpes

N° attribué par la bibliothèque

Mémoire présenté en vue de l'obtention d'une

Habilitation à diriger des recherches

Spécialité **Théorie, modèle et instrumentation pour la santé, la cognition et l'environnement**

Préparée au laboratoire **Biologie à Grande Echelle**
dans le cadre de l'École Doctorale **Ingénierie pour la santé la Cognition et l'Environnement**

Présentée et soutenue publiquement
par

Thomas Burger

le 23 janvier 2017

Titre :

Extraction de connaissances à partir de données de protéomique de découverte haut-débit

Jury

Marc-André Delsuc (D.R. CNRS)	Rapporteur
Antoine Cornuéjols (P.U. AgroParisTech)	Rapporteur
Michael Blum (D.R. CNRS)	Rapporteur
Joëlle Vinh (D.R. CNRS)	Examinatrice
Benno Schwikowski (D.R. Institut Pasteur)	Examinateur
Xavier Gidrol (D.R. INRA, CEA)	Président du Jury

Résumé

Ce mémoire présente mon travail d'encadrement d'activités de recherche pour les années 2012-2016, ainsi que des perspectives pour les cinq années à venir. Au travers de la présentation de deux de mes projets de recherche, j'analyse les différentes questions suscitées par l'animation d'un groupe de recherche dont l'objectif est le développement d'outils et de méthodes permettant l'extraction de connaissances automatisées à partir de données de quantification relative en protéomique *label-free* obtenues par spectrométrie de masse haut-débit. Ces questions concernent notamment *(i)* l'encadrement de jeunes chercheurs et la valorisation de leurs activités ; *(ii)* la recherche de financements ; et surtout *(iii)* la gestion des difficultés spécifiques au contexte interdisciplinaire (mode de diffusion/valorisation, équilibre entre recherche et ingénierie, pilotage et priorisation des sujets de recherche, etc.). Le premier projet présenté, ProStaR, est un outil logiciel permettant de faciliter l'analyse statistique de données protéomiques. Au-delà de l'important travail d'ingénierie que sa réalisation a nécessité, je montre qu'il peut être le support de nombreux petits projets relativement indépendants mais novateurs en science des données. Le second projet, Reveal-MS, propose de résoudre le démultiplexage de spectres de peptides par des méthodes innovantes de factorisation non-négative de matrices de grandes tailles. À l'inverse du précédent projet et dans une logique complémentaire, celui-ci est moins motivé par les besoins quotidiens de la protéomique que par la possibilité à long terme de permettre une rupture dans l'état de l'art.

Abstract

This dissertation presents my research activity supervision during years 2012 through 2016, as well as prospects for the next five years. Through the presentation of two of my research projects, I analyze the various issues raised by leading a research group focused on the development of tools and methods for automated knowledge extraction from relative quantification of high-throughput mass spectrometry data for label-free quantitative proteomics. Mainly, these issues relates to *(i)* advising young researchers and publishing their result ; *(ii)* grant chasing ; and most importantly *(iii)* managing the interdisciplinary context (publication strategies, balance between research and engineering, management and prioritization of research subjects, etc.). The first project presented, referred to as ProStaR, is a software tool to ease the statistical analysis of proteomic data. Beyond the important engineering workload it required, I show it is now a support to many small and independent yet innovative projects in data science. The second project, referred to as Reveal-MS, proposes to address the demultiplexing of peptide spectra by innovative non-negative factorization methods for large matrices. In contrast with the previous project, it is less motivated by the daily needs of proteomic labs than by the long-term hope of a state-of-the-art breakthrough.

Remerciements

Tout d'abord, je remercie les trois rapporteurs de ce travail. Chacun d'eux est à sa manière un modèle scientifique que j'espère pouvoir suivre, et me faire évaluer par eux est à la fois un honneur et un défi. Marc-André incarne ce à quoi la recherche interdisciplinaire devrait ressembler : focalisée sur des sujets originaux, stimulants, et indépendants des modes scientifiques ; curieuse des différents domaines de la connaissance ; et à la fois intégrative et bienveillante vis-à-vis de ceux qui y participent. Antoine est tout simplement l'auteur de l'ouvrage français de référence en apprentissage automatique : la facilité déconcertante avec laquelle il permet de comprendre la théorie comme le contexte historique de la discipline est un exemple que je cherche à suivre dans mes échanges interdisciplinaires quotidiens. Enfin, ma différence d'âge avec Michael n'est que de quelques années, mais si je pouvais atteindre avant la retraite son h-index *d'aujourd'hui, je serais déjà fier de ma carrière. Je remercie aussi chaleureusement les chercheurs ayant accepté d'être examinateurs durant la soutenance : Joëlle, qui dirige un laboratoire dont les recherches en protéomique correspondent typiquement à celles auxquelles mes outils prétendent être utiles ; Benno, le "papa" de* Cytoscape*, qui a si bien réussi cette convergence entre ingénierie et recherche, qui fait la difficulté des domaines interdisciplinaires ; et Xavier, qui dirige l'unité où tous les travaux décrits ci-après ont pu être menés avec succès.*

*Ensuite, mes égards vont à mes collègues de travail et à l'ensemble des membres contractuels ou permanents du laboratoire EDyP. Ma carrière passée est encore courte, mais je n'ai pour l'instant jamais rencontré de laboratoire où une telle largeur de spectre disciplinaire était couverte par un si petit nombre de personnes travaillant réellement ensemble au quotidien. C'est aussi le seul environnement de recherche que j'ai connu où il est universellement admis que tout chercheur, y compris le plus "capé", est forcément le néophyte de quelqu'un ; et que les échanges scientifiques doivent être conduits à la fois dans le respect de l'ignorance et de la culture scientifique des interlocuteurs. Parmi les membres du laboratoire, je tiens particulièrement à saluer, par ordre chronologique Jérôme (fondateur du laboratoire et maintenant directeur de l'institut) et Yves (responsable de l'ancien laboratoire BIM) pour avoir eu confiance en mon CV si différent des leurs, et pour avoir permis mon recrutement. Je remercie aussi Christophe et Myriam, les co-directeurs d'EDyP pour l'accueil, le soutient et la confiance qu'ils m'ont apportés. Par ce biais, ils ont contribué de manière déterminante à mon épanouissement professionnel et au succès du groupe KDPD (*Knowledge Discovery From Proteomic Data*). Naturellement, je remercie aussi ce groupe, constitué de Florence, Samuel, Thomas, Cosmin, Jimmy, Quentin, Olga et des stagiaires qui nous ont accompagnés : "Vous faites tous un super boulot,*

et je souhaite que ce rapport soit une reconnaissance de la qualité de votre travail". Enfin, je voudrais saluer Yohann, le responsable de la plateforme instrumentale, dont la perception des enjeux interdisciplinaires et de l'intérêt de la data science *pour la protéomique a été réellement déterminante pour le dynamisme du groupe KDPD.*

Pour en finir avec les remerciements d'ordre professionnel, je voudrais exprimer ma profonde gratitude à l'égard de Jimmy Wales et des nombreux contributeurs anonymes de Wikipédia : je sais qu'il ne s'agit pas d'une référence bibliographique convenable, mais cette encyclopédie reste de mon point de vue un accélérateur phénoménal d'acquisition de connaissance et de culture générale interdisciplinaire. En 1995, j'avais participé à un concours pour adolescents (le GII Junior Summit*), dans lequel il nous était demandé de décrire nos espérances quant aux révolutions qu'Internet, alors balbutiant, allait permettre. La mienne était celle d'un accès universelle à la connaissance. Jimmy Wales a concrétisé cela seulement 6 ans plus tard, en 2001.*

Pour finir, je remercie du fond du cœur mon père Jacques, ma mère Monique, ma petite sœur Marine et ma compagne Isabelle. Je tiens une preuve, expérimentale certes, mais irréfutable malgré tout, de leur amour, à savoir l'intérêt qu'ils ont porté à ce document. Regardons les choses en face : il est long, il est technique, il a été plein de fautes d'orthographe, et il parle de sujets qui ne les impactent pas vraiment au quotidien. Bref, il n'a à leurs yeux que l'intérêt d'avoir été écrit par mes soins.

Tom

Préface - retour sur mon parcours

Mon cursus académique, d'abord comme étudiant, puis en tant que Maître de Conférences, et enfin maintenant, comme chargé de recherche au CNRS, a subi de nombreux changements de direction. Ainsi, au-delà de la reconnaissance académique, l'intérêt de ce mémoire d'HDR (Habilitation à Diriger des Recherches) est de me permettre de révéler la cohérence de mon cheminement scientifique.

Mon premier cycle universitaire, en classes préparatoires, fut principalement placé sous le signe de la Physique. Celle-ci m'a été enseignée selon une approche "fondamentale" qui faisait la part-belle à la modélisation et à la mise en équation, et que j'aurais volontiers poursuivie. Cependant, j'avais l'intuition que mon second cycle, en école d'ingénieur, me décevrait, en insistant sur le génie et les procédés, au détriment de cette modélisation que j'aimais tant ; à moins que je ne m'oriente vers quelques domaines spécifiques, tels que la physique subatomique ou relativiste, les nouvelles technologies de l'information et de la communication (NTIC), la finance ou la biologie.

Parmi ces domaines, j'aurais dû m'orienter vers celui que mon goût privilégiait. C'est cependant le hasard des concours qui a choisi les NTIC. Ainsi, j'ai oscillé entre traitement du signal et recherche opérationnelle durant les deux dernières années de mon cycle d'ingénieur. À l'époque, je n'avais aucunement conscience du lien particulièrement fort qu'il y a entre la transformée de Fourier d'une distribution (au cœur de tout cours en traitement du signal), et entre l'évolution d'une marche aléatoire (au centre de la théorie des files d'attente) : il "suffit" de définir la première sur un graphe plutôt que sur $\mathbb{C}$, et d'analyser la seconde à différents instants plutôt qu'au terme de sa convergence, pour se rendre compte que l'un comme l'autre permettent de caractériser la structure multi-échelle du graphe. Malheureusement, je n'avais pas encore le recul nécessaire à la compréhension de cette "vérité supérieure", qui m'aurait ravi autant qu'un physicien découvrant que deux lois fondamentales pouvaient être unifiées. Je continuais donc mes déambulations universitaires aux allures de mouvement brownien, avec l'insatisfaisante sensation de m'éloigner de ce qui me plaisait tant, tout en passant d'un domaine à l'autre, au gré des opportunités.

C'est donc ainsi qu'en fin de DEA, je me suis tourné vers les sciences cognitives et le traitement de la parole. Le hasard des rencontres m'a fait continuer en thèse sur le traitement vidéo du langage (lecture labiale et reconnaissance de gestes). Durant ce troisième cycle, quelques réflexions se sont imposées à moi :

La première est qu'en sciences humaines, les notions de modèle, de prédiction, d'estimation de l'erreur, etc. n'ont parfois qu'une importance relative ; il arrive encore souvent que ce qui prime soient l'école de pensée ainsi que les capacités à positionner une opinion et à l'enrichir au gré des rencontres ou des obstacles.

Ensuite, j'ai compris que mon intérêt pour le traitement d'images (au centre de ma thèse) était moins fort que celui que je portais à la compréhension du système cognitif ; que celle-ci emprunte les modélisations sommaires des sciences du langage, ou les méthodes d'apprentissage automatique. J'aurais pu à l'époque me concentrer sur les sciences de l'information, et ainsi me réconcilier avec mes penchants pour la modélisation, en abordant le traitement d'image d'un point de vue plus mathématique et statistique, notamment via la théorie des ondelettes (on revient sur la caractérisation multi-échelle d'un graphe), ou les espaces de Hilbert à noyau reproduisant [1] ; ou encore, aborder l'apprentissage automatique sous l'angle des statistiques et de l'optimisation (la recherche opérationnelle me donnant une bonne base pour cela). À la place, j'ai choisi la voie de l'intelligence artificielle plus traditionnelle, probablement en raison d'un mode de fonctionnement se rapprochant un peu plus de celui des sciences humaines, au milieu desquelles je baignais alors, et qui ne me déplaisait pas.

Par ailleurs, je me suis mis à enseigner les statistiques, sans me rendre compte qu'il s'agissait du chaînon manquant entre le traitement du signal, la recherche opérationnelle, et bien d'autres domaines des sciences de l'information, dans lesquels, j'aurais pu épanouir mes penchants pour la modélisation.

Enfin, j'ai compris à quel point "l'interdisciplinarité" pouvait être compliquée. Alors qu'elle s'apparente à la promesse de grandes découvertes pour le grand public, elle est souvent considérée par les chercheurs comme un simple joint, dont l'intérêt principal est de permettre une transition lisse entre deux sections universitaires. La difficulté ne vient pas, comme on pourrait le croire, du besoin de maîtrise de deux domaines du savoir. Mais simplement du fait que deux chercheurs de deux domaines différents n'ont ni le même vocabulaire, ni les mêmes questions, de sorte qu'en plus d'être difficile, le dialogue n'est dans bien des cas, même pas nécessaire.

J'ai donc passé les trois années de ma thèse (puis une quatrième, celle des concours de Maître de Conférences) à mettre en place les briques d'un édifice que je ne pouvais pas encore voir. J'ai obtenu un poste dans une université bretonne, où j'ai été affilié, probablement sur des critères plus politiques que scientifiques, à une "future" équipe recherche (c'est-à-dire fictive) dans un laboratoire géographiquement éclaté. Dans une telle situation, il est difficile de prendre du recul sur l'édifice en construction, tellement le combat solitaire et quotidien du "publish or perish" est angoissant ; la cohérence de mon parcours scientifique n'était donc plus ma priorité. Sans thématique commune avec les rares chercheurs actifs de l'équipe, sans installation expérimentale pour continuer mes travaux de thèse, j'ai fait feu de tout bois : j'ai accepté de collaborer avec toutes les personnes qui me le proposaient, sur tous les sujets pour lesquels le rapport "publications sur temps de travail" me semblait acceptable : comme en atteste maintenant ma liste de publications, j'ai durant cette période beaucoup publié de travaux peu aboutis, sur des sujets très divers. Néan-

1. Qu'il s'agisse de la théorie des ondelettes ou de celle des noyaux, les réseaux de neurones profonds permettent d'éclairer de manière particulièrement intéressante leurs liens avec les sciences cognitives, la perception et le traitement d'image. C'est en thèse que j'ai eu cette première révélation, au moment où le goût retrouvé pour les études me poussa à envisager une carrière académique. J'ai eu la confirmation quelques années plus tard, dans un séminaire de Stéphane Mallat [1], que cette "révélation" était justifiée... Je commençais enfin à y voir claire dans les méandres de mon parcours académique.

moins, ayant été recruté entre autre pour ma capacité à enseigner les statistiques tout en faisant des recherches en informatique (les deux disciplines me considérant comme un mauvais élément, au double sens de transfuge et d'incompétent), je me suis investi dans ce domaine, et j'y ai pris un intérêt croissant.

Trois ans plus tard, j'ai réussi le concours du CNRS, grâce à un poste fléché sur le "traitement de données de spectrométrie de masse en protéomique". Face à d'autres candidats probablement tout aussi intéressants que moi, c'est principalement ma connaissance et mon attrait pour l'interdisciplinarité, et accessoirement une bonne culture générale sur les techniques de fouilles de données, accumulée aux cours des années précédentes, qui firent la différence. J'ai donc quitté mon statut de Maître de Conférences, trois ans après l'avoir durement gagné. En même temps que je quittais cette université, je distendais, avec un pincement au cœur, les partenariats de recherche que j'avais constitués avec toutes les personnes ayant eu la gentillesse de croire en ma motivation sincère plutôt qu'en mon équipe d'accueil anémique ; et j'avais le plaisir de constater que désormais, les informaticiens comme les statisticiens me considéraient un peu comme l'un des leurs plutôt que comme un étranger.

C'est ainsi qu'à l'automne 2011, j'intégrai le laboratoire où je travaille maintenant, sous l'égide de l'Institut des Sciences Biologiques du CNRS. Je pensais que j'allais ainsi tourner une page, et ne plus mettre mes appétences pour la modélisation au service des sciences de l'information (comme les concours l'avaient décidé exactement 10 ans plus tôt) mais au service de la biologie. C'était sans compter sur le *Big Data*. Les détracteurs de ce nouveau mot à la mode raillent son étrange faculté à pouvoir rassembler sous un dénominateur commun tous ceux qui le veulent bien : en effet, il suffirait d'avoir des listes de valeurs, d'observation ou de mesures, si possible stockées sous forme binaire dans un disque dur, pour correspondre au mot *Data* ; et de trouver une unité de quantification qui permette d'exprimer leur volume total avec un nombre à 5 ou 6 ordres de grandeur, pour être *Big*. Je pense pour ma part que ce *buzzword* a cependant permis la prise de conscience qu'une nouvelle discipline était en train d'émerger : la science des données (ou la *data science*). Grâce à ce mot, je n'étais plus un ex-futur physicien souhaitant exercer ses talents de modélisateur dans tout domaine le nécessitant ; mais j'étais, sans le savoir, en train de devenir un *data scientist*. Ce que je pensais être le domaine d'application de mes talents de modélisateur, devenait en fait ma discipline. Et j'allais utiliser et mettre en pratique les savoirs de celle-ci pour une nouvelle application qui en avait bien besoin : une branche de la biologie et de la physico-chimie qu'on nomme la protéomique.

Le positionnement de mon activité de recherche ne fait à l'heure actuelle aucun doute. Pourtant, le parcours qui m'y a amené depuis mes débuts en thèse à l'automne 2004 peut sembler quelque peu chaotique. Ce mémoire a donc pour objectif de faire ressortir la structure cachée de ce cheminement scientifique : ma direction de recherche.

Table des matières

Introduction - Grille de lecture

Les deux éléments essentiels de la direction de recherche sont la supervision de jeunes chercheurs, et la capacité à financer les travaux de l'équipe via des projets visibles. Cependant, dans le cas de recherches interdisciplinaires, une composante supplémentaire apparaît, aussi indispensable que les deux précédentes : la participation à la création d'une communauté scientifique. Cela implique plusieurs efforts distincts, tels que le dialogue entre les différentes disciplines impliquées, ou encore la mise en place de synergies entre disciplines connexes.

Cette interdisciplinarité a dans mon cas, deux particularités : tout d'abord, elle est émergente et n'est pas encore aussi bien structurée que d'autres domaines interdisciplinaires. Ensuite, elle s'inscrit dans un laboratoire qui possède une forte compétence technologique et réserve une part significative de sa plateforme à une activité de support à la recherche. Dans ce contexte, les difficultés de l'interdisciplinarité mentionnées ci-dessus se retrouvent renforcées ; et complétées par d'autres, telles que notamment, la prise en compte de contraintes d'ingénierie très opérationnelles, plus difficilement compatibles avec l'activité de chercheur. Finalement, je trouve cette troisième composante de la direction de recherche plus énergivore que la recherche de financement ou l'encadrement de jeunes chercheurs. C'est pourquoi, ce document a principalement été structuré en réponse à cette problématique.

Concrètement, il est constitué de quatre chapitres de longueurs à peu près égales, regroupés en deux parties.

La première cherche à rapprocher les deux communautés scientifiques aux frontières desquelles je travaille : d'un côté, celle des chercheurs en *data science*, et de l'autre celles des protéomiciens. Je parle de deux communautés, mais en réalité, il y en a beaucoup plus : les premiers se répartissent sur plusieurs disciplines (traitement du signal, statistique, informatique et mathématiques appliquées) correspondant à autant de sections universitaires ou du CNRS ; et il en est de même pour les seconds, parmi lesquels on trouve des biologistes moléculaires, des biochimistes, des chimistes-analyticiens et des physico-chimistes, eux aussi se répartissant au sein de plusieurs communautés distinctes et entre lesquels les échanges sont déjà qualifiés "d'interdisciplinaires". Dans un tel contexte, il faut accepter de passer une partie de son temps à jouer les traducteurs-interprètes. C'est donc assez logiquement que je consacre à cette activité les deux chapitres de la première partie : le premier a pour objectif de donner en exactement 30 pages une vue synthétique de la science des données, compréhensible par un lecteur n'ayant pas de culture mathématique universitaire[2]. De manière complémentaire, le second chapitre vise à décrire les élé-

2. Je pense que les grandes lignes de ce chapitre sont accessibles à un lecteur ayant un baccalauréat scientifique, des notions élémentaires de statistiques, une idée intuitive de ce qu'est une

ments de base de la protéomique à quelqu'un n'ayant pas de culture biologique ; mais plutôt une culture "math-info". Mon objectif y est donc de fournir un point d'entrée à des chercheurs en *data science* afin qu'ils puissent facilement appréhender ce domaine d'application ayant grandement besoin de leurs compétences. Il se termine donc par la description d'un certain nombre de divergences culturelles qui affectent concrètement la vie d'un laboratoire couplé à une plateforme de service, et qu'un *data scientist* s'intéressant à la protéomique devra de toute manière gérer.

La deuxième partie décrit assez naturellement pour un tel document, une sélection de mes travaux. Dans la mesure où j'envisage de continuer à travailler sur des problématiques de protéomique, cette sélection fait largement l'impasse sur mes travaux antérieurs à mon affectation à EDyP. Parmi les travaux restants, je présente deux de mes projets de recherche, auxquels je consacre un chapitre chacun. Je les ai choisis parce qu'en plus d'être des projets majeurs de mon quotidien, ils correspondent aux critères suivants : être déjà suffisamment avancés pour pouvoir décrire les résultats associés ; être encore suffisamment loin de leur clôture pour jouer une place déterminante dans mon activité de recherche future (de 3 à 5 ans), ainsi que celle des chercheurs, ingénieurs et étudiants de l'équipe ; être fortement liés aux deux communauté aux frontières desquelles je travaille ; être chacun le reflet d'un type de recherche ; soit piloté par les besoins d'applications, soit motivées par les opportunités scientifiques. Le premier type de recherche vient souvent avec la garantie de résultats utiles, voire nécessaire à la comunauté, et dont le rayonnement sera important ; avec le risque évident d'être peu innovant, et de déplacer les efforts de la recherche vers l'ingénierie. À l'inverse, le second pourra plus facilement être en rupture avec l'état de l'art, au risque de voir sa dimension applicative rester au stade de vœux pieu.

Enfin, les annexes fournissent une vue synoptique de mon activité de recherche, en suivant le format classiquement utilisé dans les dossiers d'évaluation de carrière, ou lors d'un concours de recrutement... L'usage d'une telle grille de lecture étant maintenant la norme dans bien des situations, la plupart des lecteurs de ce document s'y retrouveront facilement.

matrice (à savoir un tableau de nombres pour lequel des opérations d'addition et de multiplication sont définies), ainsi qu'une certaine motivation.

Première partie

Positionnement interdisciplinaire

Chapitre 1

Une vue d'ensemble de la science des données

L'objectif pédagogique de ce chapitre m'a amené à utiliser un style volontairement vulgarisateur, avec peu d'équation, et beaucoup de considérations annexes, d'exemples et d'analogies qu'un théoricien du domaine trouvera inutile. Par avance, je prie un tel lecteur de m'excuser pour la peine que mes écrits lui infligeront.

1 L'extraction de connaissances

1.1 Un bref historique

Le pouvoir intégratif du cerveau humain, ou sa capacité à traiter un très grand nombre de stimuli élémentaires, afin de réaliser des tâches de haut niveau d'abstraction, est depuis longtemps, et restera pour quelques temps encore, un sujet d'émerveillement, autant qu'un enjeu scientifique. Jusqu'à une époque très récente, ce sujet n'a été abordé que de manière philosophique, par des questionnements sur la nature de la connaissance, de la conscience, de l'esprit et des idées.

C'est n'est qu'à la fin du siècle des lumières, avec le balbutiement des statistiques [2], via les problèmes de dénombrement, que des techniques formelles ont été proposées afin de mimer une procédure d'inférence cognitive élémentaire. Durant les XIX^{e} et XX^{e} siècles, la statistique s'est réellement développée, et est devenue la discipline de référence quand il s'agissait de traiter des données brutes afin d'en extraire des éléments pertinents.

En parallèle, au début du XX^{e} siècle, les sciences cognitives et la systémique ont permis de mieux cerner le processus d'apprentissage, et de caractériser précisément les langages [3] et les automates [4], fournissant les bases de l'informatique théorique. À l'époque, l'informatique n'est vu que comme un outil de calcul particulièrement rapide permettant d'automatiser [1] une machine, et rien ne lie encore la communica-

1. Ce qui explique le terme *computer science* en anglais. Notons que l'étymologie du mot français *informatique* est de ce point de vue, particulièrement prophétique. En effet, ce mot vient de la concaténation de *information* et de *automatique*, le mot *information* venant lui-même du latin, et signifiant *qui donne forme*. Ainsi, de par son étymologie, le terme français *informatique* reflète mieux cette convergence de toutes les disciplines citées dans ce paragraphe, permettant de réaliser une mise en forme automatisée qui mime une des fonctions cérébrales principales.

tion humaine ou la transmission de message d'une part et la science des calculateurs d'autre part ; alors que cette dernière est pourtant basée sur les concepts des sciences du langage. Ainsi, alors que Turing [5] imaginait déjà pendant la seconde guerre mondiale, les possibilités de l'automatisation du calcul pour réaliser des tâches de plus en plus complexes et se rapprochant des capacités cérébrales (donnant naissance au concept d'intelligence artificielle), presqu'aucun lien n'apparaît avec les questions de télécommunications et les travaux de Shannon [6] (contemporain de Turing) sur la théorie de l'information. Ces dernières ne convergeront qu'avec le développement d'internet et des NTIC.

1.2 La pyramide Données - Informations - Connaissances

Dans un article publié en 2007 [7], Zins cherche à définir plus précisément les concepts de *données*, d'*information*, et de *connaissance*, sur la base d'un certain nombre de définitions proposées dans la littérature académique. Parmi les personnes qu'il interroge et dont il rapporte les définitions, celle de Quentin L. Burrell, de l'*Isle of Man International Business School* me semble la plus pertinente :

Data *are the basic individual items of numeric or [of] other information, garnered through observation ; but in themselves, without context, they are devoid of information.*

Information *is that which is conveyed, and possibly amenable to analysis and interpretation, through data and the context in which the data are assembled.*

Knowledge *is the general understanding and awareness garnered from accumulated information, tempered by experience, enabling new contexts to be envisaged.*

Concrètement, *I-L-_-P-L-E-U-V-R-A-_-D-E-M-A-I-N*, sont une succession de caractères qui constituent des données. En tant que phrase écrite en français, "*il pleuvra demain*" est porteur d'une information, ou d'un message, qui peut être véhiculé d'une manière indépendante des données précédemment mentionnées. Ainsi, cette information peut-être véhiculée par la parole. La succession de sons permettant son articulation constituent des données qui n'ont rien à voir avec la succession de caractères dactylographiés ; cependant, une fois que les deux messages sont isolés, ils apparaissent comme identiques. Malgré tout, l'extraction de cette information n'en fait pas une connaissance. En effet, un automate élaboré (un programme informatique) peut tout à fait convertir des données de type sonores en une séquence de caractères alphanumérique, tout en en préservant le message. La connaissance est la compréhension profonde du contenu de ce message, ainsi que ses conséquences.

Ces définitions sont intéressantes à plusieurs titres. Tout d'abord, en raison de l'apparition du terme de "contexte", dans chacune d'elles. Ensuite, en raison du mot "information", qui apparaît dans la première définition (*Data*), et qui doit être compris au sens étymologique du terme, c'est à dire au sens de forme. Enfin, en raison de la nature hiérarchique du lien explicite entre ces trois définitions. Tout cela permet d'aboutir à une vision globale du processus d'inférence cognitive, telle que représentée dans la Fig. 1.1 : la hauteur de la pyramide représente le niveau sémantique, et la largeur de la pyramide à un niveau donné représente le volume ou la quantité de matière correspondante (par exemple, l'espace mémoire qu'il faut pour stocker cela). Ainsi, le processus d'inférence consiste à partir de la base, constituée d'un

très gros volume d'éléments, sans aucune signification ou structure associé, avec une sémantique nulle, pour aboutir au sommet, à une quantité d'éléments extrêmement réduite, mais munie d'une sémantique très forte. La couche la plus basse de la pyramide est constituée de données. La pointe est la connaissance que l'on cherche à inférer. Toutes les couches intermédiaires correspondent à des niveaux d'informations intermédiaires. La définition de Burrell est donc cohérente avec l'étymologie : une information est un message qui prend forme dans des données. Ce message peut être transformé et peut prendre plusieurs formes différentes. Comment obtient-on de l'information ? En assemblant des données grâce à des contextes. Ainsi, quand on fournit le contexte suivant "ces symboles sont des caractères alphanumériques, et ils sont assemblés en une phrase écrite en français", il devient possible de passer de la donnée à l'information. De même, l'accumulation d'informations et leur prise en compte dans un contexte globale permet de définir des connaissances. Il reste donc à définir ces "contextes". Ceux-ci sont de plusieurs natures :

- Tout d'abord, il est explicitement dit que les connaissances constituent autant de contextes : le fait de comprendre le français me permet de déchiffrer la phrase *I-L-_-P-L-E-U-V-R-A-_-D-E-M-A-I-N*.
- Ensuite, les données peuvent servir de contextes à d'autres données : pris de manière isolée, des cercles ou une barre verticale peuvent avoir de multiple sens ; en revanche, ensemble et correctement positionnés, ils s'interprètent comme le nombre "100".
- Enfin, les informations servent aussi de contextes. Ainsi, le seul fait de savoir qu'un dessin est un symbole alphanumérique est en soit une information qui permet de l'interpréter.

Cela nous indique donc que les couches "données" et "connaissances" sont infiniment fines sur la pyramide : le seul fait de comprendre qu'une donnée est ce qu'elle est, permet une élévation sémantique. De même, toute information peut être contextualisée à un niveau sémantique plus élevé, et la frontière avec la connaissance n'est pas claire. Finalement, les contextes[2] permettant l'élévation sur la pyramide ne sont finalement que des données, des informations ou de la connaissance ; ce qui est représenté sur la pyramide par les flèches descendantes.

Dans ce chapitre, je décris de manière non exhaustive un certain nombre de disciplines que j'ai étudiées ou enseignées, et qui constituent certaines des briques de la science des données ; puis, je les positionnerai sur la pyramide de la Fig. 1.1, afin d'illustrer la démarche scientifique que je mets aujourd'hui en œuvre.

2 Les statistiques

Historiquement, les statistiques sont la branche la plus ancienne de la science des données. En France, et contrairement à la tradition anglo-saxonne, les statistiques sont souvent perçues comme une branche des mathématiques. Je n'adhère pas à ce point de vue. En mathématique, la démarche expérimentale est absente. Au contraire, la statistique constitue une discipline expérimentale, dont la matière

2. Ces contextes peuvent être extérieurs, et ne doivent pas forcément être vus comme des rétroactions. Cependant, dans la cadre de la maturation du cerveau humain, c'est précisément ce qui se passe, soulevant la question du commencement de cette boucle vertueuse.

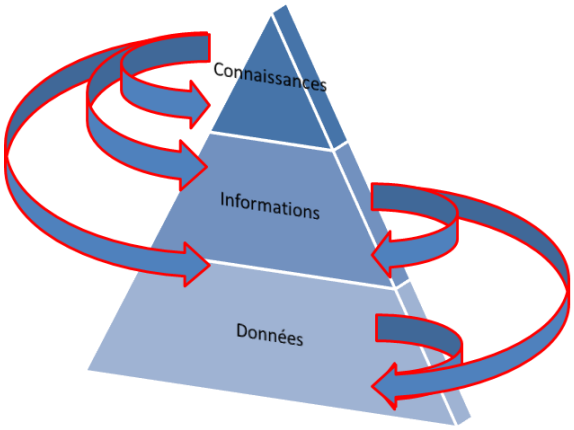

FIGURE 1.1 – *La pyramide DIC - Données, Informations, Connaissances.*

est constituée des observations collectées sur une population. Faisons un parallèle avec la physique : elle est une science expérimentale, visant à décrire le monde qui nous entoure par un modèle qui pourra être utilisé pour réaliser des prévisions. Ce modèle est formalisé par le langage mathématique ; mais il existe aussi une discipline qui constitue un pont entre les mathématiques et la physique : la physique mathématique. Il en est de même pour la statistique ; cependant, en France, la statistique mathématique et la statistique sont souvent assimilées.

Cette section ne constitue pas un cours de statistique (en voici un : [8]). Il s'agit avant tout de décrire l'évolution de la discipline et de son "état d'esprit", afin de montrer qu'elle converge avec d'autres pour constituer la *science des données*.

2.1 L'approche fréquentiste

Sous sa forme la plus simple, la formalisation probabiliste n'est pas nécessaire. En effet, les *statistiques descriptives* les plus élémentaires se suffisent du dénombrement : le comptage des occurrences (nombres de filles et de garçons dans une classe, de boules de couleur dans une urne, d'animaux malades dans un troupeau) permet de décrire une population selon une ou plusieurs modalités. Même si au fil du temps, les outils ont évolués et se sont raffinés, le concept des statistiques descriptives est resté inchangé : une *population*, constituée d'*individus* est décrite au moyen d'une ou plusieurs *variables statistiques* (par exemple, le taille, le poids et le sexe).

Cependant, dans cette approche il n'est pas possible d'effectuer des prévisions (le sujet des *statistiques inférentielles*) : il est possible de décrire la répartition en taille des membres d'une famille ou d'un village, mais pas de prédire si un enfant grandira beaucoup ou pas. Cette prédiction nécessiterait soit de connaître avec précision l'amplitude de tous les facteurs pouvant influencer cette variable statistique (la taille), soit d'accepter l'existence d'aléas non maîtrisables pouvant la modifier radicalement. Alors que la physique newtonienne construit ses modèles de la première manière, la statistique se fonde sur la seconde, en acceptant l'existence du hasard.

Le principal problème est qu'il est impossible de donner une définition mathématique du hasard. En effet, ce terme ne semble fait que pour désigner l'expression de

notre ignorance des causes exactes de certains évènements. Il est cependant possible de supposer que le hasard a été autant à l'œuvre dans le passé qu'il le sera à l'avenir. Ainsi, s'il y a beaucoup de blonds dans la famille depuis des générations, il semble naturel de supposer que le bébé à naître le sera aussi : en l'absence de définition mathématique du hasard, le comptage et les proportions (bref, le dénombrement déjà utilisé en statistique descriptive) offre donc une petite alternative.

Malheureusement, il y a des choses qui ne se comptent pas. Par exemple, lors d'un tir d'artillerie sur le champ de bataille, quelle est la probabilité de toucher les armées alliée ou adverse ? Face à de telles questions, le dénombrement ne suffit plus ; et il a fallu développer la *théorie de la mesure*, ce à quoi se sont attachés des mathématiciens comme Borel, Lebesgue ou Kolmogorov. L'élément central de la théorie de la mesure est ce qu'on appelle la *variable aléatoire*. On peut voir cela comme le pendant probabiliste de la *variable statistique*. Comme défini plus haut, une variable statistique est simplement le critère d'étude d'une population. En revanche, une variable aléatoire est un objet mathématique un peu compliqué, qui contrairement à ce qu'indique son nom, n'est pas une variable. Il s'agit d'une application, qui à toute valeur possible (aussi appelée *éventualité*) que peut prendre la variable statistique (par exemple "55kg" pour la variable statistique "poids" ou "bleu" pour la variable statistique "couleur"), va y associer, dans un espace mathématique un peu abstrait, nommée l'*espace probabilisé*, un ensemble dont on peut mesurer la taille. Dans le cas d'une variable statistique à valeur discrète, cette taille correspond à la cardinalité de l'ensemble associé ; on retombe donc sur du dénombrement. En revanche, cette taille se généralise à une longueur, une surface, un volume ou un hypervolume pour des variables statistiques continues, mono- ou multivariées. Bien sûr, au moment de définir cette variable aléatoire, c'est-à-dire en attribuant à chaque éventualité, un ensemble d'une taille spécifique, nous allons faire attention d'attribuer un ensemble plus grand aux éventualités qui auront plus de chances de se réaliser ; et inversement, un ensemble plus petit aux éventualités qui auront moins de chances de se réaliser.

Grâce à cette astuce, même si le hasard n'a toujours pas de définition mathématique, il est possible d'en supposer l'existence, et d'en décrire formellement les conséquences : certaines éventualités deviennent plus probables que d'autres, puisqu'une variable aléatoire y associe un ensemble plus grand qu'à d'autres.

2.2 Quelques notions pour apprivoiser le hasard

Cette introduction (artificielle) du hasard nous permet de définir des lois de comportements pour les phénomènes aléatoires générant les données observées.

Lois probabilistes Au-delà de la loi **Normale** (ou **Gaussienne**), il existe un très grand nombre de lois moins connues des biologistes, permettant de modéliser des processus différents (Par exemple, la loi de **Poisson** pour le comptage, les lois **Géométrique** et **Exponentielle** pour les phénomènes sans mémoire, etc.). Il est particulièrement intéressant de constater qu'à partir de la loi **Uniforme** sur l'intervalle [0,1], il est possible de concevoir un programme informatique qui simule un très grand nombre de lois, telles que celles mentionnées plus haut. La compréhension complète d'une telle procédure de simulation serait d'une grande aide pour bien des biologistes, car il y a souvent une analogie forte entre cette procédure, et la réalité des phénomènes qu'ils étudient.

Modèle de mélange Une fois que ces lois sont acceptées, il convient de supposer qu'une population statistique relève d'un tirage aléatoire selon une ou plusieurs lois particulières. Dans le cas le plus simple, il n'y a qu'une seule loi, mais ce n'est pas toujours le cas. Regardons la population humaine. Les tailles respectives des hommes et des femmes se modélisent très bien par deux lois Normales (centrées sur 1m75 et 1m65 environ). Cependant, la population entière n'est pas bien modélisée par une unique loi Normale. Heureusement, il est possible de créer une nouvelle loi, comme un **modèle de mélanges** : si π représente la proportion d'hommes ($\approx 48\%$), et si $\mathcal{N}_h$ et $\mathcal{N}_f$ représentent respectivement les lois des tailles des hommes et des femmes, alors il est possible de définir $\pi \cdot \mathcal{N}_h + (1-\pi) \cdot \mathcal{N}_f$, la loi de la taille de tous les humains.

Population i.i.d Grâce à cette astuce, il est généralement possible, quel que soit la complexité de la population étudiée, de supposer que tous ses individus correspondent à des tirages aléatoires d'une même unique loi. Par ailleurs, on supposera généralement ces tirages indépendants, c'est-à-dire que les caractéristiques d'un individu n'ont pas de conséquences ou pas de lien sur les caractéristiques d'un autre individu, ce qui dans bien des cas relève de l'évidence. Dans ce contexte, on considérera qu'une population est un ensemble d'individus i.i.d., pour *indépendants et identiquement distribués*. En pratique, comme il est particulièrement difficile de réaliser une étude statistique en dehors de ce contexte, l'hypothèse i.i.d. est fondamentale.

Estimation Une fois qu'une population i.i.d. a été identifiée, et qu'une famille de lois y a été associée (comme par exemple le modèle de mélange $\pi \cdot \mathcal{N}_h + (1-\pi) \cdot \mathcal{N}_f$), on cherche a estimer les paramètres de cette loi (dans l'exemple, les moyennes μ_h et μ_f, les écarts-types σ_h et σ_f, ainsi que la proportion π). Le terme "estimer" doit être compris au sens de "deviner". Les valeurs de ces paramètres peuvent être inconnues, mais il est possible, grâce à diverses méthodes, de trouver une valeur qui sera relativement proche. On note généralement ces valeurs estimées avec un chapeau : $\hat{\mu}_h$, $\hat{\mu}_f$, $\hat{\sigma}_h$,$\hat{\sigma}_f$ et $\hat{\pi}$.

Compromis biais-variance La littérature statistique est remplie de méthodes d'estimation, fournissant des résultats de qualité variable en fonction des problèmes. Généralement, on évalue la qualité d'un estimateur en fonction d'une part de son **biais**, et d'autre part, de sa **variance** (cf. Fig. 1.2). Le biais fait référence au fait que l'estimation produit une erreur systématique. Ainsi, pour un paramètre θ, nous aurons $|\hat{\theta} - \theta| > 0$. Au contraire, la variance traduit la *stabilité*, ou au contraire la *volatilité* de l'estimateur : si l'on prend 2 échantillons d'une même population, une méthode d'estimation stable devra fournir des valeurs $\hat{\theta}_1$ (échantillon 1) et $\hat{\theta}_2$ (échantillon 2) proches, alors qu'une méthode volatile risque de nous fournir des valeurs différentes. Malheureusement, la diminution du biais conduit généralement à une augmentation de la variance (et inversement), rendant nécessaire la recherche d'un compromis.

2.3 Le test d'hypothèse

L'estimation de paramètres permet d'aller au-delà de la simple description d'une population observée, afin de généraliser sa caractérisation à toute autre population

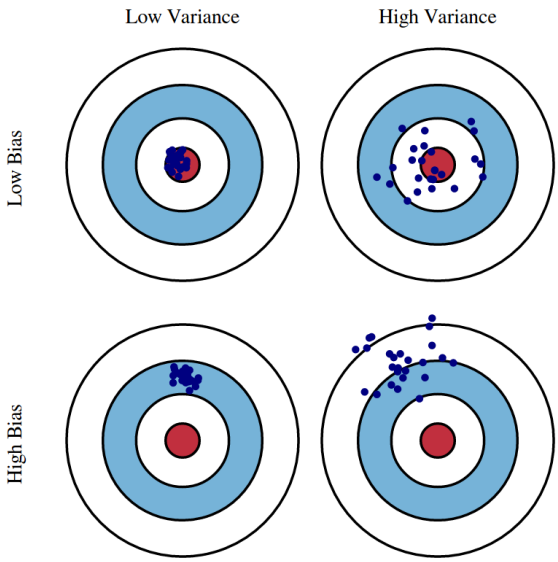

FIGURE 1.2 – *Illustration schématique du biais et de la variance - tiré de [9].*

suivant la même loi : nous passons des *statistiques descriptives* aux *statistiques inférentielles*. Une notion centrale en statistique inférentielle est celle de *test d'hypothèse*, souvent mal comprise, et dont je vais donner un petit aperçu.

Contrôler l'intérêt d'une découverte

Tout commence avec la notion de *découverte*. Il s'agit d'un phénomène saillant, original et différent de ce que l'on pourrait attendre, et qui en raison de tout cela devient intéressant, pour publication par exemple. De manière classique, un chercheur va être confronté à une *découverte potentielle*, c'est-à-dire un phénomène qui est possiblement une découverte, mais peut-être pas ; le but du test d'hypothèse est simplement d'aider le chercheur à déterminer si cette découverte potentielle est une *vrai découverte* (intéressante ou digne de publication), ou au contraire, s'il s'agit d'une *fausse découverte* (le phénomène observé n'est en réalité pas si saillant ou original que cela). De manière assez naturelle, on définit l'intérêt ou la saillance d'une découverte en fonction de la manière dont elle se différencie d'observations majoritaires et inintéressantes qui constituent la référence, ou la norme. Le but du test d'hypothèse est donc simplement de mesurer cet écart à la norme. Dans le langage des statistiques, la norme est appelée l'*hypothèse nulle* (noté $\mathbf{H_0}$), et cette dernière se définie tout simplement en observant un grand nombre de non-découvertes, qui ont un comportement similaire, et en proposant une loi pour celle-ci : on dit que ces non-découvertes suivent la *distribution nulle*.

Exemple 1 (Identification de peptides) *Un protéomicien cherche un peptide particulier* **P** *dans une grande masse de spectres MS/MS, généré au cours d'une même expérience. Il s'intéresse à un score S mesurant d'adéquation entre le spectre théorique de* **P** *et chacun des spectres MS/MS qu'il a obtenu (par exemple le score MOWSE utilisé dans le moteur d'identification de peptides Mascot [10]). Bien sûr, la plupart des spectres ne correspondent pas à* **P**, *de sorte que S sera très faible,*

mais d'autres seront plus élevés. Face au score le plus élevé, le protéomicien pensera légitimement que si $\mathbf{P}$ *est présent dans l'échantillon, alors, c'est ce spectre qui permet de l'identifier. C'est donc cela qui constitue sa découverte potentielle à tester. Cependant, il est difficile de confirmer (vraie découverte) ou d'infirmer (fausse découverte) cette découverte potentielle, dans la mesure où il est difficile d'interpréter le score (par exemple 42) sans aucune valeur de référence (échelle de 0 à 100). C'est là qu'intervient la norme, ou l'hypothèse nulle : pour cette expérience, le protéomicien va comparer la distribution de* S *sur tous les "mauvais" spectres, et ainsi constituer la distribution nulle. Dès lors, il pourra voir dans quelle mesure le meilleur score s'en éloigne.*

Voilà pour l'objectif du test d'hypothèse. Dans l'idéal, nous souhaiterions que le test d'hypothèse prenne en entrée, (1) le score à tester, et (2) la distribution nulle ; et fournisse en sortie, la probabilité que la découverte potentielle soit une vraie découverte. Cependant, nous verrons que ce n'est malheureusement pas le cas.

Interpréter la probabilité critique

Si nous notons s la valeur précise qu'a prise S pour la découverte potentielle à tester, et H_0 la variable binaire qui indique si oui ou non la norme $\mathbf{H_0}$ est respectée, nous souhaitons simplement pouvoir calculer $\mathbb{P}(H_0 = 0|S \geq s)$: il s'agit de la probabilité que la découverte potentielle ne corresponde pas à la norme (et donc qu'il s'agisse d'une vraie découverte), sachant que son score S vaut exactement s, ou tout score encore meilleur (indiquant un écart à la norme encore plus fort). Pour calculer cette probabilité, nous pouvons commencer par remarquer que :

$$\mathbb{P}(H_0 = 0|S \geq s) = 1 - \mathbb{P}(H_0 = 1|S \geq s)$$

puis appliquer le théorème de Bayes, qui stipule que pour deux évènements A et B :

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(B|A) \cdot \mathbb{P}(A)}{\mathbb{P}(B)} \tag{1.1}$$

Cela nous donne donc :

$$\mathbb{P}(H_0 = 0|S \geq s) = 1 - \left(\mathbb{P}(S \geq s|H_0 = 1) \cdot \frac{\mathbb{P}(H_0 = 1)}{\mathbb{P}(S \geq s)}\right) \tag{1.2}$$

Le second membre de cette équation comporte le produit de deux termes : tout d'abord $\mathbb{P}(S \geq s|H_0 = 1)$, puis $\mathbb{P}(H_0 = 1)/\mathbb{P}(S \geq s)$. Le premier terme se calcule très facilement. Il s'agit simplement d'estimer la probabilité qu'une fausse découverte donne par hasard un résultat au moins aussi élevé que s. Pour cela, il suffit d'évaluer pour $\mathbf{H_0}$ la proportion de l'aire sous la courbe à droite de la valeur s, comme cela est représenté sur la Fig. 1.3. En revanche, on ne sait pas calculer le second terme [3].

3. En effet, $\mathbb{P}(H_0 = 1)$ correspond à la probabilité qu'une découverte potentielle quelconque soit fausse (en gros, donnée par la proportion de vraies et fausses découvertes parmi toutes les découvertes potentielles possibles), et $\mathbb{P}(S \geq s)$ correspond à la probabilité globale (aussi bien sur les vraies que les fausses découvertes) d'un score au moins supérieur à s. Cependant, il existe des estimateurs pour cette quantité, mais ceux-ci sont particulièrement volatiles, et ne peuvent être utilisés que dans des conditions particulièrement strictes. Le plus utilisé d'entre eux est le *local FDR* [11], dont l'étude sort largement du cadre introductif de ce chapitre.

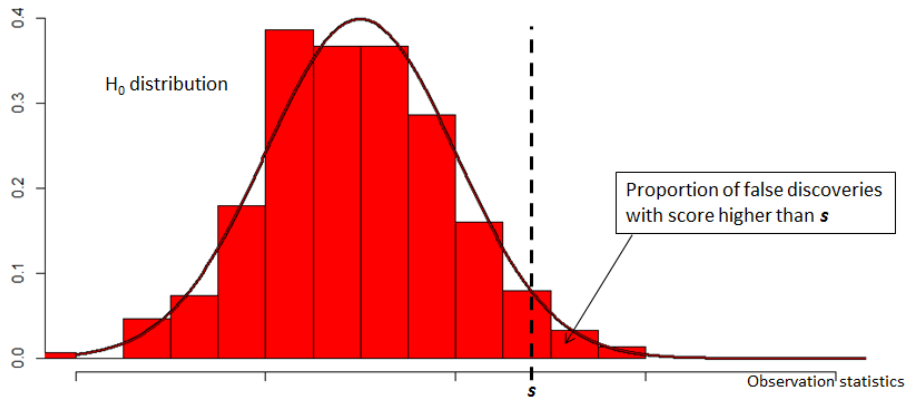

FIGURE 1.3 – *Illustration graphique de l'évaluation de* $\mathbb{P}(S \geq s|H_0 = 0)$.

Finalement, il n'est pas possible de calculer la grandeur attendue par le praticien, à savoir la probabilité $\mathbb{P}(H_0 = 0|S \geq s)$ que la découverte soit vraie, et le test d'hyptohèse ne permet que d'obtenir la probabilité $\mathbb{P}(S \geq s|H_0 = 1)$, classiquement appelé *probabilité critique*, ou par anglicisme, p-valeur. Le praticien doit donc prendre sa décision, non pas sur la probabilité $\mathbb{P}(H_0 = 0|S \geq s)$ dont il a besoin, mais sur une autre probabilité, $\mathbb{P}(S \geq s|H_0 = 1)$, dont l'interprétation est différente.

Quelle interprétation doit-on donner à la probabilité critique ? Il s'agit de la probabilité d'observer un score au moins aussi extrême que s (par rapport à notre norme), en supposant que cette norme est respectée. Naturellement, si cette probabilité est particulièrement faible, disons 0.001, cela voudra dire qu'il y a très peu de chances pour que notre découverte potentielle donne un score si élevé si elle n'est pas "vraie". On aura donc tendance à supposer qu'il s'agit d'une vraie découverte. Cependant, la probabilité qu'il s'agisse d'une vraie découverte est inconnue. Une erreur fréquente consiste à croire qu'avec une probabilité critique de 0.001, il y a une chance sur mille que la découverte potentielle soit une fausse découverte, et donc 999 chances sur mille qu'elle soit une vraie découverte. Cela est faux. En effet, dans le cas d'une hypothèse nulle très probable (et d'une probabilité très faible de découverte), la p-valeur ne signifie rien en ce qui concerne la probabilité de découverte, comme l'illustre l'exemple suivant :

Exemple 2 *Supposons que nous ayons 3000 découvertes potentielles, mais supposons que les découvertes réelles soient extrêmement rares (de l'ordre d'une sur un million, c'est-à-dire qu'un des termes inconnus dans l'Equation 1.2, à savoir* $\mathbb{P}(H_0 = 1)$*, soit très proche de 1). Sur les 3000 découvertes potentielles, il est probable d'obtenir 3 fois un score tellement extrême qu'il ne se produit pour une fausse découverte que dans 1 cas sur 1000. Autrement dit, nous aurons 3 découvertes potentielles avec une p-valeur de 0.001. Il n'est pourtant pas cohérent de supposer que chacune de ces 3 découvertes potentielles a 999 chances sur 1000 d'être une vraie découverte, si les vraies découvertes sont rares de l'ordre d'une pour un million.*

La difficulté d'interprétation de la p-valeur est donc liée à la confusion entre $\mathbb{P}(A|B)$ et $\mathbb{P}(B|A)$. De nos jours, de nombreux logiciels grand-publics permettent d'appliquer facilement un très grand nombre de tests statistiques à n'importe quel

jeu de données. De plus, tous ces tests et les distributions nulles associées sont décrits en détails sur Wikipédia [12]. Finalement, la compréhension profonde de la p-valeur est le seul élément vraiment problématique [13–16], sur lequel les enseignements universitaires de statistiques appliqués devraient se concentrer.

Grande dimensionnalité et contrôle des fausses découvertes

Au cours du test d'hypothèse, nous pouvons déjà être confrontés aux difficultés de travailler en grande dimensionnalité. En effet, dans l'exemple 2, nous avons supposé qu'au lieu de ne considérer qu'une seule découverte potentielle, nous avions à notre disposition 3000 d'entre elles. C'est un cas fréquent en protéomique, où chacun des milliers de spectres peptidiques d'une analyse MS/MS est une découverte potentielle.

De manière plus générale, si l'on possède un ensemble de N découvertes potentielles triées par p-valeur croissante, et que nous ne retenions que les $n < N$ plus faibles, combien risque-t-il d'y avoir de fausses découvertes parmi les n ? Ce type de question a révolutionné les statistiques dans le milieu des années 90 : de nombreux travaux théoriques (résumé par Bradley Efron dans [17]) se sont penchés sur l'estimation d'une telle proportion (que l'on appelle le *False Discovery Rate*) ou plus généralement, ont proposé des méthodes de correction de p-valeurs pour tenir compte de la multiplicité des tests. En parallèle, ces travaux trouvaient un écho fort en génomique (notamment sous la direction de Tibshirani [18]) : à partir de problèmes de sélections de gènes biomarqueurs, les questions se sont généralisées, ouvrant la voie au paradigme "*small n, large p*", caractérisant les problèmes de statistiques où le nombre de variables est grand devant le nombre d'individus de la population, que nous retrouverons au paragraphe 5.2.

3 L'intelligence artificielle

En parallèle de l'approche fréquentiste des probabilités, très intuitive, d'autres approches, dites subjectives, se sont développées avec une filiation plus forte avec l'intelligence artificielle (IA) qu'avec les statistiques. Voici donc quelques généralités sur l'IA, suivi de notions de probabilités subjectives et d'inférence bayésienne, que j'étendrai ensuite vers la théorie des fonctions de croyance [4].

3.1 Les deux principaux courants de l'IA

Comme son nom l'indique, l'objectif initial de l'intelligence artificielle était de réaliser des machines pensantes, ou du moins capables de reproduire les comportements intelligents observables. Cela s'est fait de deux manières différentes amenant à deux courants distincts :

La bioinspiration où il s'agit de de copier les comportements intelligents ayant naturellement émergés : par exemple, reproduire des réseaux de neurones artificiels un peu schématisés et simplifiés [19], ou programmer des agents simples

4. En toute objectivité, cette théorie est trop confidentielle pour mériter d'être introduite dans un résumé de 30 pages sur l'extraction de connaissances ; et je ne m'attarde dessus qu'en raison des travaux que je lui ai consacrés ces dernières années.

dotés de capacités d'interaction, afin de permettre l'émergence d'une intelligence collective, comme avec les colonies de fourmis artificielles [20]. *In fine*, tous ces algorithmes permettent la définition implicite d'une fonction de décision, et peuvent être tout à fait interprétés sous l'angle de l'apprentissage automatique (cf. Sec. 5.3) ; cependant, une analogie naturelle est utilisée pour la définition de celle-ci plutôt que la *minimisation du risque empirique*.

La logique Ce courant a pour objectif de décomposer analytiquement les différentes étapes d'un raisonnement cognitif afin de le reproduire, ou simplement de le décrire avec un formalisme adapté. À ce titre, une difficulté réelle concerne l'encodage de notions qualitatives et vagues dans une machine ne manipulant que des données binaires. C'est dans ce contexte que la théorie des ensembles flous [21], ou que les probabilités subjectives [22] sont apparues.

N'étant pas familier du premier courant, je me concentre dans la suite sur le second. Par ailleurs, le premier a de nos jours complétement fusionné avec celui de l'apprentissage automatique [23], du traitement du signal [24], des méta-heuristiques pour l'optimisation [25] de sorte que j'utiliserai le terme générique "intelligence artificielle" pour désigner le second exclusivement.

Le principe de la logique floue est qu'une proposition n'est plus soit vraie, soit fausse, mais qu'elle peut correspondre à un état de véracité intermédiaire. Le principe des probabilités subjectives est assez proche : il y a beaucoup d'évènements qui sont uniques et qui ne vont pas se reproduire (par exemple, l'élection présidentielle de 2022), et pour lesquels il semble qu'il n'y ait aucun sens à parler de probabilité de réalisation, puisqu'il n'est pas possible de répéter cet évènement afin de faire des statistiques. En revanche, tout le monde est d'accord pour considérer qu'il est plus probable qu'à l'issue de l'élection de 2022, un politicien soit élu, plutôt qu'un artiste. La probabilité ne doit donc pas nécessairement correspondre à une fréquence, mais peut, plus généralement correspondre à un score de confiance dans la réalisation de tel ou tel évènement. Ainsi, un score de confiance de 1 indique que l'on est sûr que l'évènement va se produire, alors qu'un score de 0 indique l'on est sûr qu'il ne va pas se produire ; l'entre-deux permettant de graduer le niveau de confiance.

Finalement, une probabilité subjective est quelque chose d'assez simple : elle est définit comme une probabilité classique, sur un espace mesurable ; chaque évènement est muni d'une probabilité comprise entre 0 et 1 ; et la somme des probabilités associées à des évènements exclusifs et exhaustifs (c'est-à-dire qu'un et un seul d'entre eux se réalisera) vaut 1. Cette définition est naturellement plus générale, puisqu'une confiance "subjective" peut éventuellement être fondée sur des considérations statistiques. Elle peut cependant être fondée sur d'autres choses, comme la déduction finale d'un raisonnement basé sur de la logique floue, ou encore, sur l'avis formulé *ex nihilo* par un expert que l'on interroge. De même, il est possible d'utiliser ce formalisme, non plus pour encoder une confiance, mais une autre notion subjective, comme par exemple une possibilité (un score de 1 indiquant une possibilité totale, mais malgré tout incertaine) [26].

Le subjectivisme ouvre l'idée à la fusion d'informations, pour la simple raison qu'il peut être intéressant de combiner l'expertise de plusieurs personnes ou systèmes afin d'obtenir un résultat plus fin ou plus fiable. C'est notamment pour cela que l'IA, au contraire des statistiques, fait la part-belle aux approches subjectives.

3.2 Les réseaux bayésiens

Il est difficile de résumer ce courant de pensée en une page ([27] est plus exhaustif), tant ses domaines d'applications sont vastes, ses ancrages épistémologiques nombreux, et son rattachement théorique difficile à cerner. En effet, la communauté bayésienne a eu de fortes interactions avec celle de l'intelligence artificielle tout comme avec celle des statistiques, tout en restant indépendante de ces deux-là. À l'origine de l'inférence bayésienne, on trouve bien sûr, le théorème de Bayes, que nous avons rapidement évoqué à l'occasion de l'approche fréquentiste des probabilités inhérentes aux statistiques, et qui pour rappel, permet d'inverser le conditionnement d'une probabilité (cf. Eq. 1.1) ; mais qui donne l'impression trompeuse qu'il est possible d'exprimer une sorte de "causalité probabiliste" [28, 29].

En inférence bayésienne, on s'intéresse à une distribution jointe (c'est-à-dire des distributions multidimensionnelles faisant intervenir plusieurs variables), dont la manipulation permettra de répondre à toutes sortes de questions. Par définition, une distribution jointe peut aussi se formuler à partir d'une distribution conditionnelle,

$$\mathbb{P}(A, B) = \mathbb{P}(A|B) \cdot \mathbb{P}(B), \tag{1.3}$$

sur laquelle le théorème des probabilités totales donne

$$\mathbb{P}(A) = \sum_i \mathbb{P}(A|B_i) \cdot \mathbb{P}(B_i). \tag{1.4}$$

Il devient possible, à partir d'une distribution jointe, de calculer la distribution conditionnelle se rapportant à n'importe quelle variable, puis de l'évaluer pour un évènement particulier. Par ailleurs, dans le cas de plusieurs variables indépendantes, les équations se simplifient puisque le conditionnement mutuel n'a plus d'influence.

Finalement, un réseau bayésien peut être simplement vu comme un moyen de simplifier une distribution jointe faisant intervenir un très grand nombre de variables : dans un premier temps, il s'agit de faire l'hypothèse qu'un certain nombre de variables sont indépendantes entre elles. Ensuite, il s'agit de réécrire la distribution jointe comme un produit de distributions jointes ou conditionnelles (ainsi que quelques sommes, correspondant à l'application du théorème des probabilités totales) relativement compact en raison de toutes les simplifications induites par les hypothèses d'indépendance. Un réseau bayésien est donc avant tout une manière de rendre facilement calculable par une factorisation appropriée une distribution initialement complexe à manipuler.

Le terme de "réseau" provient du fait que si l'on représente les variables par des sommets et les relations de dépendances par des arêtes (l'indépendance étant représentée par une absence d'arêtes), la factorisation de la distribution jointe est encodée sous la forme d'un graphe (comme cela apparaît sur la Fig. 1.4). Par ailleurs, il apparait que ce réseau devient beaucoup plus difficile à manipuler si celui-ci contient des boucles, rendant la représentation graphique nécessaire à l'utilisateur. Enfin, il apparait que tous les calculs susnommés, que l'on peut qualifier de "globaux", puisqu'ils réfèrent à la distribution jointe dans son ensemble, peuvent se ramener à deux calculs "locaux" pour chacun des sommets du réseau [30–32]. Bien qu'il ne s'agisse en pratique que d'une astuce calculatoire, cet aspect "calcul distribué" et l'analogie de surface avec une architecture neuronale qu'il véhicule, ont beaucoup participé à la diffusion de cet outil, comme de sa représentation graphique.

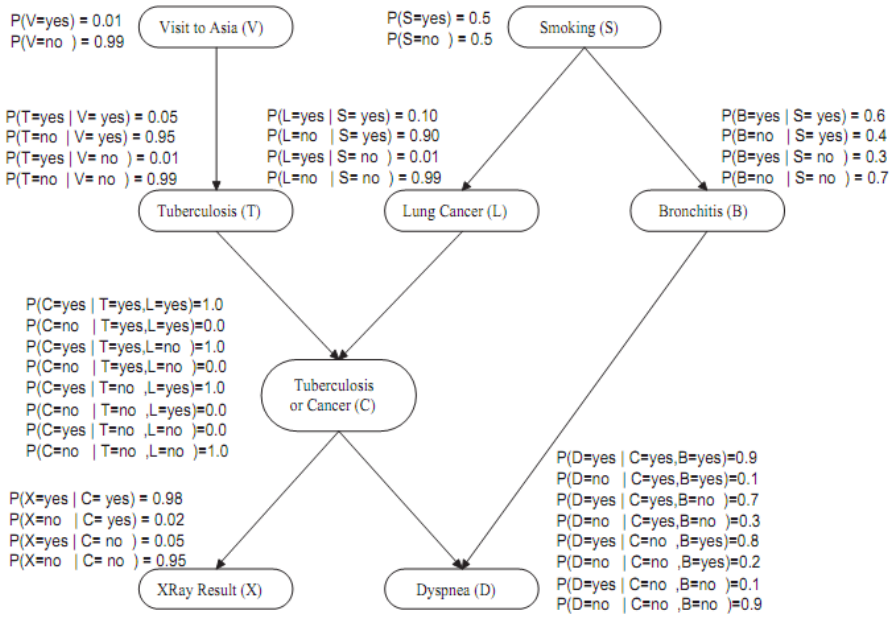

FIGURE 1.4 – *Exemple classique d'un réseau bayésien encodant une distribution de probabilité associée à des variables décrivant l'état sanitaire d'un patient, ses symptômes cliniques et son histoire ; tiré de [33].*

Par ailleurs, comme tout ce qui a été décrit plus haut ne se base que sur le formalisme probabiliste, il n'est pas nécessaire d'avoir une base statistique fréquentielle aux données manipulées, et il est possible d'inclure des probabilités subjectives, afin d'utiliser l'outil comme un moyen de réaliser des déductions probabilistes, voire de combiner des sources d'informations subjectives et complémentaires. Par exemple, le réseau de la Fig. 1.4 modélise le diagnostic d'une maladie respiratoire, en fonction de symptômes cliniques et de l'histoire du patient.

Enfin, les réseaux bayésiens possédant une structure spécifique (celle d'un peigne) sont nommé Modèle de Markov Caché (ou HMM pour *Hidden Markov Model* [34]) et constituent un outil particulièrement adapté au traitement de données séquentielles ou temporelles (notamment l'alignement de séquences ADN et reconnaissance de la parole en sont les applications reines ; cependant il en existe de nombreuses autres, notamment la reconnaissance de gestes ou d'écritures manuscrites, sur lesquelles j'ai travaillé avec de m'intéresser à la protéomique).

3.3 Les fonctions de croyances

Ce sont les travaux qu'Arthur Dempster a mené en statistique dans les années soixante qui ont abouti plus tard à la théorie des fonctions de croyance [35–37]. Initialement, Dempster s'est intéressé à des problèmes de statistiques simples (sur des données catégorielles) où il était possible que les observations soient imprécises. Par exemple, un problème de tirage de boules colorées dans une urne où le statisticien lui-même n'effectue pas les tirages, mais laisse la place à un ami, doté d'un sévère daltonisme. Ainsi, de temps en temps, il indique "la boule est soit rouge, soit verte. Je sais juste qu'elle n'est pas bleue". Au-delà de l'exemple un peu scolaire des boules

et de l'urne, un tel cadre statistique s'applique naturellement aux situations où un capteur est soumis à des conditions rendant fluctuante sa précision. Dans ce contexte, Dempster a fourni plusieurs résultats fondamentaux.

Le premier est qu'un tel problème peut tout aussi bien être traité qu'un problème statistique classique, même si les résultats seront d'autant moins précis que les observations le sont aussi. En pratique, plutôt que de considérer l'univers (l'ensemble des observations possibles, noté Ω, tel que par exemple les couleurs des boules de l'urne, à savoir $\Omega = \{R, V, B\}$ pour Rouge, Verte et Bleue), il suffit de considérer l'ensemble des parties non-vides de cet univers (l'ensemble de toutes les combinaisons d'observations possibles) : $\mathcal{P}(\Omega) = \{\{R\}, \{V\}, \{B\}, \{R \text{ ou } V\}, \{V \text{ ou } B\}, \{B \text{ ou } R\}, \{V \text{ ou } R \text{ ou } B\}\}$. Ensuite, il suffit d'effectuer des statistiques classiques sur ces nouvelles (unions d') observations, puis de remarquer qu'il est possible de calculer des bornes supérieures P^* et inférieures P_* pour les probabilités ou proportions désirées. Par exemple, la probabilité de tirer une boule qui est rouge ou bleue, notée $\mathbb{P}(\{bleu \text{ ou } rouge\})$ est comprise entre :

$$P_*(\{bleu \text{ ou } rouge\}) = m(\{rouge\}) + (\{bleu\}) + m(\{bleu \text{ ou } rouge\}) \tag{1.5}$$

et

$$\begin{aligned} P^*(\{bleu \text{ ou } rouge\}) &= m(\{rouge\}) + (\{bleu\}) + m(\{bleu \text{ ou } rouge\}) \\ &\quad + m(\{vert \text{ ou } bleu\}) + \{vert \text{ ou } rouge\} \\ &\quad + \{vert \text{ ou } rouge \text{ ou } bleu\} \\ &= 1 - m(\{vert\}) \end{aligned} \tag{1.6}$$

ou m (nommée la *masse de croyance*) est une distribution de probabilités [5] sur $\mathcal{P}(\Omega)$.

Le second résultat fondamental de Dempster est qu'il est possible de considérer plusieurs observateurs imprécis, et de combiner les masses de croyance qu'ils génèrent, afin de "recouper leurs informations", et ainsi gagner en précision. Cela ouvre directement la porte au subjectivisme, c'est donc naturellement que son troisième résultat d'importance est qu'il est possible de généraliser l'inférence bayésienne décrite plus haut à des situations d'observations imprécise. Ainsi, en pratique, il devient possible de remettre en cause le postulat selon lequel une absence d'information ($m(\{vrai \text{ ou } faux\}) = 1$) et l'information de l'équiprobabilité ($m(\{vrai\}) = m(\{faux\}) = 0.5$) sont équivalents : maintenant, il est possible de quantifier l'incertitude relevant de l'imprécision et celle relevant de l'aléas de manières différentes. Au travers un tel formalisme plus raffiné, il est naturel d'attendre des inférences plus justes, et cette attente a été la cause de très nombreux développements récents de la théorie des fonctions de croyance.

À la suite des travaux de Dempster, de nombreux autres auteurs ont développé des outils supplémentaires permettant de manipuler des fonctions de croyance, jusqu'à aboutir à une théorie complète (parfois nommée théorie de Dempster-Shafer en référence à ses deux principaux promoteurs). Cependant, de nos jours, peu de problèmes de statistique se focalisent sur des données catégorielles avec observations imprécises, alors qu'au contraire, ils sont fréquents en fusion de données. Finalement, alors qu'elles se sont initialement développées dans un contexte statistique, les fonctions de croyance sont maintenant principalement un outil relevant de l'IA.

5. Notons que P_* et P^* ne peuvent être assimilés à des probabilités sur Ω (notamment ces fonctions ne sont pas additives sur cet espace).

4 Traitement du signal et analyse harmonique

Intéressons-nous maintenant à un domaine un peu plus technique, et dont le lien avec la science des données et les problèmes d'inférence de connaissances n'est pas immédiat. Cependant, celui-ci apparaîtra plus tard, au moment où nous ferons le lien avec les questions statistiques évoquées plus haut. Tout d'abord, nous allons passer en revue les *opérateurs intégraux* : il s'agit d'une classe assez large permettant d'instancier de nombreuses transformations, comme par exemple la transformée de Fourier, et dont l'expression générale est donnée par l'équation suivante,

$$T_k[f](y) = \int_{\mathcal{X}} k(y,x) f(x) d\mu(x) \tag{1.7}$$

que je tâcherai de démystifier. Ensuite, nous nous focaliserons sur ce qu'on appelle le noyau de l'opérateur, à savoir la fonction k, impliquée dans l'Eq. 1.7 ; puis plus précisément, sur une classe particulière de noyaux, qui de par leurs propriétés sont attrayants pour la science des données.

4.1 Opérateur intégral

Transformée de Fourier

Le principe de la transformée de Fourier [24] est généralement connu des protéomiciens, celle-ci étant régulièrement utilisée dans des spectromètres de masse : il s'agit d'analyser un signal $f(t)$, initialement décrit au moyen d'une fonction f de la variable temporelle t, afin de le redécrire différemment, au moyen d'un ensemble d'harmoniques [6] ; ensuite, il est possible de déduire des harmoniques, la masse des ions analysés par le spectromètre. Formellement, la transformée de Fourier $T_F[f]$ de f est une fonction définie par :

$$T_F[f](\nu) = \int_{t=-\infty}^{+\infty} e^{-i\nu t} f(t) dt \tag{1.8}$$

Il s'agit donc de l'intégrale du produit de deux fonctions : tout d'abord, le signal d'intérêt $f(t)$, dépendant du temps ; ensuite, une autre fonction, impliquant le temps t et la fréquence ν, et dont la forme est $e^{-i\nu t}$. Cette dernière est appelée le *noyau* de la transformée. Notons que le produit noyau-signal n'est intégré que sur le temps, de sorte qu'il est normal d'obtenir comme résultat, une fonction de la fréquence ν.

Analyse temps-fréquence

La transformée de Fourier s'intéresse au signal globalement. Or si celui-ci varie dans le temps (ou dans l'espace, dans le cas d'un signal bidimensionnel comme une image par exemple), il peut être intéressant de ne se focaliser que sur une fenêtre particulière du signal, centrée sur un instant τ, puis de faire varier la valeur de τ afin d'avoir une caractérisation local du signal. Concrètement, pour ce faire, il suffit d'annuler le signal partout, sauf dans la fenêtre de temps que l'on considère. Cette

6. Les harmoniques sont les différentes fréquences composant le signal, que l'on peut séparer les unes des autres, sur le principe des éclairages de boites de nuit, dont les pulsations à différentes fréquences suivent la musique.

fenêtre se modélise facilement par une fonction w (pour "window") qui vaut 1 pour l'intervalle d'intérêt et 0 ailleurs, avec laquelle on va multiplier le signal :

$$T_{FL}[f](\tau, \nu) = \int_{-\infty}^{+\infty} e^{-i\nu t} w(t-\tau) f(t) dt \tag{1.9}$$

Dans le cas d'un signal bidimensionnel, nous aurons :

$$T_{FL}[f](\mathbf{x}, \mathbf{y}, \nu_x, \nu_y) = \int_{x=-\infty}^{+\infty} \int_{y=-\infty}^{+\infty} e^{-i(\nu_x x + \nu_y y)} w(x-\mathbf{x}, y-\mathbf{y}) f(t) dx dy \tag{1.10}$$

avec $(\mathbf{x}, \mathbf{y})$ caractérisant la localité de la fenêtre d'étude, et (ν_x, ν_y) correspondant aux fréquences sur chacune des dimensions de l'image. Finalement, il est aussi possible d'avoir des fenêtres plus ou moins larges, afin de caractériser le signal à différentes échelles, les fenêtres étroites capturant tous les détails, et les fenêtres plus larges se rapprochant d'une description globale du signal : c'est ce que l'on appelle l'analyse temps-fréquence (ou par analogie, l'analyse espace-fréquence).

Ondelettes

La transformée de Fourier peut intuitivement être comprise comme une généralisation des *séries de Fourier* : ces dernières furent développées par Joseph Fourier durant ses travaux sur la propagation de la chaleur : il s'agissait pour lui de modéliser la température à l'extrémité d'une barre de fer, quand on applique une source de chaleur dont la température varie périodiquement au cours du temps à l'autre extrémité. La transformée de Fourier peut être vue comme le cas extrême où la période de variation de la température de la source de chaleur est tellement longue que, finalement, elle n'est plus périodique. Cependant, le lien avec les signaux périodiques est encore visible dans la transformée de Fourier. En effet, le noyau $e^{-i\nu t}$ peut aussi s'écrire sous la forme $\cos(-\nu t) + i \sin(-\nu t)$: les fonctions sinusoïdales sont une *famille génératrice* des signaux périodiques. En effet, il est possible de décomposer tout signal périodique en une série (une somme avec éventuellement un nombre infini de termes) de fonctions sinus et cosinus.

Dans le cas où le signal d'intérêt n'est pas une variation de chaleur au cours du temps, mais la variation d'intensité de la couleur sur une image, le noyau $e^{-i\nu t} = \cos(-\nu t) + i \sin(-\nu t)$ n'a aucune raison de fournir une décomposition du signal qui soit "adaptée". C'est pourquoi, la théorie de Fourier a été généralisée à d'autres noyaux pour donner la *théorie des ondelettes* [24]. Le terme *ondelette* désigne en fait simplement un noyau, localisé dans l'espace ou dans le temps (le noyau inclut donc la fonction fenêtre w) et défini par un niveau de granularité. D'un point de vue imagé, on peut voir le noyau comme une fonction nulle partout à l'exception d'une localisation autour de laquelle le signal prend la forme d'une petite vague. Celle-ci peut être plus ou moins "zoomée" afin de permettre une caractérisation multi-échelle ; et déplacée, afin de caractériser localement le signal en tout point. Formellement, la transformée en ondelette de noyau ψ est définie par :

$$T_{\psi}[f](p, s) = \int_{\mathcal{X}} \psi(p, s, x)^* f(x) d\mu(x) \tag{1.11}$$

Dans cette formule, p et s permettent de caractériser le signal en termes de position et d'échelle (s pour "scale" en anglais) tout comme la fréquence ν et la position

τ de la fenêtre dans le temps pour l'analyse de Fourier temps-fréquence. L'étoile en exposant indique le complexe conjugué [7] ; quant à l'intégrale, si elle prend une forme légèrement différente, c'est simplement parce qu'elle est écrite de manière plus générale. Plutôt que d'intégrer entre $-\infty$ et $+\infty$, éventuellement sur les deux dimensions de l'image, l'intégrale est ici définie sur un domaine quelconque $\mathcal{X}$ (pouvant représenter le temps, la surface d'une image ou quoi que ce soit) qui a la propriété [8] d'être mesurable ; à cet égard $d\mu(x)$ ne représente que des parties mesurables de $\mathcal{X}$ qui sont sommées par l'intégrale.

Apprentissage de dictionnaires

Une fois le principe de l'ondelette établi, il reste encore à choisir la famille de noyaux permettant d'obtenir cette caractérisation multi-échelle (en faisant varier p et s). Naturellement, il n'est pas possible de prendre n'importe quoi. Comme nous en avons eu l'intuition à propos de la transformée de Fourier, une analyse harmonique par un opérateur intégral n'est intéressante que si elle permet de reconstruire le signal d'origine à partir de la description qu'en donne sa transformée ; ainsi tout signal périodique peut être reconstruit à partir de sa série de Fourier. De même, à partir de la transformée de Fourier d'un signal, il est possible de retrouver le signal d'origine, en appliquant la *transformée de Fourier inverse*, dont la formule est particulièrement proche de la "vraie" transformée (attention au signe de l'exponentielle) :

$$f(t) = \int_{-\infty}^{+\infty} e^{+i\nu t} \cdot T_F[f](\nu) d\nu \tag{1.12}$$

Comme expliqué plus haut, pour que cette reconstruction soit possible, il faut que la famille de noyaux soit génératrice des fonctions que l'on cherche à décrire : en d'autres termes, en sommant les individus de cette famille, il est possible de reconstruire n'importe quelle fonction parmi celles que nous avons la prétention de décrire. À titre d'exemple, la famille des Gaussiennes (dont on fait varier la moyenne et la variance pour ajuster les paramètres de position p et d'échelle s) est génératrice des fonctions $\mathcal{C}^\infty$ (fonctions continues, infiniment dérivables et de dérivées successives continues) sur $\mathbb{R}$. Les Gaussiennes peuvent donc servir d'ondelettes. De même pour la distribution dite du *chapeau mexicain*, ou l'ondelette de Haar, adaptée à la nature discrète des images numériques, telles qu'illustrées sur la Fig. 1.5.

Ces ondelettes sont particulièrement "génériques" : il est possible de décrire un très grand nombre de fonctions grâce à elles. Cette généricité se fait bien sûr au détriment de la spécificité. Ainsi, si nous ne nous intéressons qu'à une classe réduite de signaux ayant une forme spécifique, il peut être avantageux de prendre une famille de noyaux moins générique, mais permettant d'obtenir une description plus fine (erreur de reconstruction plus faible) et plus compacte (faisant intervenir moins d'éléments à sommer dans la reconstruction du signal d'origine). Ceci étant d'autant plus vrai que $\mathcal{X}$, le domaine de définition des signaux, est de grande dimension. Les signaux à

7. Un détail mathématique que l'on retrouve sous la forme de la négativité dans l'exponentielle du noyau de Fourier. La compréhension de son origine n'est pas nécessaire à la suite de lecture.

8. Cette notion de mesurabilité n'est pas indispensable. Cependant, nous l'avons déjà abordée lors de la modélisation du hasard en statistique : une variable aléatoire associe à chaque évènement, un ensemble dont on peut mesurer la taille. L'espace probabilisé est lui aussi mesurable.

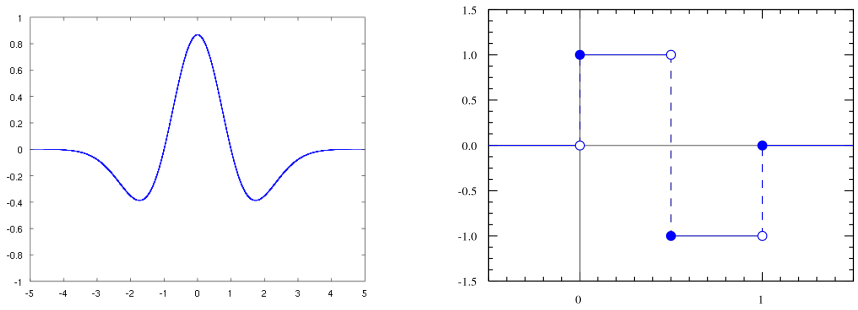

FIGURE 1.5 – *Ondelettes "Chapeau mexicain" et de Haar - tirés de [38, 39].*

décrire constituent eux-mêmes une famille génératrice et très compacte des signaux à décrire, mais la redescription résultante n'aura pas d'intérêt. Il est plus intéressant de trouver un petit sous ensemble de signaux, nommé *dictionnaire*, qui permet de regénérer l'ensemble, à la manière de ce que l'on cherche à faire en compressant l'information[9] : un exemple à la fois parlant et rigoureux consiste en la description d'un flocon de neige à l'aide du motif élémentaire qui le compose, ainsi que des différents facteurs d'échelle à appliquer pour reconstruire la figure complète. Cette recherche d'un petit sous-ensemble permettant la génération des signaux est qualifiée de *description parcimonieuse*, et a trouvé beaucoup d'application en vision par ordinateur [43, 44]. Nous reviendrons plus tard sur ce concept de parcimonie ; pour l'instant, attardons-nous sur un autre élément d'importance, à savoir que l'espace de redescription des données est induit par les données elles-mêmes.

4.2 Noyaux semi-définis positifs

Du signal à la donnée

Jusqu'à présent, quel que soit l'opérateur intégral utilisé, il était particulièrement facile d'identifier les domaines respectivement de définition du signal et de sa description harmonique. Le lien entre ces deux domaines est naturellement fait par le noyau qui est une fonction (généralement noté k pour "kernel" en anglais) définie sur le produit cartésien des deux domaines. Maintenant que nous souhaitons décrire nos signaux à partir d'une famille génératrice elle-même issus de nos signaux, nous ne pouvons plus vraiment faire la différence entre le domaine d'origine et le domaine de redescription. Cela nous amène à chercher à définir les noyaux $k(.,.)$ tels que :

$$T_k[f](x') = \int_{\mathcal{X}} k(x', x) f(x) d\mu(x), \quad x' \in \mathcal{X} \tag{1.13}$$

De tels opérateurs intégraux sont plus intéressants qu'ils n'y paraissent. Par exemple, si le signal f représente une photo, et que k est une gaussienne, alors $T_k[f]$ correspondra à une version floutée de f : il s'agit simplement de la forme analytique du "flou gaussien" disponible dans la plupart des logiciels de retouche, et dont l'intensité

9. L'élévation sémantique de l'interprétation d'un signal, la recherche d'éléments saillants au sein de ce signal, et la compression de celui-ci entretiennent d'ailleurs d'étroites relations [40–42] qui ne seront que brièvement illustrées plus loin, sur la norme JPEG.

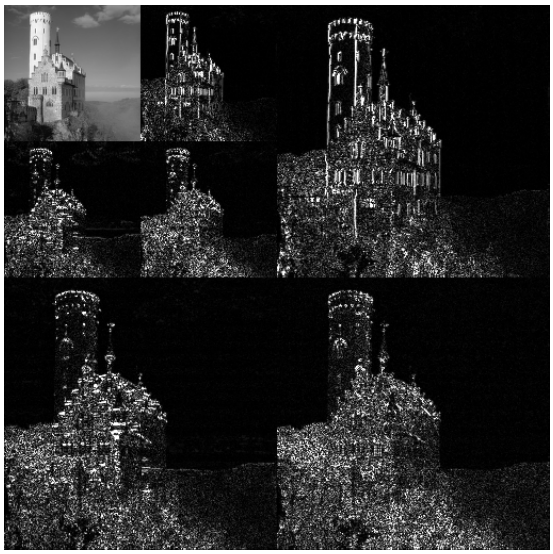

FIGURE 1.6 – *Illustration de l'encodage JPEG2000 - tiré de [45].*

du flou dépend de la variance de la gaussienne utilisée. Si l'on souhaite compresser une grande photographie numérisée, une stratégie particulièrement efficace consiste à réaliser des flous successifs de plus en plus importants. Sur ces-derniers, il y aura peu de détails et beaucoup de pixels de couleurs identiques, fournissant une image lisse et très facile à compresser. Ensuite, il suffit de prendre les flous moins importants, d'y soustraire ce qui est déjà décrit dans les flous plus importants, afin de ne garder que les détails, et de les compresser à leur tour. C'est sur un principe similaire que fonctionne la célèbre norme JPEG2000, illustrée sur la Fig. 1.6.

Il y a des applications encore plus intéressantes pour l'analyse de données. Jusqu'à présent, le domaine de définition du signal $\mathcal{X}$ était le temps (dans le cas de la transformée de Fourier), les fréquences (pour la transformée de Fourier inverse), l'espace bidimensionnelle d'une photographie analogique (pour la transformée en ondelettes) ou numérique (JPEG2000). Dans ce dernier cas, $\mathcal{X}$ constitue une grille, dont la largeur et la longueur sont déterminées par le nombre de pixels de l'image. D'un point de vue formel, une telle grille est un graphe avec autant de sommets que de pixels, chacun étant relié à ses quatre voisins par des arêtes. Ainsi, le signal f n'est plus défini sur $\mathbb{R}_+$ (comme avec un signal classique) mais sur un graphe.

Naturellement, il est possible de généraliser à des graphes qui n'ont pas la régularité d'une grille. Par exemple, un réseau social : les sommets représentent les individus, et les arêtes les relations entre eux. Nous pouvons définir le domaine $\mathcal{X}$ comme l'ensemble des individus du réseau social, et f comme une fonction sur ces derniers. Par exemple, une fonction binaire qui indique si chaque individu est au courant d'une information, ou une fonction réelle positive indiquant le temps qu'il aura fallu pour que l'information diffuse jusqu'à chacun d'eux. Plus généralement, toute population statistique peut servir à définir le domaine $\mathcal{X}$, et l'on peut étudier une propriété liée à cette population, encodée sous la forme d'une fonction f définie sur $\mathcal{X}$, au moyen des outils de l'analyse harmonique.

Alors qu'en traitement d'images, le signal f est généralement connu, ce n'est souvent pas le cas en extraction de connaissances. Les problèmes y sont légèrement

différents, puisqu'on cherche en général plutôt à estimer cette fonction : on peut chercher par exemple une fonction qui prend la valeur 0 ou 1 suivant qu'un individu appartienne à un groupe ou à un autre. Cependant, même si l'objectif diffère fondamentalement, il apparaît que les outils de l'analyse harmonique restent un excellent cadre de travail : finalement, que l'on s'intéresse à la diffusion de la chaleur dans une structure particulière (ce qui a motivé les développements de Fourier), ou que l'on s'intéresse à la diffusion d'information dans un réseau social, ou encore, que l'on utilise les processus de diffusion afin de caractériser les relations entre les individus d'une population statistique, le même formalisme mathématique s'applique.

Dans la suite, afin de nous rapprocher des objectifs que nous poursuivons en science des données, nous supposerons que le domaine $\mathcal{X}$ est celui sur lequel est définie une population de n individus, notés $\{x_1, \ldots, x_n\}$, telle que dans l'exemple ci-dessus du réseau social, ou telle que définie dans la Sec. 2. De plus, nous allons restreindre notre étude aux opérateurs intégraux dont le noyaux s'interprête naturellement en termes de similarités.

Produit scalaire et similarité

Quand nous décrivons une population humaine de n individus $\{x_1, \ldots, x_n\}$ au moyen de p variables physiologiques (la taille, le poids, le sexe, etc.), nous pouvons assimiler chaque individu x_i à un vecteur $\mathbf{x}_i$ caractérisé par p composantes (ses coordonnées dans un repère en quelques sorte). En faisant cela, nous munissons le domaine $\mathcal{X}$ d'une structure d'espace vectoriel. Dès lors, il est possible de confondre la population $\{x_1, \ldots, x_n\}$, et les vecteurs $\mathbf{x}_1, \ldots, \mathbf{x}_n$ la représentant. Si les p variables prennent des valeurs réelles, alors cet espace vectoriel se denote $\mathbb{R}^p$. Comme beaucoup d'espaces vectoriels "classiques", $\mathbb{R}^p$ est un *espace de Hilbert*, c'est-à-dire qu'il est muni d'un *produit scalaire*, et qu'il est *complet* (en gros, "il n'y manque aucun point"). Ce produit scalaire se note $\langle ., . \rangle_{\mathbb{R}^p}$, ou encore $\langle ., . \rangle_{\mathcal{X}}$, puisqu'ici $\mathbb{R}^p = \mathcal{X}$.

Formellement, un produit scalaire $\langle ., . \rangle_{\mathcal{X}}$ est une application de $\mathcal{X} \times \mathcal{X}$ dans $\mathbb{R}_+$, telle que, $\forall \mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3 \in \mathcal{X}$ et $\forall a \in \mathbb{R}$: **1.** $\langle \mathbf{x}_1, \mathbf{x}_2 \rangle_{\mathcal{X}} = \langle \mathbf{x}_2, \mathbf{x}_1 \rangle_{\mathcal{X}}$; **2.** $\langle \mathbf{x}_1, \mathbf{x}_1 \rangle_{\mathcal{X}} = 0 \Rightarrow \mathbf{x}_1 = 0$; **3.** $\langle a\mathbf{x}_1 + \mathbf{x}_2, \mathbf{x}_3 \rangle_{\mathcal{X}} = a \cdot \langle \mathbf{x}_1, \mathbf{x}_3 \rangle_{\mathcal{X}} + \langle \mathbf{x}_2, \mathbf{x}_3 \rangle_{\mathcal{X}}$. En termes plus prosaïques, un produit scalaire permet, à un facteur d'échelle près, de mesurer une similarité entre les individus d'une population : en effet, elle concerne deux individus et leur associe un score positif ou nulle (la similarité étant d'autant plus forte que le score l'est) ; d'après **1.**, la similarité entre les individus x_1 et x_2 est la même qu'entre les individus x_2 et x_1 ; d'après **2.**, la similarité entre un individu et lui-même est forcément non-nulle ; enfin, d'après **3.**, la mesure de similarité est linéaire par rapport à un facteur multiplicatif, de sorte qu'il est possible de rééchelonner la similarité entre 0 (dissimilarité parfaite) et 1 (similarité parfaite entre un individu et lui-même).

Tout produit scalaire $\langle ., . \rangle$ défini sur un espace de dimensionnalité finie p a l'intéressante propriété suivante : il est possible de le réécrire sous la forme d'un produit hermitien : il existe une matrice $\mathbf{H} \in \mathbb{R}^{p \times p}$, symétrique (en raison de la symétrie du produit scalaire) dite *matrice hermitienne*, telle que $\langle \mathbf{x}_1, \mathbf{x}_2 \rangle = \mathbf{x}_1^\top \mathbf{H} \mathbf{x}_2$, où $.^\top$ désigne la transposition (le vecteur-colonne devient un vecteur-ligne, et inversement). Réciproquement, toute matrice carrée de taille $d \times d$ *symétrique définie positive*, c'est-à-dire telle que $\mathbf{x}^\top \mathbf{H} \mathbf{x} \geq 0$ (et $\mathbf{x}^\top \mathbf{H} \mathbf{x} = 0$ seulement pour $\mathbf{x} = 0$), est une matrice hermitienne, et donc définit un produit scalaire sur $\mathbb{R}^d$.

Les individus deviennent des fonctions

Intéressons-nous maintenant aux opérateurs intégraux de la forme donnée dans l'Eq. 1.13, mais en nous restreignant au cas d'un noyau k *symétrique semi-défini positif*. Cela signifie que pour toute population $\{x_1, \ldots, x_n\}$ de $\mathcal{X}$, la matrice

$$\mathbf{K} = \begin{pmatrix} k(x_1, x_1) & \cdots & k(x_1, x_n) \\ \vdots & k(x_i, x_j) & \vdots \\ k(x_n, x_1) & \cdots & k(x_n, x_n) \end{pmatrix} \tag{1.14}$$

est *symétrique semi-définie positive*. Une telle matrice est presque identique à une matrice *symétrique définie positive* : la seule différence réside dans la relaxation de la contrainte "$\mathbf{x}^\top \mathbf{H}\mathbf{x} = 0$ seulement pour $\mathbf{x} = 0$" mentionnée plus haut. À ce détail près, un noyau symétrique semi-défini positif permet de générer la matrice hermitienne d'un produit scalaire. Or, nous venons de voir que la sémantique naturelle du produit scalaire est celle d'une mesure de similarité. Dès lors, celle-ci se retrouve dans la formule de l'opérateur intégral. Pour illustrer cela, considérons que pour tous les individus de la population $\{x_1, \ldots, x_n\}$, nous ayons connaissance de la valeur prise par une fonction d'intérêt f (qui par exemple indique pour chaque individu x, le temps qu'il faut pour que x voit la dernière vidéo virale à la mode sur le réseau social) : ainsi, nous avons accès aux valeurs $\{f(x_1), \ldots, f(x_n)\}$. Maintenant, considérons un individu x_{n+1} pour lequel $f(x_{n+1})$ est inconnu. Il est naturel d'estimer cette valeur en fonction des valeurs $f(x_i)$ associées aux individus x_i ayant des caractéristiques similaires à celles de x_{n+1}, voire même de calculer une moyenne des valeurs $f(x_i)$, pondérée par les similarités entre x_{n+1} et x_i. Cette similarité étant fourni par le noyau semi-défini positif (que par simplicité nous supposerons normalisé, c'est-à-dire que $k(x_i, x_i) = 1,\ \forall i$), nous obtenons :

$$\hat{f}(x_{n+1}) = \frac{1}{n} \sum_{i=1}^{n} k(x_{n+1}, x_i) f(x_i) \tag{1.15}$$

Maintenant, supposons que n tende vers l'infini, de sorte que la population $\{x_1, \ldots, x_n\}$ couvre "continument" l'espace $\mathcal{X}$. La somme devient une intégrale (f est donc pondérée par $d\mu(.)$ plutôt que par $\frac{1}{n}$), et nous retombons sur la formule de l'Eq. 1.13 :

$$T_k[f](x') = \int_{\mathcal{X}} k(x', x) f(x) d\mu(x), \quad x' \in \mathcal{X},$$

Maintenant que l'intérêt des noyaux semi-définis positifs ne fait plus de doute, revenons aux principaux résultats qui y sont associés [46]. Notamment, il a été montré que pour tout noyau k de ce type, nous avons les résultats suivants : (i) il existe au moins un espace de Hilbert $\mathcal{H}$, dit *à noyau reproduisant*, et au moins une fonction $\Phi : \mathcal{X} \mapsto \mathcal{H}$ qui permet de représenter tout élément x de $\mathcal{X}$ par un élément $\Phi(x)$ de $\mathcal{H}$; (ii) Le produit scalaire $\langle ., . \rangle_{\mathcal{H}}$ est tel qu'il *reproduit* k, de sorte que nous avons :

$$\langle \Phi(x_i), \Phi(x_j) \rangle_{\mathcal{H}} = k(x_i, x_j), \quad \forall (x_i, x_j) \in \mathcal{X}^2. \tag{1.16}$$

(iii) L'ensemble $\mathbb{R}^{\mathcal{X}}$ des fonctions de $\mathcal{X}$ dans $\mathbb{R}$ est un candidat admissible pour $\mathcal{H}$. Dans ce cas, cela signifie que $\Phi(x)$ est en fait une fonction, que nous noterons

$\Phi_{[x]}(.)$ par commodité. Il est donc possible de représenter tout élément x de $\mathcal{X}$, non plus par un vecteur au sens classique du terme, mais par une fonction $\Phi_{[x]}(.)$ définie sur le domaine $\mathcal{X}$. (*iv*) Enfin, il est possible d'identifier précisément cette fonction $\Phi_{[x]}(.)$, puisque nous avons $\Phi_{[x]}(.) = k(., x)$. Le noyau k étant semi-défini positif, il garde la même sémantique de similarité. Ainsi $\Phi_{[x]}(.)$ est une fonction qui mesure la similarité de x avec tous les éléments pouvant être défini sur le domaine $\mathcal{X}$.

Cela vient confirmer notre intuition selon laquelle il est possible d'utiliser les données pour les décrire elles-mêmes, dans une logique auto-structurante telle que décrite sur la pyramide DIC. Ainsi, pour paraphraser [46], "*c'est en quelque sorte comme si maintenant, un individu était représenté par sa similarité à tous les autres individus qu'il est possible de définir sur le domaine* $\mathcal{X}$". Nous sommes donc bien dans une situation où les individus se définissent les uns par rapport aux autres : ils sont chacun mis en contexte par leurs pairs.

L'astuce du noyau (ou "*kernel trick*")

Quand $\mathcal{X}$ est un espace de Hilbert, il est possible de définir une matrice particulièrement intéressante pour l'étude des similarités entre les individus $\{x_1, \ldots, x_n\}$. Il s'agit de la *matrice de Gram* des vecteurs $\mathbf{x}_1, \ldots, \mathbf{x}_n$; elle contient tous les couples de produits scalaires possibles, pour une population donnée :

$$\mathbf{G} = \begin{pmatrix} \langle \mathbf{x}_1, \mathbf{x}_1 \rangle_{\mathcal{X}} & \cdots & \langle \mathbf{x}_1, \mathbf{x}_n \rangle_{\mathcal{X}} \\ \vdots & \langle \mathbf{x}_i, \mathbf{x}_j \rangle_{\mathcal{X}} & \vdots \\ \langle \mathbf{x}_n, \mathbf{x}_1 \rangle_{\mathcal{X}} & \cdots & \langle \mathbf{x}_n, \mathbf{x}_n \rangle_{\mathcal{X}} \end{pmatrix} \tag{1.17}$$

Maintenant, intéressons-nous à la matrice de Gram, non plus des vecteurs $\mathbf{x}_1, \ldots, \mathbf{x}_n$, mais à celle de leurs images dans $\mathcal{H}$, à savoir $\Phi(\mathbf{x}_1), \ldots, \Phi(\mathbf{x}_n)$:

$$\mathbf{G}_\Phi = \begin{pmatrix} \langle \Phi(x_1), \Phi(x_1) \rangle_{\mathcal{H}} & \cdots & \langle \Phi(x_1), \Phi(x_n) \rangle_{\mathcal{H}} \\ \vdots & \langle \Phi(x_i), \Phi(x_j) \rangle_{\mathcal{H}} & \vdots \\ \langle \Phi(x_n), \Phi(x_1) \rangle_{\mathcal{H}} & \cdots & \langle \Phi(x_n), \Phi(x_n) \rangle_{\mathcal{H}} \end{pmatrix} \tag{1.18}$$

Comme k reproduit $\langle ., . \rangle_{\mathcal{H}}$, nous avons $\mathbf{G}_\Phi = \mathbf{K}$ (où $\mathbf{K}$ est la matrice définie en Eq. 1.14). En d'autres termes, il est possible de définir explicitement la matrice de Gram associée à l'image dans $\mathcal{H}$, sans faire intervenir les fonctions $\Phi_{[x]}(.)$ qui n'ont pas forcément une définition analytique connue.

Ainsi, tout algorithme qui effectue un calcul sur la population $\{x_1, \ldots, x_n\}$ (afin d'estimer une fonction sur celle-ci par exemple), en ne se basant que sur les relations pair à pair entre individus (qui peuvent être encodées par une matrice de Gram $\mathbf{G}$) peut être modifié, pour effectuer ce même calcul dans un autre espace (en l'occurrence $\mathcal{H}$) à la "topographie" radicalement différente. Pour cela, il suffit d'utiliser la matrice $\mathbf{K}$ à la place de $\mathbf{G}$. C'est exactement en cela que consiste le *kernel trick*. L'intérêt est le suivant : le changement de "topographie" entre $\mathcal{X}$ et $\mathcal{H}$ correspond à des distorsions de la géométrie de l'espace de travail. Ainsi, l'évaluation d'une fonction non-linaire sur $\mathcal{X}$ qui par nature est un problème difficile, peut, si l'on "déforme" correctement l'espace, se transformer en l'évaluation d'une fonction linéaire, comme cela est intuitivement illustré dans la vidéo [47]. Les non-linéarités étant particulièrement présentes dans le monde vivant, les algorithmes utilisant cette astuce du noyau ont eu un succès retentissant en bioinformatique.

5 L'apprentissage automatique

Comme cela est raconté dans [48], l'apprentissage automatique avait une saveur très cognitive à ses débuts, induisant naturellement un rapprochement avec les deux courants de l'IA (cf. Sec. 3) au cours des années 80. Ensuite, à partir des années 90, sous l'influence principale de Vapnik, qui développait la théorie statistique de l'apprentissage [49], un glissement vers les statistiques et vers l'optimisation convexe s'est opéré, tendant à rendre l'apprentissage beaucoup plus mathématique, tout en s'éloignant de l'IA [10]. Cependant, l'histoire décrite dans [48] affilie aussi l'apprentissage automatique à la *reconnaissance de formes*, un domaine qui faisait déjà le lien entre statistique et traitement du signal, tout en adoptant un formalisme algébrique qui a été particulièrement développé en *analyse de données*. Ainsi, au regard de l'évolution actuelle de l'apprentissage automatique, il me semble approprié de positionner l'analyse de données au centre de sa généalogie.

5.1 L'analyse de données

Racines algébriques

Outre-Atlantique, la *data analysis* fait simplement et logiquement référence aux différentes étapes du traitement des données (collecte, stockage, nettoyage, interprétation, etc.) ; alors qu'en France, l'*analyse de données* est un courant méthodologique en tant que tel, qui s'est développé sous la bannière du bourbakiste Jean-Paul Benzécri [50]. Il s'agit une approche alternative et particulièrement mathématisée des statistiques qui s'est développée dans les années 60 et 70, et qui a redécouvert, enrichi et étendu les méthodes connues aux Etats-Unis sous le nom de *multivariate analysis*, dont l'*analyse en composantes principales* (ACP) reste l'outil le plus connu.

La principale spécificité de ce courant réside dans l'absence de modèle probabiliste expliquant les données : celles-ci existent, simplement, et de manière pragmatique, il s'agit de les étudier, sans questionner l'existence du hasard ou la difficulté de sa modélisation. C'est pourquoi, le formalisme des probabilités a été remplacé par celui de l'algèbre linéaire. En effet, celui-ci est particulièrement utile à la description d'une population par plusieurs variables. Dans la section précédente, nous avons immergé les données dans un espace vectoriel, de sorte que la population $\{x_1, \dots, x_n\}$ peut-être représentée par un ensemble de vecteurs. Ici, nous les concaténons au sein d'une matrice $\mathbf{X} = [\mathbf{x}_1, \dots, \mathbf{x}_n] \in \mathbb{R}^{p \times n}$, dont ils forment les colonnes. Celle-ci encodant toute l'information disponible sur cette population, il devrait être possible de réaliser tous les traitements souhaités via des calculs appropriés sur $\mathbf{X}$. L'analyse de données fournit simplement cette boîte à outils calculatoire.

Changement de base

Bien que cela soit complétement artificiel, il est possible de réécrire la matrice $\mathbf{X}$ en termes de produits scalaires entre les vecteurs représentant les individus de la population $\mathbf{x}_1, \dots, \mathbf{x}_n$ et les variables statistiques $\mathbf{e}_1, \dots, \mathbf{e}_\ell$ (appelées *vecteurs de*

10. Ensuite, à partir des années 2000-2010, le courant bioinspiré de l'IA et l'apprentissage automatique ce sont fortement rapprochés, à l'inverse du courant logique de l'IA .

base, ou simplement *base*) qui sont utilisées pour engendrer le domaine $\mathcal{X}$:

$$\mathbf{X} = \begin{pmatrix} \langle \mathbf{e}_1, \mathbf{x}_1 \rangle & \cdots & \langle \mathbf{e}_1, \mathbf{x}_n \rangle \\ \vdots & \langle \mathbf{e}_i, \mathbf{x}_j \rangle & \vdots \\ \langle \mathbf{e}_\ell, \mathbf{x}_1 \rangle & \cdots & \langle \mathbf{e}_\ell, \mathbf{x}_n \rangle \end{pmatrix} \tag{1.19}$$

En revanche, cela s'interprète facilement en termes de statistiques descriptives : la cellule (i, j) de la matrice $\mathbf{X}$ indique le niveau de similarité de l'individu $\mathbf{x}_j$ (par exemple "Pierre") par rapport à la variable $\mathbf{e}_i$ (par exemple "yeux bleus", ou "grand"). Cependant, il est tout à fait possible de décrire la même population $\{x_1, \ldots, x_n\}$ avec d'autres, variables $\mathbf{e}'_1, \ldots, \mathbf{e}'_L$ (associées par exemple au caractère des individus, plutôt qu'à leur physiologie). La matrice résultante serait différente :

$$\mathbf{X}' = \begin{pmatrix} \langle \mathbf{e}'_1, \mathbf{x}_1 \rangle & \cdots & \langle \mathbf{e}'_1, \mathbf{x}_n \rangle \\ \vdots & \langle \mathbf{e}'_i, \mathbf{x}_j \rangle & \vdots \\ \langle \mathbf{e}'_L, \mathbf{x}_1 \rangle & \cdots & \langle \mathbf{e}'_L, \mathbf{x}_n \rangle \end{pmatrix} \tag{1.20}$$

Cependant, l'algèbre matricielle nous indique qu'il est possible de passer de l'une à l'autre, par la formule suivante :

$$\begin{pmatrix} \langle \mathbf{e}'_1, \mathbf{x}_1 \rangle & \cdots & \langle \mathbf{e}'_1, \mathbf{x}_n \rangle \\ \vdots & \langle \mathbf{e}'_i, \mathbf{x}_j \rangle & \vdots \\ \langle \mathbf{e}'_L, \mathbf{x}_1 \rangle & \cdots & \langle \mathbf{e}'_L, \mathbf{x}_n \rangle \end{pmatrix} = \begin{pmatrix} \langle \mathbf{e}'_1, \mathbf{e}_1 \rangle & \cdots & \langle \mathbf{e}'_1, \mathbf{e}_\ell \rangle \\ \vdots & \langle \mathbf{e}'_i, \mathbf{e}_j \rangle & \vdots \\ \langle \mathbf{e}'_L, \mathbf{e}_1 \rangle & \cdots & \langle \mathbf{e}'_L, \mathbf{e}_\ell \rangle \end{pmatrix} \times \begin{pmatrix} \langle \mathbf{e}_1, \mathbf{x}_1 \rangle & \cdots & \langle \mathbf{e}_1, \mathbf{x}_n \rangle \\ \vdots & \langle \mathbf{e}_i, \mathbf{x}_j \rangle & \vdots \\ \langle \mathbf{e}_\ell, \mathbf{x}_1 \rangle & \cdots & \langle \mathbf{e}_\ell, \mathbf{x}_n \rangle \end{pmatrix} \tag{1.21}$$

Ou encore, avec une notation plus compact :

$$\mathbf{X}' = \mathbf{P}\mathbf{X} \qquad \text{où} \qquad \mathbf{P} = \begin{pmatrix} \langle \mathbf{e}'_1, \mathbf{e}_1 \rangle & \cdots & \langle \mathbf{e}'_1, \mathbf{e}_\ell \rangle \\ \vdots & \langle \mathbf{e}'_i, \mathbf{e}_j \rangle & \vdots \\ \langle \mathbf{e}'_L, \mathbf{e}_1 \rangle & \cdots & \langle \mathbf{e}'_L, \mathbf{e}_\ell \rangle \end{pmatrix} \tag{1.22}$$

est appelée *matrice de passage*, et encode les relations entre les deux bases $\mathbf{e}_1, \ldots, \mathbf{e}_\ell$ et $\mathbf{e}'_1, \ldots, \mathbf{e}'_L$. C'est ce que l'on appelle un *changement de base*, une opération qui est au centre des développements décrits au Chap. 4.

Diagonalisation

Il est possible d'aller plus loin, et de chercher un couple de matrices de passage $\mathbf{U}$ et $\mathbf{V}$ telles que $\mathbf{XV} = \mathbf{UD}$,où $\mathbf{D}$ est une matrice diagonale, c'est-à-dire :

$$\mathbf{D} = \begin{pmatrix} \alpha_1 & \cdots & 0 \\ \vdots & \alpha_i & \vdots \\ 0 & \cdots & \alpha_n \end{pmatrix} \tag{1.23}$$

Autrement dit, si l'on note $\{\mathbf{e}''_1, \ldots, \mathbf{e}''_p\}$ la base associée à la matrice $\mathbf{D}$, le vecteur $\mathbf{e}''_i$ est une combinaison des variables statistiques telles qu'à un facteur multiplicatif près, noté α_i, il corresponde à l'individu x_i. Ainsi, les caractéristiques des individus servent à la constitution d'une base, dans une logique systématiquement soulignée depuis le début de ce chapitre, selon laquelle les données servent à se décrire elles-mêmes. Plus prosaïquement, si nous multiplions à droite l'expression précédente par $\mathbf{V^{-1}}$, nous obtenons $\mathbf{XVV^{-1}} = \mathbf{UDV^{-1}}$ ce qui est équivalent à $\mathbf{X} = \mathbf{UDV^{-1}}$. Cette dernière expression est connue sous le nom de *décomposition en valeurs singulières* de $\mathbf{X}$ (ou *diagonalisation*, si $\mathbf{X}$ est carrée). Dans le cas particulier où $\mathbf{X}$ est semi-définie positive (tel que la matrice $\mathbf{K}$ définie en 4.2, ou encore une matrice de Gram), on a en plus $\mathbf{U} = \mathbf{V}$.

Matrice de covariance

Maintenant que ces notions sont introduites, il est possible de faire le lien entre les méthodes statistiques classiques, décrites au début de ce chapitre, et l'analyse harmonique, décrite dans la section précédente : pour ce faire, nous allons mettre en parallèle deux matrices distinctes. La première, notée $\mathbf{\Sigma}$, est la matrice de covariance associée à la matrice $\mathbf{X}$. Elle a pour expression $\mathbf{\Sigma} = \frac{1}{n}(\mathbf{X}-[\overline{\mathbf{x}}])(\mathbf{X}-[\overline{\mathbf{x}}])^\top$ où $.^\top$ désigne la transposée d'une matrice (cela consiste à "inverser" les lignes et les colonnes, par une rotation le long de la diagonale), et où $[\overline{\mathbf{x}}]$ désigne une matrice constituée de la répétition de la colonne moyenne de $\mathbf{X}$, notée $\overline{\mathbf{x}}$: $(\mathbf{X} - [\overline{\mathbf{x}}])$ est donc simplement la version "centrée" de $\mathbf{X}$. Si $\mathbf{X}$ est une matrice de n colonnes (les individus) et p lignes (les variables), alors, $\mathbf{\Sigma}$ est une matrice carrée de taille $p \times p$, symétrique, et dont chaque cellule (i, j) contient la covariance entre les $i^{\text{ème}}$ et $j^{\text{ème}}$ variables. Cette matrice est très utilisée en statistique pour déterminer quelles sont les variables décrivant une population statistique donnée, qui sont corrélées, ou au contraire, qui ne le sont pas. Cette matrice est aussi au cœur de l'analyse en composantes principales (ACP), dont l'une des étapes fondamentales est la diagonalisation de $\mathbf{\Sigma}$.

La seconde matrice qui nous intéresse est simplement la matrice de Gram $\mathbf{G}$ associée à $\mathbf{X}$, telle que précédemment définie dans la section dédiée à l'analyse harmonique. Il se trouve que si l'on diagonalise $\mathbf{\Sigma}$ et $\mathbf{G}$, les matrices diagonales résultantes auront exactement le même ensemble de valeurs non-nulles (on dit que $\mathbf{\Sigma}$ et $\mathbf{G}$ ont le même *spectre*). Cela est d'autant plus surprenant que (1) $\mathbf{G}$ est de taille $n \times n$ contrairement à $\mathbf{\Sigma}$ qui est de taille $p \times p$; (2) $\mathbf{\Sigma}$ décrit les relations entre variables sans explicitement faire référence à la population étudiée en tant que telle, alors qu'on contraire, $\mathbf{G}$ s'intéresse aux similarités entre les paires d'individus, en l'absence de tout référentiel (et sans intervention explicite des variables statistiques).

La conséquence de cette égalité est heureuse : dans bien des cas, une analyse "statistique" peut se conduire avec une approche "harmonique", en substituant la matrice de covariance par une matrice de Gram, voire par un noyau, afin d'appliquer le *kernel trick*. Ainsi, il est possible de définir une ACP non linéaire [46]. Cependant cela va bien au-delà et permet de réaliser un véritable pont entre statistique, analyse de données, et traitement du signal, que nous allons continuer à explorer.

5.2 Gérer la grande dimensionnalité

Par son approche algébrique, l'analyse de données semble particulièrement adaptée à la grande dimensionnalité. En effet, il "suffit" de rajouter des lignes à la matrice : on change donc la taille de la matrice, mais pas la manière de la traiter. Malheureusement, il apparaît que les espaces de grande dimensionnalité n'ont pas les mêmes comportements que les espaces auxquels notre intuition se confronte normalement, de sorte qu'il n'est pas toujours possible d'appliquer les raisonnements classiques à la base d'une inférence cognitive, tels que le regroupement d'individus similaires, ou la recherche de corrélations. Il est classiquement fait référence à ce phénomène contre-intuitif sous le nom de *malédiction de la dimensionnalité*. L'illustration la plus étonnante de ce phénomène est ce qu'on appelle la compression de la norme [51, 52] : si l'on jette des points au hasard dans un espace de grandes dimensions, les distances entre tous les couples de points, les proches comme les distants,

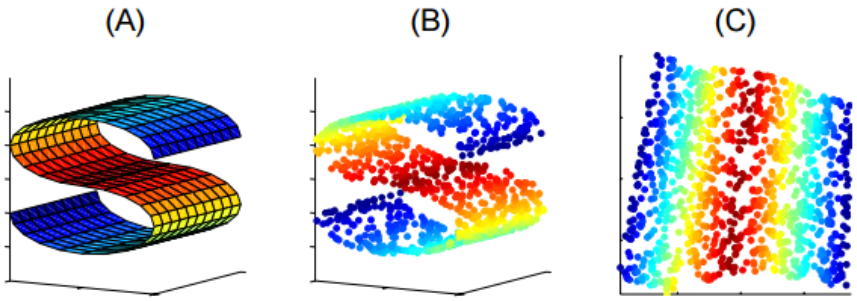

FIGURE 1.7 – *Un ensemble de points (B) habitent sur une variété (A) dans un espace à 3 dimensions : il n'est pas possible d'opérer une réduction de dimensionnalité linéaire. En revanche, il est possible de "déplier" la variété, et d'obtenir un espace à 2 dimensions, qui préserve localement la géométrie des points (C) - tiré de [53].*

seront à peu près les mêmes, de sorte que la distance perd de son intérêt.

Face à cette difficulté, il existe malgré tout plusieurs parades : la première consiste à choisir la matrice qui de $\mathbf{\Sigma}$ ou $\mathbf{G}$ est la plus petite. Ainsi, dans le cas d'un problème en grande dimension "*large p*" avec peu d'individus "*small n*", il peut être plus efficace de travailler sur $\mathbf{G}$. Bien sûr, cela ne résout pas les problèmes ; cela les cache, tout au plus. Une seconde parade possible est déconcertante de brutalité : il s'agit simplement de transformer de force le problème pour que celui-ci s'exprime dans un espace de dimensionnalité plus faible. C'est ce que l'on appelle la *réduction de dimensionnalité*. Une manière élémentaire de réduire la dimensionnalité est de supprimer les variables qui sont "redondantes" (trop fortement corrélées à d'autres). Cela peut se faire à la main, en observant la matrice $\mathbf{\Sigma}$; ou de manière plus systématique en diagonalisant $\mathbf{\Sigma}$, puis en ne gardant que les k valeurs diagonales (appelées les *valeurs propres*) $\alpha_1, \ldots, \alpha_k$ les plus importantes, tout en annulant les autres, avant de reconstruire une version de dimensionnalité k du problème :

$$\mathbf{X}_{(dim)(k)} = \mathbf{P}\mathbf{D}_k\mathbf{P}^{-1} \quad \text{avec} \quad \mathbf{D}_k = \begin{pmatrix} \alpha_1 & & & & \\ & \ddots & & (0) & \\ & & \alpha_k & & \\ & (0) & & 0 & \\ & & & & \ddots \end{pmatrix} \tag{1.24}$$

C'est tout simplement (et encore une fois) ce que l'on appelle une ACP ; qui permet de réduire la dimensionnalité de manière optimale sous contrainte de linéarité.

Cependant, une réduction de dimension linéaire n'est pas toujours adaptée à un problème réel, qui peut présenter de fortes non-linéarités, car elle pourrait induire une perte d'informations pertinentes. Heureusement, il est possible d'opérer une réduction non-linéaire : en pratique, on observe que les données qui habitent dans un espace de grande dimensionnalité se concentrent souvent sur des surfaces courbes, dont la dimension intrinsèque est plus faible [54, 55], tel que cela est illustré schématiquement sur la Fig. 1.7. Malheureusement, il est particulièrement difficile de travailler sur de telles courbes (aussi appelé *variétés*). Cependant, il est possible d'utiliser le *kernel trick*, afin de travailler dans un espace fictif qui correspondrait au dépliement de l'espace d'origine. Pour ce faire, il convient d'abord de se convaincre

qu'entre la Fig. 1.7(B) et Fig. 1.7(C), l'ensemble des relations entre les pairs de points est localement préservé, de sorte que les points qui étaient voisins avant le dépliement le reste après. Il convient ensuite de se rendre compte que durant ce dépliement, la référence à un repère extérieur (les variables du problème) est inutile, de sorte qu'il devrait être possible de travailler seulement à partir de la matrice de Gram. Concrètement, dans le cas du "S", on peut modifier cette matrice comme suit : la similarité (ou la distance correspondante) entre toutes les paires de points ne doit plus être calculée selon un chemin qui nécessite de "sortir" du "S", mais en suivant un chemin "de proche en proche". La matrice de Gram ainsi modifiée va correspondre aux similarités dans un nouvel espace où certains points ne sont plus connectés que par une succession d'autres points, exactement comme si la variété avait été dépliée[11]. En pratique, le rang de cette matrice de Gram (c'est-à-dire la taille de son spectre) va être beaucoup plus faible, confirmant que l'on a bien réussi à réduire la dimension du problème.

Finalement, le principe de la réduction de dimensionnalité est élémentaire : il s'agit de modifier des valeurs de la matrice $\mathbf{X}$ ou de la matrice de Gram associée, de telle sorte que le problème en est simplifié, tout en en préservant les caractéristiques intrinsèques, qui nous intéresse. Dans le cas où ces modifications impliquent de forcer beaucoup de valeurs à zéro, on parle de représentation *parcimonieuse*. Avant de rentrer plus dans le détail des représentations parcimonieuses, il est cependant nécessaire d'introduire quelques notions d'optimisation.

5.3 Entre statistique, optimisation et géométrie

Jusqu'à présent, nous avons surtout fait le lien entre statistique et analyse harmonique, la notion même d'*apprentissage* n'étant pas particulièrement présente. Celle-ci fait simplement référence à l'estimation d'une fonction (permettant d'encoder la réalisation d'une tâche ou la prise d'une décision), sur la base de l'expérience acquise par l'observation de plusieurs exemples, pour lesquels la valeur de la fonction est connue : par exemple, si toutes les personnes d'une taille supérieure à 1m80 sont qualifiées "grandes", il doit être possible de généraliser, et d'associer ce label à quelqu'un mesurant 182.38cm, même si cela n'a jamais été observée auparavant.

Il y a deux concepts particulièrement importants dans l'apprentissage : le premier est la capacité à généraliser des conclusions observées sur un nombre fini d'exemples. Le second est la définition de la fonction que l'on cherche à apprendre. Comme dit plus haut, il peut s'agir d'une action à réaliser, d'une décision à prendre, de la discrimination des individus en plusieurs catégories, de la prédiction d'une observation à venir, etc. Ces deux éléments sont à la base de la théorie statistique de l'apprentissage, telle que définie par Vapnik, dans le milieu des années 90 [49].

11. Notons ici encore, une analogie profonde entre l'analyse de donnée et l'étude des processus de diffusion en physique : le calcul d'une distance "de proche en proche" correspond en fait à la version discrète d'un processus de diffusion (par exemple, la diffusion de la chaleur, déjà mentionnée). Le Laplacien est l'opérateur mathématique au cœur des équations caractérisant un tel processus de diffusion sur des structures continues (espaces Euclidiens, ou variétés Riemannienne, pour sa généralisation de Beltrami [56]), et il est possible d'en donner une version discrète, adaptée à des jeux de données statistique. Cette version discrète, appelé Laplacien du graphe a la forme d'une matrice, dont l'inverse est une matrice de Gram, appelé *noyau de diffusion*.

Nous supposerons l'existence d'une fonction $f : \mathcal{X} \mapsto \mathcal{Z}$, inconnue, mais qui correspond aux concepts que l'on souhaite apprendre. Cette fonction est capable d'associer à des individus observés (des vecteurs de $\mathcal{X}$) ces concepts, que nous supposerons faisant partie d'un domaine $\mathcal{Z}$. Le but de tout algorithme d'apprentissage est de trouver une fonction $\hat{f} : \mathcal{X} \mapsto \mathcal{Z}$ qui approxime au mieux f. Autrement dit, nous souhaitons que f et $\hat{f}$ soient le moins possible différentes. Comme la différence entre f et $\hat{f}$ correspond à un risque de se tromper de concept (dans $\mathcal{Z}$), nous pouvons définir la quantité suivante, dénommée *risque réel* :

$$R_r(\hat{f}) = \int_{\mathcal{X}} dist(f(x), \hat{f}(x)) d\mu(x) \tag{1.25}$$

Ou $dist()$ est une mesure de distance sur $\mathcal{Z}$ qui permet de quantifier l'écart entre les concepts prédits par f et $\hat{f}$. Malheureusement, comme $\hat{f}$ ainsi que les différentes instances possibles de x dans $\mathcal{X}$ ne sont pas connues, il n'est pas possible de calculer ce risque. Cependant, si l'on dispose d'un nombre suffisant d'exemples d'apprentissage de la forme $(x, f(x))$, c'est à dire constitués de l'observation d'origine accompagnée de la "réponse" souhaitée, et si ces différents exemples sont i.i.d. (cf. Sec. 2.2), il est néanmoins possible de l'estimer. Concrètement, la suite de n exemples $\{(x_1, f(x_1)), \dots (x_n, f(x_n))\}$ peut être considérée comme un échantillon représentatif de $\mathcal{X}$ (comme quand on effectue un sondage), afin d'estimer R_r. Cet estimateur est classiquement appelé *risque empirique*, et a pour expression :

$$R_e(\hat{f}) = \frac{1}{n} \sum_{i=1}^{n} dist(f(x_i), \hat{f}(x_i)) \tag{1.26}$$

Dès lors, le problème d'apprentissage se résume à trouver parmi un ensemble de fonctions $\mathcal{F}$ la fonction $\hat{f}$ qui permettra de minimiser le risque empirique :

$$\hat{f} = \min_{g \in \mathcal{F}} R_e(g) \tag{1.27}$$

Cette formalisation de l'apprentissage par la minimisation du risque empirique (MRE) a été une véritable révolution : alors que f était d'abord recherchée en se basant sur des considérations cognitives et algorithmiques, voire des analogies naturelles (cf. l'IA bioinspirée, cf. p. 14), le principe de la MRE a brutalement ancré l'apprentissage dans les mathématiques appliquées et l'optimisation. Mais aussi, nous allons le voir, dans les statistiques, telles que décrites à la Sec. 2.

Dans le cas le plus simple, $\mathcal{F}$ est une famille de distributions, et le choix de $\hat{f}$ se résume à l'estimation de ses paramètres, tel qu'expliqué dans la Sec. 2 ; à cette occasion, il avait été brièvement fait mention du compromis biais-variance, qui peut s'illustrer facilement, ici. Comme mentionné plus haut, l'objectif de l'apprentissage est de trouver une fonction $\hat{f}$ qui se généralise bien : les conclusions de l'apprentissage doivent pouvoir être transposées à d'autres observations que les exemples qui ont servi à l'apprentissage. Le moyen le plus simple pour garantir une bonne généralisation, est d'avoir suffisamment d'exemples d'apprentissage pour bien couvrir $\mathcal{X}$, de sorte que $R_r \approx R_e$. Cependant, à nombre constant d'exemples d'apprentissage, le pouvoir de généralisation dépend directement du compromis biais-variance. Une minimisation très fine de R_e entraine une très forte sensibilité aux exemples d'apprentissage, de sorte que s'ils changent légèrement, la fonction $\hat{f}$ peut elle aussi

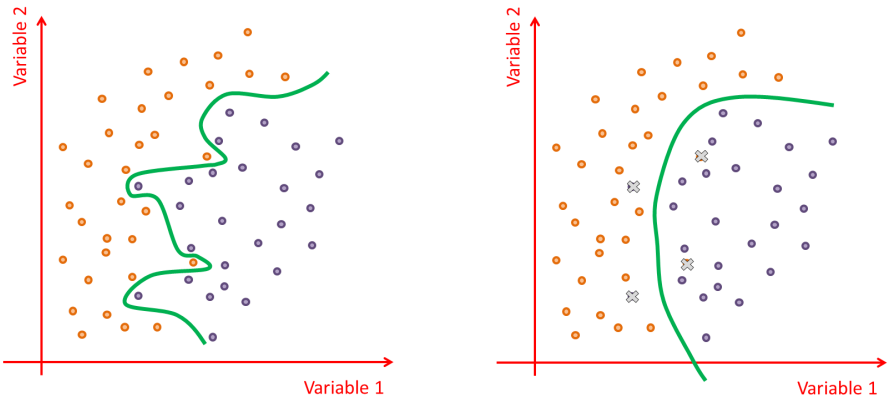

FIGURE 1.8 – *À gauche, la fonction de démarcation entre les deux groupes est très irrégulière, mais très fine : elle "colle au données", mais ne sera pas bien généralisable à d'autres situations, au contraire de celle de droite, beaucoup plus régulière.*

changer (cela correspond à une situation de forte variance, mais de faible biais). En revanche, une minimisation moins fine et plus régulière se généralisera mieux (biais plus important, mais variance faible), tel qu'illustré sur la Fig. 1.8.

Maintenant, faisons le lien avec la malédiction de la dimensionnalité, brièvement décrite plus haut : dès que la dimensionnalité augmente un peu, il devient beaucoup plus difficile d'avoir un échantillonnage suffisant de $\mathcal{X}$, de sorte que le compromis biais-variance prend toute son importance. Ainsi, dans les premières analyses génomiques du milieu des années 90, l'objectif était d'apprendre une fonction permettant de discriminer les patients atteints d'une maladie génétique des personnes saines, sur la base de la liste de leurs gènes. Face aux milliers de gènes humains, qui dans notre problèmes d'apprentissage, correspondent à autant de variables, le nombre de patients atteints d'une maladie et dont on dispose du génome peut être incroyablement faible (une dizaine), notamment pour des maladies rares. Dans un tel contexte, il est impossible d'espérer des performances correctes. La solution consiste donc à réduire la dimensionnalité de $\mathcal{F}$, en forçant une certaine régularité dans la fonction apprise, tel qu'illustré sur la Fig. 1.8. Concrètement, la complexité de chaque fonction candidate doit être prise en compte, et incluse dans l'optimisation, afin de choisir, à risque empirique égal, le modèle le plus simple, et donc le plus généralisable :

$$\hat{f} = \min_{g \in \mathcal{F}} R_e(g) + \rho(g) \tag{1.28}$$

où ρ *pénalise* les fonctions de $\mathcal{F}$ en fonction de leur complexité. Dans un tel contexte, le minimiseur $\hat{f}$ sera naturellement de complexité moindre, et se généralisera mieux. Historiquement, dans les premières études génomiques susnommées, ρ pénalisait les fonctions en proportion du nombre de variables qu'elles faisaient intervenir [18], et la pénalité correspondante, dénommée LASSO (*Least Absolute Shrinkage and Selection Operator*) est encore de nos jours parmi les pénalités les plus populaires.

Finalement, la réduction de dimensionnalité peut soit être explicite, soit passer par la recherche d'un modèle *parcimonieux*, avec peu de degrés de liberté. Cette dernière option nous permet un ultime lien vers l'analyse harmonique (Sec. 4) : afin de remplacer une famille génératrice quelconque par un dictionnaire spécifique, on apprend ce dernier en optimisant son pouvoir expressif pénalisé par sa complexité.

6 Retour sur la pyramide DIC

Maintenant que cet inventaire est terminé, positionnons ces différentes disciplines sur la pyramide DIC (Fig. 1.1), dans un ensemble définissant la science des données.

Les outils issus du courant logique de l'IA, basés sur l'analogie cognitive, se focalisent sur le sommet de la pyramide, avec pour objectif de traiter un faible volume de données au niveau sémantique élevé. À l'heure du *big data*, et alors qu'un cerveau humain n'est pas capable d'appréhender de tête un fichier de quelques milliers d'entrées, tout en étant la seule machine connue permettant de manipuler des connaissances, les outils issus de l'intelligence artificielle ne remplissent pas vraiment les attentes utilitaristes suscitées par les données. Bien qu'absolument passionnant pour l'étude des processus de décision et absolument nécessaire quand certaines tâches de diagnostic doivent être réalisées de manière automatique, le courant de pensée dans lequel s'inscrit la conception de ces outils reste pour l'instant en marge de la science des données, principalement guidée par des objectifs pragmatiques.

Les statistiques et le traitement du signal partagent de nombreuses parentés. Ces disciplines sont apparues au XX^{e} siècle, et ce sont largement développées après la guerre, avec une même philosophie : fournir un modèle explicatif (ou génératif) des données ainsi qu'une théorie largement basée sur les mathématiques, afin de les traiter. Cependant, ces disciplines sont nées à une époque où le *big data* n'existait pas, de sorte qu'il n'était pas nécessaire de monter bien haut sur la pyramide DIC pour se retrouver avec un volume d'information assimilable par un cerveau humain. C'est pourquoi, tous les concepts indispensables à l'inférence en étaient initialement absents, contrairement à l'IA : comparaison, généralisation, induction, transfert, etc.

Enfin, l'apprentissage automatique permet de combler le trou restant, et de permettre l'élévation du niveau sémantique des informations proches de la donnée, aux informations proches des connaissances. Cependant, en raison d'une histoire fortement marquée par l'IA d'une part et les statistiques d'autre part, cette discipline a tendance à venir élargir aux extrémités son domaine d'application.

Le point de jonction entre ces trois disciplines qu'est la théorie des noyaux, et plus largement l'ensemble des méthodes algébriques adaptées à la grande dimensionnalité permet de définir réellement la science des données, grâce à la large place qu'elle laisse à la notion de *contexte* : en effet, ces outils mathématiques permettent de considérer qu'un *datum* ne se décrit plus forcément par un ensemble de variables, mais par sa proximité avec d'autres *data*, dans une optique relativiste, dénuée de tout référentiel explicite, que l'on pourrait qualifier d'auto-structurante [57].

Chapitre 2

La protéomique vue par un *data scientist*

À l'inverse du précédent, ce chapitre s'adresse au *data scientist*, et vise à lui fournir les éléments de protéomique (vocabulaire, notions, etc.) nécessaires à la compréhension du reste du manuscrit. Encore une fois, le ton adopté n'est pas celui classiquement utilisé en biologie : j'ai fondé mon explication sur des analogies facilement compréhensibles par des informaticiens, j'insiste sur la notion de modèle, et je ne prétends pas à l'exhaustivité ; de sorte que beaucoup de mes assertions sembleront particulièrement réductrices au spécialiste, qui ne trouvera pas vraiment d'intérêt à la lecture de ce chapitre, et qui pourra y regretter la prédominance de Wikipédia parmi les références bibliographiques.

1 Quelques éléments de biologie

1.1 Quelques définitions

Protéine Si l'on compare un organisme vivant à une usine très complexe, les protéines sont à la fois le bâti et les installations de cette usine (les murs, les chaines de montage, les outils), les ouvrières travaillant dans cette usine, et une partie des matières que cette usine produit ou transforme. La caractérisation des protéines est donc naturellement au centre des préoccupations de tout biologiste. Concrètement, une protéine est une macromolécule que l'on peut se représenter comme un très long collier de perles replié sur lui-même en une forme compacte, telle que cela est illustrée sur la Fig. 2.1.

Acide aminé Il s'agit des perles constituant le collier qu'est la protéine. Il existe 20 types différents d'acides aminés, que l'on peut décrire à l'aide d'un codage alphabétique. Une protéine peut donc être représentée par un mot d'autant de lettres que d'acides aminés la constituant. La séquence d'enchainement des acides aminés le long du collier va déterminer les possibilités de repliement de la protéine, et donc ses fonctions. Concrètement, un acide aminé est une molécule constituée de deux éléments : tout d'abord, une chaîne carbonée avec à une extrémité, un groupe amine ($-NH_2$), et à l'autre extrémité, un groupe carboxyle (-COOH) ; ensuite, une chaîne latérale (souvent notée R), dont la

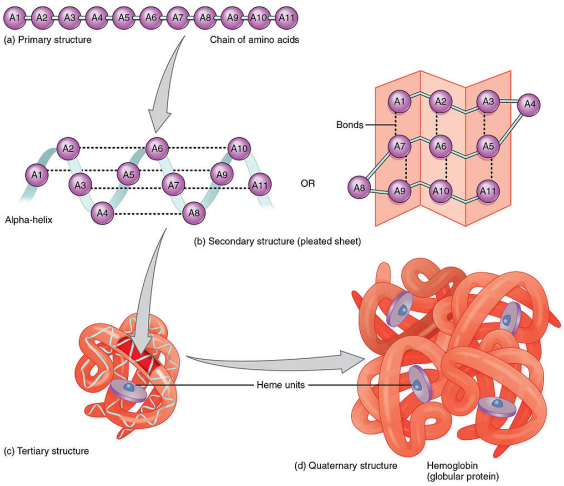

FIGURE 2.1 – *Illustration du repliement d'une protéine - tiré de [58].*

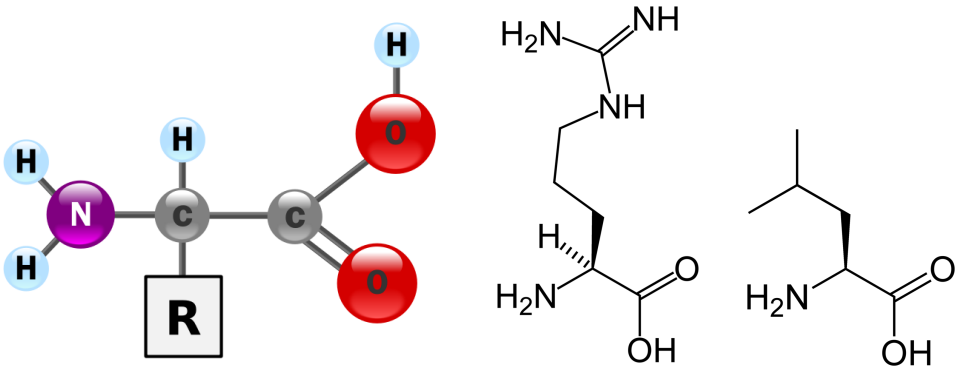

FIGURE 2.2 – *À gauche, architecture d'un acide aminé. À droite, deux exemples d'acides aminés : l'arginine (noté R) et la leucine (noté L) - tiré de [59].*

composition varie entre les vingt types d'acides aminés, tel que cela est illustré sur la Fig. 2.2.

Peptide Un peptide est, tout comme une protéine, une molécule constituée de l'assemblage de plusieurs acides aminés. Cependant, le terme *peptide* est en général réservé aux chaînes d'acides aminés qui sont particulièrement courtes (moins d'une cinquantaines d'acide aminés), ou aux chaînes de longueur quelconque qui sont un fragment d'une protéine complète. Dans le reste de ce document, le terme peptide n'est utilisé qu'au sens de fragment de protéine.

Protéome Le protéome désigne l'ensemble des protéines contenu dans un échantillon biologique. La protéomique est donc l'étude systématique des protéines d'un échantillon.

1.2 Cycle de vie des protéines

Comme indiqué plus haut, les protéines ne sont pas des entités statiques et figées (telles que la structure d'une usine), mais sont au contraire éphémères et dynamiques : elles apparaissent, interagissent entre elles, se modifient mutuellement,

disparaissent, etc. (d'où l'analogie précédente avec les ouvrières et la matière première traitée dans l'usine), de sorte qu'il est tout à fait adapté de parler de *cycle de vie* pour les protéines, même s'il ne s'agit pas d'entités vivantes à proprement parler.

Expression L'*expression* (ou la naissance) des protéines est souvent expliquée de manière simpliste à un informaticien, en prenant la mémoire d'un ordinateur pour modèle : l'ADN (ou *acide désoxyribonucléique*), en tant que très longue molécule constituée de la concaténation d'un nombre gigantesque de paires de bases, peut être comparé à la mémoire physique d'un ordinateur, constituée elle aussi d'un grand nombre de bits. Là où le codage digital est binaire, le codage génétique est quaternaire (c'est-à-dire que pour chaque paire de bases, il y a 4 possibilités, classiquement noté A-T, T-A, G-C et C-G). À une étape de l'expression (qui sera précisée ultérieurement), l'équivalent d'une tête de lecture, l'ARN-polymérase, vient lire ce code par groupes de 3 paires de bases, appelés codons (l'équivalent des octets formant des mots de 8 bits). Chaque codon permet d'identifier de manière unique un acide aminé, de sorte que l'enchaînement des codons permet la définition d'une protéine. Ainsi, il y aurait une correspondance entre une protéine, succession d'acides aminés, et un gène, succession de codons. Pour l'instant, tenons-nous en à cette analogie simpliste selon laquelle les protéines constituent l'exécution par un ordinateur biologique d'un programme écrit dans le code génétique. Nous reviendrons dessus dans le paragraphe suivant.

Epissage alternatif Lors du processus permettant l'expression des protéines, certaines parties du code génétique peuvent être oblitérées, conduisant à des enchaînements tronqués de codons, et donc, donnant naissance à des protéines différentes. Ce processus relativement complexe augmente considérablement le nombre de protéines résultant d'un ruban d'ADN donné.

Modifications post-traductionnelles (ou PTM) Il s'agit d'une modification de la formule chimique de la protéine au cours de son cycle de vie ; par exemple, l'ajout d'un groupe phosphate (PO_4) à un acide aminé (la phosphorylation) ou d'un groupe glycosyl (la glycosylation), la création d'un pont disulfure entre deux cystéines (un acide aminé particulier), etc. Il apparait qu'une modification aussi simple qu'une PTM, ne venant modifier que marginalement un acide aminé, peut considérablement modifier la fonction d'une protéine.

Déplacements Une fois exprimées, les protéines peuvent se déplacer ; elles peuvent aussi être des transporteurs ayant vocation à permettre le déplacement d'autres molécules (entrer et sortir d'une cellule par exemple). Par ailleurs, comme toutes les protéines ne sont pas exprimées dans toutes les cellules d'un organisme, la ou les localisation(s) d'une protéine constitue(nt) un élément majeur de la description de son cycle de vie.

Formation de complexes Des interactions chimiques peuvent se former entre plusieurs protéines qui forment alors un complexe. L'étude des réseaux d'interactions protéines-protéines est un enjeu particulièrement important de la protéomique. Sur de tels réseaux, les méthodes traditionnelles en statistiques sont mises en échec, car les protéines ne peuvent plus être considérées comme des individus i.i.d. (cf. Chap. 1).

Dégradations Finalement, les protéines en fin de vie (devenues obsolètes ou ayant subi une transformation inappropriée) sont détruites par digestion.

Régulation Malgré tous les processus qui entrent en jeu (expression, modifications, déplacements, dégradation, etc.), la quantité de chaque protéine est à tout instant précisément régulée. Pour ce faire, la concentration de chaque protéine a une influence sur les processus d'expression et de dégradation, ainsi que sur ceux d'autres protéines. Cela permet l'apparition de boucles de rétroactions particulièrement complexes, telles que décrites dans le paragraphe suivant.

1.3 Epigénétique et régulation du protéome

Revenons-en à l'analogie informatique permettant de décrire l'expression des protéines. La réalité est en fait beaucoup plus complexe : notamment, la lecture de l'ADN (que l'on appelle la *transcription*) ne donne pas naissance directement à des protéines, mais à des copies de fragments d'ADN, les ARN (ou *acide ribonucléique*) de nature multiple (ARN messagers, ARN de transfert), dont ensuite la *traduction* donnera des protéines.

Pour que la transcription soit possible, la molécule d'ADN s'ouvre en deux (comme une fermeture-éclair), de sorte que le codage n'est pas directement induit par les paires (A-T, T-A, G-C et C-G), mais par la moitié gauche ou droite de celles-ci (A, T, G, C). Ainsi, tout un ensemble de fonctions particulièrement complexes permettent de réguler cette lecture : dans quel sens faut-il lire ? De quel côté de l'ouverture ? En partant d'où et en s'arrêtant où ? Etc.

Mais le fait que l'ADN puisse ainsi s'ouvrir en deux pour permettre la transcription est déjà un processus particulièrement complexe, qui dépend de l'état de l'ADN et des molécules qui l'entourent, le structurent et le "rangent". En effet, la longue chaîne de l'ADN est stockée de manière "efficace", enroulée sur elle-même et sur un ensemble de protéines (les histones), telle que cela est illustré sur la Fig. 2.3. Les histones, en tant que protéines, sont bien sûr elles-mêmes issues de la transcription, puis de la traduction de gènes, et évoluent selon le cycle de vie décrit plus haut. Leur présence ou absence à certains emplacements permet à l'ADN de s'enrouler en des zones plus ou moins compactes, en un ensemble que l'on appelle la *chromatine*. Du niveau de condensation de la chromatine, directement relié à la présence de certaines protéines, découle la capacité de l'ADN à s'ouvrir et à être transcrit.

Ainsi, une partie de l'information génétique, notamment le fait que certains gènes soient accessibles à la transcription ou pas, ne relève pas directement de l'ADN lui-même, mais de la structure complexe de la chromatine permettant son stockage et sa mise en forme. L'étude de cette génétique "hors ADN" et des processus de régulation permettant l'expression des gènes se nomme l'*épigénétique*, et son importance croissante devrait finir par mettre à mal l'image ultra-vulgarisée d'un ADN-programme, et de protéines résultant de l'exécution de ce programme.

Cependant, si une analogie informatique doit être conservée, celle du FPGA (*field-programmable gate array*, ou plus généralement, de tout autre circuit électronique reconfigurable) est plus juste. Cependant, sa complexité serait démentielle, et il aurait la capacité de reprogrammer ses circuits, en fonction du contenu des informations qu'il possède en mémoire, ou résultant d'un calcul, permettant une évolution proactive et autorégulée.

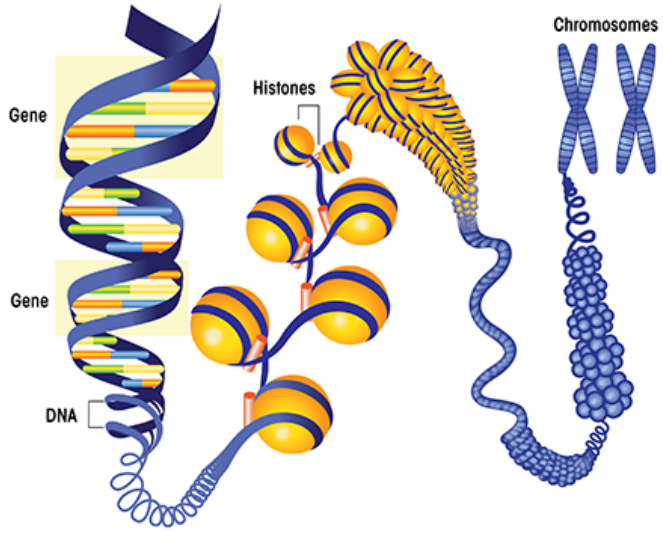

FIGURE 2.3 – *Structure de la chromatine - tiré de [60].*

2 Une brève histoire de la “-omique”

2.1 De la génétique à la génomique

La génétique a commencé par l'étude de la transmission héréditaire des phénotypes (les célèbres petits pois de Gregor Mendel en 1866), principalement à des fins botaniques. Suite à la découverte, en 1944, selon laquelle l'ADN est le support de l'hérédité, la génétique a été associée à l'isolation de gènes spécifiques, afin de comprendre le développement embryonnaire, l'évolution des espèces, des maladies héréditaires, ou des spécificités physiologiques d'une population. En un demi-siècle, la génétique est devenue un élément central de la biologie, et s'est rapidement diversifiée en plusieurs branches, chacune trouvant toujours plus d'applications et autant d'échos dans la société civile.

Cependant, quel que soit le niveau de cette diversification, nous pouvons constater que les découvertes en génétique ont principalement résulté d'une démarche analytique : isolement d'un gène, étude de tous ses interactants possibles dans un grand nombre de situations, analyse détaillée de ses différentes formes, etc. afin de pouvoir, pour le gène d'intérêt, en faire une description la plus précise et la plus exhaustive possible. À ce titre, la *génomique* fait exception. En effet, la génomique ne procède pas de cette démarche analytique, mais au contraire, d'une démarche à la fois synthétique et systémique. Il s'agit non plus de s'intéresser à un gène unique pour le décrire et le comprendre de manière exhaustive, mais plutôt d'avoir une vision d'ensemble, permettant d'appréhender tout le *génome* (c'est-à-dire l'ensemble des gènes pris simultanément) en une seule fois.

2.2 De la génomique à la protéomique

La protéomique se positionne dans la continuité directe de la génomique : la première est à la protéine, ce que la seconde est au gène.

Dans la mesure où le génome est la cause directe de l'expression des protéines, la finalité de la génomique et celle de la protéomique sont particulièrement proches, et les deux disciplines largement interdépendantes. Ainsi, il est courant d'utiliser des bases de données génomiques pour définir à la volée et de manière systématique les séquences protéiques (la liste des acides aminés composant chaque protéine) attendues, afin de pouvoir plus facilement les identifier dans un échantillon. De même, plutôt que de s'intéresser aux gènes présents sur l'ADN, la *transcriptomique* s'intéresse au code porté par les ARN messagers (ARNm), dont le rôle est central dans l'expression des protéines. Ainsi, en étant capable d'identifier des ARNm et de quantifier le nombre de copies présentes dans le noyau d'une cellule pour chacun d'eux, il est possible d'avoir accès à l'intensité de transcription des gènes correspondants, et donc indirectement, d'avoir une idée de la production des protéines correspondantes, avec un point de vue complémentaire de celui fourni par la protéomique.

Il y a cependant deux types de différences notables entre les données de génomique et de protéomique. Tout d'abord, celles qui relèvent du protocole expérimental et de la chaîne d'acquisition de ces données, que je décris dans la suite de ce document ; ensuite, il y a le changement d'échelle de complexité entre génome et protéome, et qui s'opère au moment des différentes phases de l'expression (transcription et traduction). Au-delà de l'aspect dynamique et de l'évolution dans le temps et dans l'espace de l'abondance des différentes protéines, il y a toutes leurs interactions, formations de complexe et PTM qui viennent multiplier la complexité de la tâche de séquençage. Enfin, d'un point de vue combinatoire, alors que les gènes peuvent être assimilés à des mots écrits avec un alphabet de 4 lettres (les paires de bases), l'alphabet des protéines contient 20 lettres (les acides aminés), rendant leur séquençage d'autant plus complexe.

Cela explique en grande partie pourquoi la protéomique est une discipline beaucoup plus récente que la génomique (il y a approximativement 15 ans de décalage entre leurs balbutiements respectifs, et ce décalage est maintenu dans le temps en termes de résultats ou de fiabilité). Mais cela explique aussi pourquoi les méthodes de traitement de données en protéomique sont directement issues de la génomique ou de la transcriptomique, selon un schéma particulièrement répétable :

1. Mise en place d'une nouvelle méthode d'acquisition haut débit ;
2. Phase de questionnement quant au traitement de ces données (il est important de comprendre que la question du traitement des données n'apparaît qu'après la maîtrise de leur production) ;
3. Identification dans l'état de l'art d'une méthode issue de la transcriptomique ou de la génomique, suivie de son application.
4. Remise en cause de la méthode : à y regarder de plus près, des modifications sont nécessaires afin de tenir compte des spécificités de la protéomique.

Mon rôle au sein d'un laboratoire de protéomique est de renforcer ce dernier point.

2.3 L'explosion des *omics*

Comme mentionné un peu plus haut, l'ADN, en tant que support initial du génome, a d'abord été au centre des attentions. Ensuite, une fois celui-ci séquencé, l'attention s'est déportée vers les ARNm, dont la structure chimique reste très proche

de celle de l'ADN, tout en fournissant des informations essentielles sur le processus de traduction ; et donc ouvrant une porte sur le protéome. Ainsi, on a naturellement trois "règnes" : celui des gènes et de l'ADN (le génome), celui des transcrits et de l'ARN (le transcriptome), et celui des protéines et des acides aminés (le protéome). De nombreuses recherches en biologie nécessitent une validation à chacun de ces trois niveaux, de sorte que le terme "approche multi-omique" est maintenant courant.

Les personnes intéressées par un inventaire des approches multi-omiques peuvent se référer à des pages web qui les dénombrent [61, 62], en un long inventaire à la Prévert. À côté du lipidome (qui s'intéresse à la grande classe des lipides) et du métabolome (les métabolites sont les petites molécules issues du métabolisme) qui font naturellement sens, il y a des définitions plus originales, telles que le connectome (l'ensemble des connexions neuronales) ou l'allergome (le protéome des allergènes) voire même franchement exotiques : le bibliome (la bibliographie en biologie), le foodome (qu'il est difficile de définir autrement que par le "ome" de la nutrition), et même le researchsome, (qui a été proposé initialement de manière ironique, avant d'être consacré comme le domaine de recherche d'un individu ou d'une organisation).

Clairement, les suffixes "ome" et "*omics*" semblent être à la mode, et servir à définir des *buzzwords* à la pelle, plus dans une optique de communication scientifique, que de réelle science. Cependant, au-delà de cette caricature, l'augmentation des vrais termes en *omics* est réelle, et de mon point de vue, ils ne sont à la biologie que ce que le *big data* est à l'informatique ou aux statistiques : le marqueur communicationnel d'une évolution de la discipline.

2.4 Le paradigme de la biologie à grande échelle

Du point de vue du biologiste, chaque "bidule-omics" est une voie d'étude supplémentaire, ou un outil additionnel, parmi tant d'autres pour mieux comprendre le sujet "bidule". Du point de vue du *data scientist*, il s'agit d'une véritable révolution de la discipline biologique.

Selon mon image d'Épinal (qui doit être partagée par un certain nombre de néophytes en biologie), un biologiste est avant tout un observateur particulièrement patient (capable de décrire précisément, voire de dessiner son sujet d'étude) muni d'un savoir encyclopédique permettant de mettre ces observations en contexte (au sein de la célèbre taxonomie animale par exemple), que j'imagine en action en train d'observer un papillon ou une bactérie, à la loupe ou au microscope. Finalement, dans mon imaginaire, ce qui définit le biologiste tient autant dans sa méthode (faite de descriptions et de comparaisons, mais avant tout analytique) que son sujet d'étude (le vivant). Les *omics* ne rentrant à priori pas dans ce cadre de méthodologie analytique, on comprendra sans peine qu'elles suscitent tant l'intérêt, et qu'on parle de révolution. Si le modèle synthétique et systémique mis en place dans les *omics* se généralise, on pourra bientôt distinguer deux types de biologie : la biologie traditionnelle, fondée sur une approche analytique, et une nouvelle biologie, tour à tour nommée *biologie à grande échelle*, *biologie des systèmes*, ou *biologie à haut débit d'analyse*[1] (de l'anglais *high throughput biology*), et que le terme "approches

1. Ici le terme "analyse" est intéressant : il souligne que dans tous les cas, l'aspect analytique de la biologie est indispensable, et qu'il ne pourra pas être remplacé par une autre forme de biologie. Cependant, il pointe aussi du doigt le fait que cet aspect analytique peut être automatisé, détachant

omiques" définit très mal.

De mon point de vue, l'évènement qui a marqué réellement la naissance de cette discipline et l'acceptation de sa méthodologie pour investiguer le vivant, est une publication datant de fin 2008 dans la revue Nature [63], cosignée par des chercheurs de Google (et déjà citée près de 2200 fois, soit environ trois fois tous les quatre jours depuis sa parution), décrivant le logiciel Google Flu Trends [64] qui permet de prédire l'arrivée annuelle de l'épidémie grippale sur la base des recherches effectués (via le moteur Google) par les internautes avec des mots-clefs comme "paracétamol", "maux de gorge", etc. et dont les prédictions sont plus fiables que celle du réseau Sentinelle, chargé de collecter, de faire remonter et d'interpréter toutes les observations réalisées par les médecins généralistes.

Que peut-on en conclure ? Il sera toujours absolument nécessaire d'observer le fond de la gorge d'un patient pour déterminer la nature de sa grippe et le soigner en conséquence. Cependant, au-delà de la question médicale individuelle, la question de santé publique ou d'épidémiologie a aussi besoin d'être considérée. Dans la mesure où il s'agit de questions différentes, il est normal d'utiliser des méthodes différentes pour y répondre. Jusqu'à présent, les cursus universitaires de biologie privilégiaient la démarche analytique à juste titre, celle-ci ayant été la seule à permettre une réelle compréhension du vivant. Avec l'avènement des outils d'analyse automatique d'échantillons biologiques (dont certains sont décrits dans la suite de ce chapitre) et avec les outils de traitement automatique de l'information ainsi générée, il devient possible d'espérer généraliser ces méthodes. Peut-être dans quelques années, sera-t-il communément accepté qu'il y ait deux types de biologies, correspondant à des méthodes et des métiers différents. Les cursus universitaires proposeront les deux, et il sera naturel de faire collaborer des experts de ces deux types de biologies dans les laboratoires de recherche.

2.5 Protéomique de découverte

L'avènement de la biologie à grande échelle n'est encore que partiel, et pour l'instant, cette discipline est encore subordonnée à la biologie traditionnelle. Ainsi, la manière de conduire une analyse protéomique et de l'insérer dans un projet de recherche dépend principalement de la question biologique, et des outils autres que la protéomique qui sont mis en œuvre pour y répondre. Dans ce contexte, je me suis focalisé sur la *protéomique de découverte* : il s'agit d'un type particulier d'analyses protéomiques, qui possèdent trois caractéristiques : tout d'abord, induire la production de gros volumes de données, sur lesquelles mes compétences peuvent s'exercer ; ensuite, être suffisamment générique pour pouvoir s'insérer dans un grand nombre d'investigations biologiques, et donc être transposable à de nombreux projets ; enfin, être suffisamment mature pour être considéré comme une technologie relativement standard (dans une logique de plateforme de soutien à la recherche) à adapter à différent projets de recherche biologique, plutôt que comme une méthode dont le développement est un sujet de recherche en tant que tel.

Le principe de la protéomique de découverte est le suivant : comparer les protéomes d'au moins deux conditions biologiques différentes, afin de pouvoir dans

le biologiste de ces tâches pour pouvoir se concentrer sur d'autres formes de raisonnements.

un premier temps, relier la différence de nature des conditions à une variation des protéines exprimées (en termes d'identités ou de quantités relatives) dans ces conditions ; puis dans un second temps, transmettre au biologiste une caractérisation de la différence entre ces protéomes ; étant entendu que l'usage que ce dernier fera d'un tel résultat est très variable, et dépendra principalement de la question biologique à laquelle il cherche à répondre. La nature des conditions dont on compare les protéomes est très variable. Il peut s'agir par exemple :

- De cohortes de patients, atteints de différentes variations d'une même maladie (stade plus ou moins avancé, résultant de différentes sources de pathogènes, etc.) avec en plus pour référence, une cohorte de personnes saines ;
- Souris ayant reçu différentes drogues (un groupe de souris non droguées servira alors à établir le protéome de référence) ;
- D'espèces quelconques (plantes vertes, levures, etc.) dont un gène a été muté ou supprimé (nommées *mutant* ou *K.O.*), par rapport à des individus de la même espèce, mais non mutés (appelés "*wild type*") ;
- D'échantillons biologiques dans lesquels on a respectivement "péché" par affinités physico-chimiques (cf. Chap. 3, Sec. 3.1), ou non, les interactants de certaines protéines clefs, afin de définir, par soustraction, l'ensemble des interactants ;
- De compartiments biologiques différents, afin de déterminer où se trouvent certaines protéines.

Bien sûr, la protéomique ne se limite pas à la protéomique découverte. Cependant, comme mon travail se focalise sur celle-ci, je ne décris pas les autres approches.

3 Spectrométrie de masse et protéomique

Initialement, l'étude des protéines était principalement conduite par des méthodes biochimiques, nécessitant d'isoler la protéine d'intérêt avant de l'étudier sous l'angle de son comportement chimique. C'est l'arrivée de méthodes physiques, telles que la spectrométrie, qui a permis de considérablement augmenter le débit de protéines analysées ; et qui a donc indirectement nécessité la mise en place de méthodes de traitement de données adaptées.

3.1 Principe de la spectrométrie de masse

Un spectromètre de masse peut être principalement vu comme une balance d'une très grande précision (de l'ordre d'une petite dizaine de parties par million). L'usage du spectromètre en protéomique repose sur le principe que si l'on arrive à peser précisément un (ou plusieurs) acide(s) aminé(s), il est possible de le(s) identifier, puisque pour la plupart, les acides aminés n'ont pas la même masse (à l'exception de la Leucine et de l'Isoleucine). Bien sûr, ce principe de base doit être un peu raffiné pour identifier de longues séquences d'acides aminés.

Il existe de nombreuses technologies de spectrométrie de masse, cependant, toutes respectent globalement la même architecture, constituée de trois éléments : la *source d'ionisation*, l'*analyseur* et le *détecteur*. Les molécules d'intérêt (la plupart du temps, des peptides, mais pas toujours comme nous le verrons plus loin) sont d'abord diri-

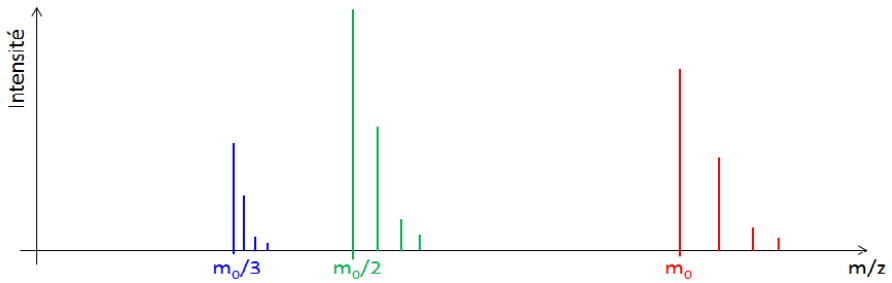

FIGURE 2.4 – *Représentation schématique du spectre de masse pour une molécule de masse m_0 : la molécule a atteint trois états de charges différents : 1+, 2+, et 3+, ce qui explique les 3 groupes de raies de couleurs différentes, aux abscisses m_0, $m_0/2$ et $m_0/3$. Par ailleurs, en raison des propriétés physico-chimiques de cette molécule, il y a beaucoup d'entités qui ont une charge 2+, et peu qui ont une charge 3+, ce qui explique la hauteur des raies correspondantes. Enfin, Il y a plusieurs raies pour chaque état de charge en raison de la présence d'isotopes : il y a environ 1.1% du carbone terrestre qui est du ^{13}C pesant 13/12 fois plus que le carbone classique (le ^{12}C), de sorte que dans une grosse molécule contenant une longue chaîne carbonée, il peut y avoir un ou plusieurs isotopes. Cela explique pourquoi on observe des* massifs isotopiques *de la forme d'une distribution binomiale (avec un intervalle entre les raies inversement proportionnel à l'état de charge).*

gées dans la *source d'ionisation*, afin de les faire passer d'un état électrique neutre à un état chargé (éventuellement, la molécule peut contenir plusieurs atomes chargés). Ensuite, elles sont dirigées vers l'*analyseur*, dans lequel un champ magnétique permet une déviation de la course de toutes les molécules chargées. Cette déviation dépendant principalement du rapport m/z (la masse de la molécule divisée par son état de charge), il est possible d'après la trajectoire de la molécule, de déterminer sa masse, modulo sa charge. Enfin, le *détecteur* compte les ions analysés.

La sortie du spectromètre est un *spectrogramme*, ou un *spectre* : il s'agit d'un graphe sur lequel sont indiqués le rapport m/z en abscisse, et une mesure d'intensité en ordonnée. Ainsi, si une seule espèce chimique est introduite dans un spectromètre, le spectrogramme résultant correspondra à l'illustration de la Fig. 2.4. En pratique, les échantillons classiquement analysés contiennent de nombreux peptides, de sorte que leur spectres ressemblent plutôt à celui de la Fig. 2.5.

3.2 Spectres de fragmentation

Naturellement, si la molécule que l'on va "peser" avec le spectromètre est particulièrement complexe (ce qui est classique en protéomique, puisqu'il s'agit d'analyser une longue chaîne d'acides aminés), son poids ne permet plus de l'identifier : le nombre de combinaisons d'acides aminés dont la somme est égale à celle mesurée peut être élevé. De plus, l'ordre des acides aminés peut modifier radicalement la protéine sans conséquence sur son poids total.

C'est pour cela que l'on rajoute dans le spectromètre une chambre de collision, entre la source d'ionisation et l'analyseur. Les différentes copies de la chaîne d'acides aminés étudiée (qu'il s'agisse d'un peptide ou d'une protéine) vont pénétrer dans

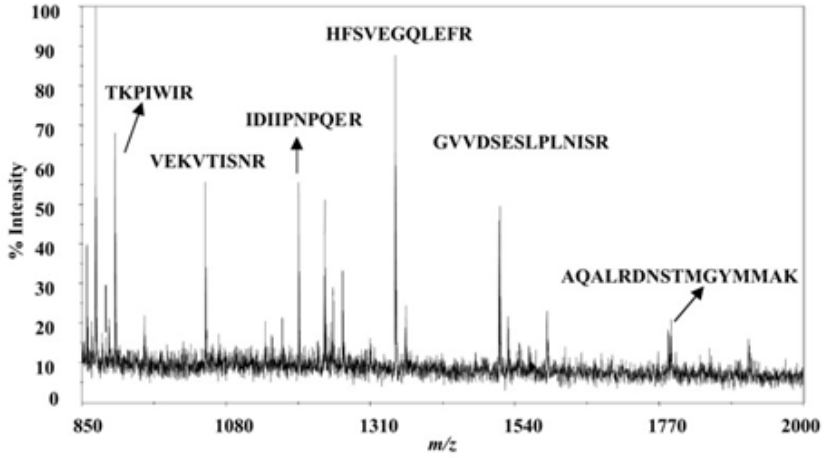

FIGURE 2.5 – *Spectre de masse d'un échantillon contenant plusieurs peptides (sur une telle échelle permettant de visualiser une large gamme de m/z, les massifs isotopiques ne sont pas discernables, mais présents malgré tout) - tiré de [65].*

cette chambre et rentrer en collision avec les molécules d'un gaz inerte, conduisant à leur fragmentation sous l'effet du choc. Nous pouvons espérer que dans le cas idéal :

- Toutes les copies de la chaîne d'acides aminés vont se scinder en exactement deux parties (ni plus, ni moins) ;
- Cette fragmentation n'aura lieu qu'entre deux acides aminés, et ne scinde pas un acide aminé en deux (rappelons que les acides aminés sont en eux-mêmes déjà des molécules complexes) ;
- Le lieu de la fragmentation sera aléatoire et se répartira de manière uniforme sur les liaisons entre acides aminés, de sorte que sur l'ensemble des copies injectées, toutes les possibilités de fragmentation seront représentées ;
- la charge résultante de l'ionisation (réalisée juste avant) se trouvera avec une probabilité égale sur le premier ou le second fragment.

Si toutes ces hypothèses étaient vérifiées, il serait trivial d'identifier une chaîne d'acides aminés sur la base de son spectre de fragmentation : celui-ci contiendrait exactement $2N$ pics, soit deux pics pour chacun des N lieux de fragmentation possibles, pour une chaîne de $N+1$ acides aminés. Les fragments étant respectivement de longueur n et $N-n+1$, $\forall n \in [1, N]$, et leur appariement étant possible (en cherchant les paires de fragments dont le poids est constant et égal à la molécule d'intérêt), le calcul des différences de masse entre des pics successifs permettrait de décoder la séquence d'acides aminés, tel qu'illustré sur la Fig. 2.6.

Malheureusement, les hypothèses ci-dessus ne sont en pratique pas vérifiées, de sorte que les raies spectrales n'ont pas toutes la même hauteur (certaines étant inexistantes) et que certains pics ne correspondent qu'à un fragment central de la chaîne ; de plus, il faut rajouter les massifs isotopiques et les multiples états de charges précédemment décrits, ainsi que le fait que les fragmentations peuvent avoir lieu à différents endroits sur la liaison entre acides aminés (cf. Fig. 2.7) impliquant des variations sur les écarts de masses. Finalement, le spectre de fragmentation réel d'un peptide est beaucoup plus difficile à lire (cf. Fig. 2.8).

En conséquence de quoi, l'identification d'une chaîne d'acides aminés, même si

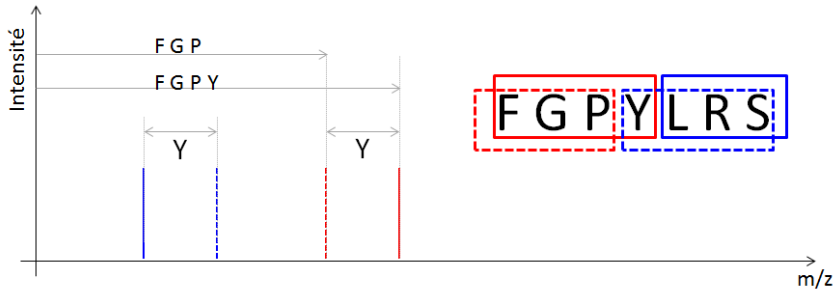

FIGURE 2.6 – *Principe de lecture d'un spectre de fragmentation.*

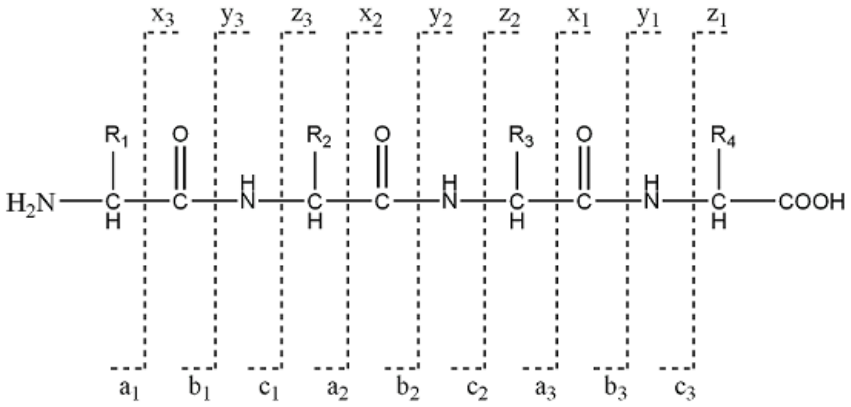

FIGURE 2.7 – *Entre 2 acides aminés, il y a 3 lieux de fragmentation possibles, référencés respectivement par ax, by et cz, donnant un fragment de type a (respectivement b ou c) appareillé avec un fragment de type z (respectivement y ou x). Ainsi, pour une chaîne de 4 acides aminés (avec les radicaux* $\mathbf{R}_1, \mathbf{R}_2, \mathbf{R}_3$ *et* $\mathbf{R}_4$*), il y a 9 lieux de dissociations possibles - tiré de [66].*

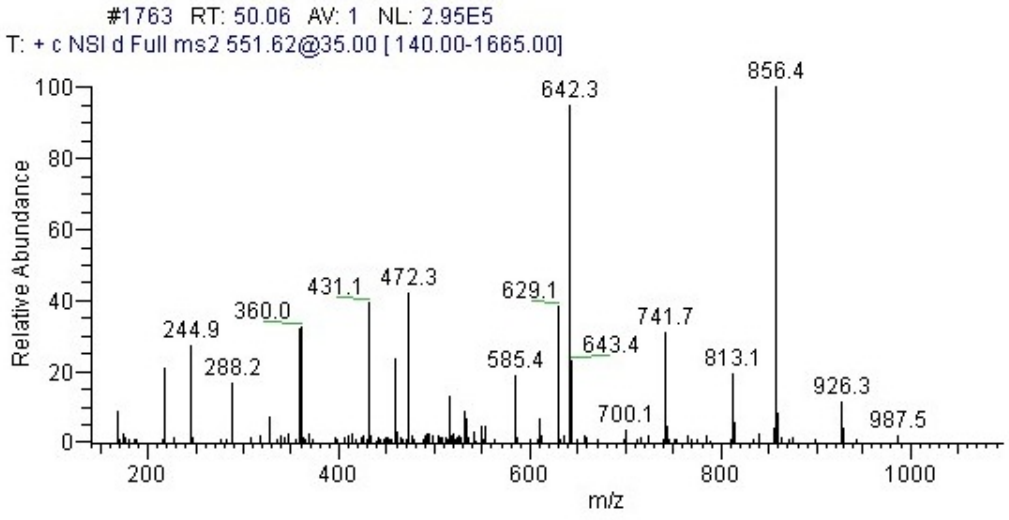

FIGURE 2.8 – *Spectre de fragmentation réel - tiré de [67].*

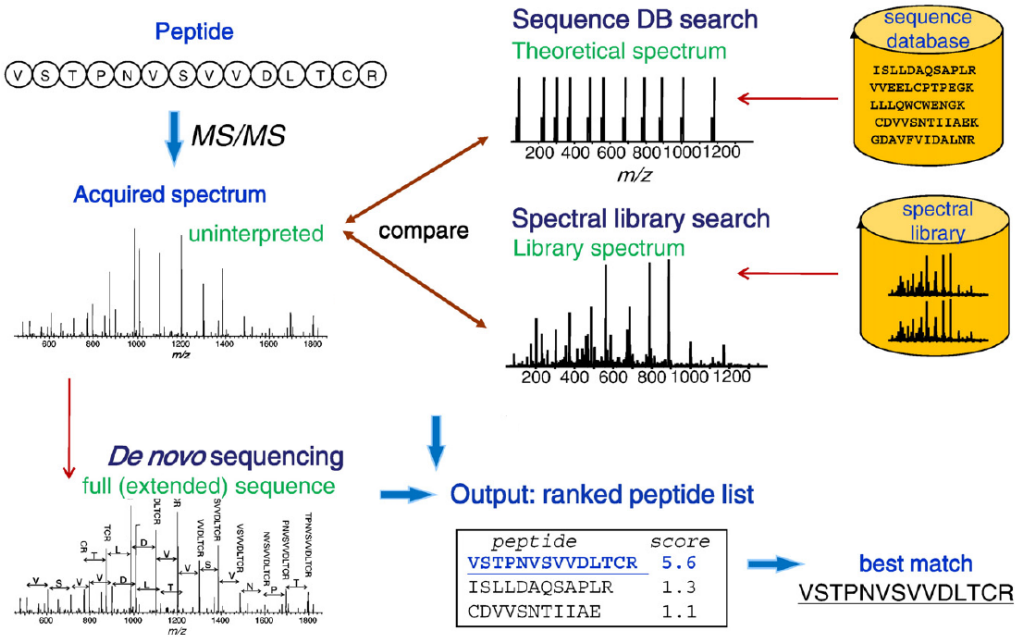

FIGURE 2.9 – *Les trois principales méthodes d'identification - tiré de [68].*

elle est courte, est un problème difficile. Il y a en gros 3 familles d'algorithmes permettant de réaliser cette identification (cf. Fig. 2.9, [68]). La première consiste tout simplement à décoder le spectre sur le principe de ce que j'ai décrit plus haut (on appelle cette méthode *identification De Novo*, ou *séquençage De Novo*). Une seconde solution consiste à utiliser des spectres précédemment identifiés pour en identifier des nouveaux, dans une logique d'apprentissage selon l'algorithme du plus proche voisin (Ces méthodes sont dites "basées sur des *librairies spectrales*"). La dernière méthode fonctionne sur le même principe que la précédente, mais en utilisant à la place des librairies spectrales, des spectres théoriques produit par une simulation d'analyse spectrométrique, sur la base des protéines que peuvent produire un génome ; ces approches sont appelées "*database search*", et sont de nos jours les plus populaires. Au laboratoire EDyP, une telle stratégie a été mise en place, via l'utilisation du logiciel commercial Mascot (un "*database search engine*' ou DBSE').

3.3 Principe de la chromatographie

Nous avons vu précédemment que le spectre d'un échantillon ne contenant qu'une seule entité est déjà relativement complexe. Dès lors, il est compréhensible qu'un échantillon biologique réellement complexe, c'est à dire contenant plusieurs centaines de protéines ou plusieurs milliers de peptides, va donner naissance à un spectrogramme complétement illisible. C'est pour éviter cela qu'une étape préliminaire de séparation est presque toujours nécessaire. L'objectif de celle-ci est de regrouper entre elles les différentes copies d'une même entité chimique, et de séparer autant que possible les différentes entités chimiques, afin de pouvoir, dans le cas idéal, les injecter dans le spectromètre les unes à la suite des autres. Bien que ce cas idéal ne soit pas atteignable pour des échantillons trop complexes, il est néanmoins possible d'étaler dans le temps, l'arrivée des différentes entités, afin qu'à un instant donné, un

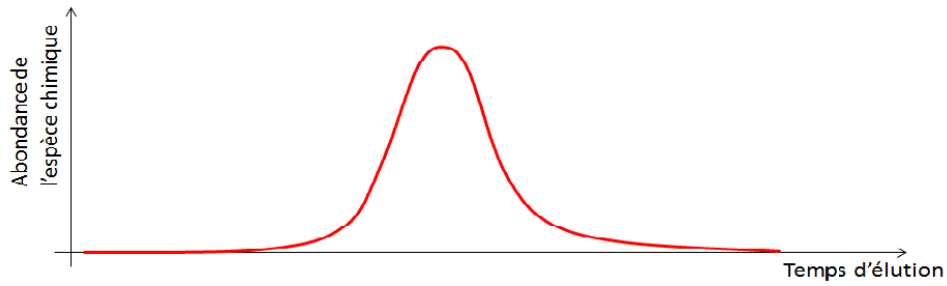

FIGURE 2.10 – *Représentation schématique de l'abondance d'une espèce chimique en sortie de chromatographie, au cours de ton temps d'élution.*

nombre plus faible d'entre elles soient co-analysées. Pour cela, il est courant d'utiliser une *chromatographie liquide* (de nombreuses autres techniques existent, mais c'est celle privilégiée au laboratoire, et donc la seule que je présente ici).

Une chromatographie liquide est constituée principalement d'une colonne d'élution, et d'un injecteur. Ce dernier servant avant tout à la connexion avec le spectromètre, concentrons-nous sur la colonne d'élution : il s'agit d'une sorte de long tube que les différentes entités chimiques ne vont pas parcourir à la même vitesse en fonction de leurs propriétés physico-chimiques (un peu à la manière d'un signal dont les différentes harmoniques ne se propagent pas à la même vitesse dans un matériau, induisant une déformation du signal d'origine). Au regard du temps d'analyse nécessaire à un spectromètre, il va se passer un temps particulièrement long entre l'arrivée de la première et la dernière molécule d'un échantillon ; de sorte que dans l'intervalle, le spectromètre pourra réaliser plusieurs milliers d'analyses, chacune contenant seulement une fraction de l'échantillon biologique complet.

En pratique, à une légère variance prés, toutes les copies d'une entité chimique se déplacent à la même vitesse dans la colonne d'élution, de sorte que pour chaque entité, il est possible de définir un chromatogramme (cf. Fig. 2.10). Malheureusement, entre deux expériences, même si la colonne d'élution est la même, le chromatogramme n'est que très difficilement reproductible : il arrive souvent qu'il y ait des distorsions, rendant difficile (mais pas impossible [69]) l'usage du temps d'élution pour l'identification de l'espèce chimique en question.

3.4 Quantification avec ou sans marquage isotopique

Comme détaillé plus haut, il est possible d'identifier les espèces chimiques injectées dans le spectromètre. Maintenant, examinons les possibilités que cet outil offre pour quantifier les molécules identifiées [70, 71].

Comme cela apparaît sur la Fig. 2.4, l'intensité des raies du spectrogramme sont d'autant plus importantes que l'espèce chimique est présente. Ainsi, si Q désigne sa quantité, et I l'intensité de la raie, il est possible de définir une fonction bijective f qui approxime le lien entre Q et I, de sorte qu'il est possible de déduire l'un à partir de l'autre et inversement : $I = f(Q)$ et $Q = f^{-1}(I)$. Malheureusement, f n'est pas linéaire, dépend de l'entité chimique (chacune a une réponse particulière), mais aussi de la présence d'autres entités chimiques dans le spectromètre (en raison, notamment du fait que les différentes entités rentrent en compétition dans la phase

d'ionisation). Il en résulte que :

- Il n'est en pratique pas possible de convertir d'intensité du signal en une quantité de molécules ou en une concentration dans l'échantillon. L'intensité du signal correspond à une *abondance*, une grandeur sans unité, qui ne reflète que grossièrement la quantité de l'espèce chimique ;
- Les abondances de différentes entités chimiques analysées simultanément ne peuvent être comparées les unes aux autres.

Dans ce contexte, il ne reste que deux solutions : la première consiste à rajouter des entités chimiques similaires, simplement modifiées par la présence de quelques isotopes, en quantités précisément connues. Ainsi, pour chaque espèce chimique, il sera possible de mesurer l'intensité des raies des versions originale et marquée, d'en faire le ratio et de multiplier cela par la quantité connue de la version marquée pour en déduire l'abondance de la version non marquée. Naturellement, cela nécessite un dispositif expérimental lourd qui ne peut être appliqué à un grand nombre de peptides ou de protéines. Dès lors, ces approches sont réservées aux problèmes dits *ciblés*, où l'on ne s'intéresse qu'à très peu d'espèces (comme dans les études précliniques, où le nombre de biomarqueurs candidats pour une maladie donnée est faible), et ne correspondent pas aux contraintes de la protéomique de découverte.

L'alternative consiste à n'utiliser la mesure d'abondance que comme un moyen de comparer relativement des échantillons biologiques entre eux. Ainsi, il est possible de regrouper plusieurs échantillons représentant une première condition, et de les comparer à une autre condition de référence, afin de regarder pour chaque protéine, le ratio d'abondance entre les deux conditions (ou du moins la différence dans les distributions d'abondance conditionnellement aux conditions). C'est ce que l'on appelle de la *quantification relative label-free* (du fait de l'absence de marquage isotopique), à laquelle la suite de ce document est consacrée.

3.5 Les principaux pipe-lines

Bien qu'il soit tout à fait possible d'analyser des protéines entières au spectromètre de masse (analyses dites *top-down* [72]), cela ne constitue pas la méthode la plus répandue en protéomique de découverte. En effet, les protéines sont des molécules particulièrement grosses et complexes : chaque acide aminé ayant une masse molaire comprise entre 75 et 210 g/mol (un acide aminé est donc entre 6 et 18 fois plus lourd qu'un atome de carbone), les protéines font classiquement entre quelques milliers et un million de g/mol. Il est donc difficile de les étudier au spectromètre de masse, car leurs spectres de fragmentation sont extrêmement complexe. Ainsi, la méthode privilégiée (dite *bottom-up*) consiste d'abord à digérer *in vitro* les protéines avec une enzyme (classiquement la trypsine), afin de les découper en peptides, puis d'injecter ces derniers dans le spectromètre [2]. À partir de l'identification des peptides, on remonte ensuite à l'identification des protéines avant digestion [68]. Cette démarche simplifie l'analyse spectrométrique, mais complique l'interprétation des résultats (peptides communs à plusieurs protéines, propagation des erreurs au niveau peptidique vers le niveau protéique, etc.).

2. Attention, cette digestion en plus petites molécules ne remet pas en cause la fragmentation, qui reste une étape nécessaire au séquençage de la chaîne d'acide aminés.

Algorithm 1: Algorithme d'analyse en spectrométrie en tandem.

```
for t_i ∈ {t_1, t_2, ..., t_endElution} do
    MS_{t_i} ← getMS1() ;
    W_{t_i} ← defineListOfWindows(MS_{t_i}) ;
    repeat
        for each w_j ∈ W_{t_i} do
            MSMS_j ← getMS2(w_j, t) ;
    until t = t_{i+1};
```

L'identification de spectres de peptides est un processus complexe donnant lieu à de nombreuses erreurs. Afin de limiter celles-ci, la plupart des outils d'identification se basent sur deux informations : bien sûr, le spectre de fragmentation du peptide en question, mais aussi la masse totale du peptide chargé (aussi appelé *ions précurseur*) avant sa fragmentation. Cela explique pourquoi les spectromètres sont utilisés en alternance pour réaliser des spectres du type de la Fig. 2.5 (dits *spectres MS1* ou *spectres MS*), et des spectres du type de la Fig. 2.8 (dits *spectres MS2* ou *spectres MS/MS*) : c'est ce qu'on appelle la *spectrométrie en tandem*.

Concrètement, le spectromètre itère à très grande vitesse selon l'Algorithme 1 : chaque itération commence par l'obtention d'un spectre MS1 fournissant la masse de tous les ions précurseurs présent dans l'analyseur à un instant t_i (au cours de l'élution). Ensuite, un ensemble de fenêtres centrées sur des masses d'intérêt est défini. Enfin, un cycle d'analyse MS2 est enclenché (jusqu'à la prochaine analyse MS1) : durant ce cycle, toutes les fenêtres d'intérêt sont considérées tour à tour (éventuellement plusieurs fois), et un spectre de fragmentation est généré pour chacune d'elles. Ainsi, pour chaque spectre de fragmentation MS2, la liste des fenêtres de masse ainsi que le spectre MS1 des ions précurseurs (réalisé en début du cycle) permet de retrouver la masse des peptides fragmentés.

Une part importante de la stratégie d'analyse est déterminée par la manière dont la liste des fenêtres de masse est construite. Il y a deux familles de stratégies :

Data Dependent Analysis (DDA) : La liste des fenêtres est définie contextuellement, de manière itérative en parcourant le spectre MS1. Cela commence par la recherche du pic le plus élevé, correspondant au précurseur le plus intensément mesuré (probablement un peptide très facilement mesurable, et par ailleurs, très abondant) : Une fenêtre très étroite est simplement définie autour de sa masse, afin d'avoir une tolérance quant à la précision de mesure, tout en évitant de sélectionner d'autres précurseurs ayant une masse proche. Une fois cela réalisé, on fait de même avec le deuxième pic le plus abondant, et ainsi de suite. L'avantage de cette méthode est indéniable : elle permet de limiter le risque d'obtenir un spectrogramme contenant les spectres de fragmentation entrelacés de plusieurs précurseurs. L'inconvénient est cependant qu'il n'est pas possible d'être exhaustif, et que seul les N précurseurs les plus abondants sont fragmentés, les autres étant oubliés.

Data Independent Analysis (DIA) : Les fenêtres sont définies à l'avance, indépendamment du contenu des spectres MS1. Par exemple, il est possible de découper la gamme m/z en N fenêtres de largeur comparable, et de fragmenter

le contenu intégral de l'une des fenêtres à chaque itération MS2. Ainsi, avec autant de spectres de fragmentation qu'en mode DDA (ici N), la couverture des ions précurseurs est exhaustive. En revanche, chaque spectrogramme de fragmentation contient le spectre de plusieurs précurseurs, et leurs raies spectrales sont entremêlées, de sorte qu'il n'est plus possible de lire la masse de chaque acide aminé par différence successive, comme expliqué sur la Fig. 2.6.

À l'heure actuelle, les méthodes DDA sont prépondérantes, en raison principalement de leur facilité de mise en œuvre. Cependant les méthodes DIA, plus récentes, commencent à gagner du terrain [73], pour la seule raison qu'elles sont perçues comme un passage nécessaire à l'amélioration de la couverture des analyses protéomiques.

4 Finalement, qu'est-ce que la protéomique ?

4.1 Entre biologie et chimie - entre recherche et ingénierie

Jusqu'ici, un lecteur *data scientist* doit avoir du mal à comprendre précisément en quoi consiste la recherche en protéomique, et à quoi peuvent bien s'occuper les membres d'un laboratoire de ce domaine. La raison est simple : alors que le sujet d'étude relève clairement de la biologie, les méthodes mises en œuvre relèvent sans ambiguïté de la chimie. Qu'est donc un chercheur en protéomique ? Un biologiste cherchant à mieux comprendre certaines protéines ? Un chimiste cherchant à améliorer les méthodes d'analyse des protéines ? Initialement, il était un peu des deux. Cependant, avec l'augmentation de la complexité des méthodes d'analyse, avec l'apparition puis la multiplication de méthodes à haut-débit devenant complémentaires les unes des autres, et avec l'industrialisation de la fabrication des instruments, les rôles se sont clairement séparés : des chimistes et des physiciens travaillent sur l'instrumentation et la méthode, alors que les biologistes se détachent de ceux-ci au profit de la question biologique pure. Dans ce contexte, un laboratoire de protéomique peut être perçu comme assis entre deux chaises qui s'éloignent inexorablement.

Afin de pallier cela, les personnels travaillant dans les laboratoires de protéomique se sont peu à peu repositionnés. Certains se concentrent sur les outils, alors que d'autres se rapprochent de la biologie. Enfin, une part significative d'entre eux sert de liant, et s'est spécialisée dans la maîtrise technologique : c'est pour cette raison que la plupart des grands laboratoires de protéomique, partout dans le monde, ont vu leur activité se partager entre recherche (en biologie, en chimie, ou plus rarement, en *data science*) et service, fournissant le support et la maîtrise de leur plateforme instrumentale : ainsi, un laboratoire de protéomique est souvent amené à jouer le rôle de partenaire au sein d'un projet d'envergure, multi-technologique et dont il ne maîtrise pas l'objectif final.

C'est cette activité de plateforme qui a clairement le plus besoin des avancées que peut permettre la *data science*. Cependant, l'intégration des résultats interdisciplinaires qui en résultent ne se fait pas de la même manière que dans le cadre d'une activité de recherche traditionnelle. Cela sera discuté d'ici à la fin du chapitre ; mais auparavant, approfondissons un peu les conséquences de l'évolution disciplinaire susmentionnée.

4.2 L'évolution des métiers

En réalité, cette possible transformation d'une activité de recherche vers une activité de support n'est qu'une des facettes de l'évolution du métier de protéomicien. Il en existe une autre, qui est plus générale, au sens où elle impacte l'activité de nombreux laboratoires interdisciplinaires. En effet, ceux-ci fédèrent souvent des chercheurs issues de différentes disciplines, mais travaillant sur un sujet commun. C'est le cas de mon laboratoire actuel, mais aussi, par exemple de l'Institut de la Communication Parlée (ICP), au sein duquel j'ai effectué mon stage de DEA. Dans les deux cas, j'ai observé le même phénomène quant aux difficultés induites par l'évolution du sujet d'étude : Ainsi, à l'ICP, la compétence en électronique qui a longtemps été au cœur du traitement de la parole est devenue soudainement obsolète avec la démocratisation des outils informatiques. De même, au cours des dernières décennies, le passage de méthodes biochimiques à des méthodes physiques comme la spectrométrie de masse a considérablement changé le métier de protéomicien. À l'heure actuelle la protéomique vit une seconde révolution liée à la place toujours croissante qu'y prend la science des données. Cette évolution rapide des besoins-métiers autour d'une question de recherche interdisciplinaire soulève des questions quant à la gestion des compétences des personnels du laboratoire. Concrètement, un chercheur compétent en biochimie ou en physique des spectromètres a la plupart du temps, choisi cette discipline par goût, voire par passion. Quand la protéomique évolue vers la science des données, que doit-il faire ? Quitter la protéomique pour une autre discipline ? Ou se résigner à l'idée que ses compétences, bien qu'encore nécessaires, ne sont plus aussi indispensables qu'à une époque antérieure ? Ou encore, doit-il évoluer avec la protéomique et changer de métier ? Actuellement, ma position au carrefour entre protéomique et science des données est confortable, cependant, une troisième révolution de la protéomique peut changer cela à l'avenir.

Parmi les effectifs du laboratoire EDyP sur les 10 dernières années (une période courte au regarde de la longueur de la carrière d'un chercheur), la proportion de chercheurs "sur paillasse" (biologistes, biochimistes, analyticiens) ou "sur ordinateur" (informaticiens, statisticiens, mathématiciens) a considérablement évolué. Dans un tel contexte, la question de l'identité d'une discipline et d'un laboratoire devient cruciale. Et cela explique aussi les tensions ou craintes que j'ai pu ressentir dans la communauté (durant des congrès, des soumissions d'articles, ou des recherches de financements) à propos de cette évolution disciplinaire.

Ce problème existe potentiellement pour tout laboratoire interdisciplinaire dont l'unité se fait autour d'un sujet d'étude plutôt qu'autour d'une compétence disciplinaire ; et à plus fort titre dans un laboratoire ayant une forte activité de plateforme. Or, cela ne me semble que peu pris en compte (dans le mode de recrutement des chercheurs, dans la gestion de carrière des personnels permanents, ou encore dans le fonctionnement des laboratoires) ; et c'est probablement un frein au développement des recherches pluridisciplinaires. De plus, ce problème ne concerne pas que les personnels permanents. Comme cela est discuté un peu plus loin, cela impacte les post-doctorants qui sont la variable d'ajustement la plus naturelle pour adapter les compétences de l'équipe aux récentes évolutions disciplinaires ; au détriment parfois de leur carrière.

4.3 Des vertus de la pédagogie

Cette évolution du métier de la protéomique et de la composition des laboratoires de recherche a des conséquences sur la communication scientifique.

La biologie et les mathématiques sont deux extrêmes de la communication scientifique : Le biologiste sait que sa présentation doit être percutante, vendeuse, qu'elle doit raconter une histoire, voire même constituer un bon sujet pour de la vulgarisation. Les diapositives sont un support visuel, dont la qualité des illustrations est essentielle, quitte à cacher certains détails "moins sexy". À l'inverse, le mathématicien, s'il projette des diapositives (ne contenant que formules et listes énumérées), c'est qu'il n'est pas "un vrai", car celui-ci aurait déroulé ses théorèmes à la craie et démontré sa "maîtrise de conférences" au même titre que la puissance de son raisonnement qui, au pire, n'aurait perdu que quelques auditeurs.

En résumé, dans un échange scientifique, il est d'usage chez les biologistes que l'effort de compréhension soit porté par les orateurs, alors qu'il est invariablement porté par l'auditoire chez les mathématiciens. Alors qu'en physique, traitement du signal ou intelligence artificielle, l'effort est plus partagé, les statisticiens, du moins en France, tendent à reproduire l'exemple des mathématiques. Dès lors, il n'est pas étonnant que les présentations de statisticiens soient rarement bien accueillies par les biologistes. Ils leurs préférent des "biostatisticiens", plus au fait des efforts à faire sur la dimension visuelle de la présentation, ou de l'usage selon laquelle une présentation doit être introduite par les retombées attendues par le grand public ("ces travaux s'inscrivent dans le cadre de la lutte contre le cancer.")

Cependant, il y a un écueil, induit par cet usage vulgarisateur de l'illustration dans une présentation technique à destination de biologistes. Un protéomicien versé dans le traitement de données peut facilement penser qu'il a compris le concept ou sa théorie sous-jacente, simplement parce qu'il a assimilé l'analogie de l'explication. Il peut alors souhaiter partager cet éclairage, et transmettre l'analogie lui-même auprès de ces collègues ; qui lui préféreront sa présentation à celle, plus austère, d'un statisticien. Cela est aussi dommageable que fréquent. Ainsi, de nombreux tutoriels, plus "bio" que "stat" illustrent la puissance de l'analyse en composante principale (ACP, cf. Chap. 1, Sec. 5) comme le moyen de mieux visualiser le partitionnement d'un jeu de données. Il en résulte une croyance presque superstitieuse, extrêmement répandue dans le milieu de la protéomique, qui stipule qu'il faut absolument projeter les données dans l'espace engendré par les deux directions principales d'une ACP, avant de réaliser un partitionnement, sous peine d'obtenir de "mauvais" *clusters* se chevauchant ; ce qui est bien évidemment une contre-vérité.

Si les *data scientists* ne font pas l'effort de communication qui est attendu d'eux, cela sera nécessairement demandé à d'autres, dont ce n'est pas le métier, ouvrant la voie à la dérive que je viens d'illustrer. Cela implique donc que nous acceptions d'adopter un ton vulgarisateur, de passer du temps à rédiger ou présenter des tutoriels, et à inclure une dimension "formation" dans l'expertise que nous fournissons dans le cadre de notre activité de recherche. En réalité, ce besoin de vrais experts est déjà clairement affiché dans les différents domaines de recherche de la biologie à grande échelle : les conférences incluent volontiers des tutoriels, voire de vrais cours, et de nombreuses écoles d'été ou de printemps existent ; de plus, les journaux publient des articles ayant une vocation pédagogique. De sorte que l'acceptation d'une

telle activité reste avant tout un choix de chaque *data scientist*. Ce travail peut sembler un peu redondant à l'enseignant-chercheur, qui a déjà une charge conséquente de travail à l'université. De même, certains chercheurs puristes peuvent avoir l'impression de faire de la "leçon de chose", au détriment de leur activité de spécialiste et trouver cela "dégradant". Cependant, il me semble assez logique de considérer qu'une telle activité doit être le marqueur de ce que représente l'interdisciplinarité académique. Et c'est notamment pour cette raison que ce document y accorde une telle importance, au détriment partiel des projets de recherche proprement dits (qui après tout sont déjà détaillés dans des publications).

4.4 Recherche de découverte ou méthodologique

Les manières de communiquer propres à chaque communauté scientifique ne se distinguent pas seulement à l'oral. Elles transparaissent aussi dans les publications scientifiques.

En arrivant dans un laboratoire sous la tutelle de l'Institut National des Sciences Biologiques au CNRS, j'ai eu la grande surprise de découvrir que la notion de recherche n'y était pas du tout comprise de la même manière qu'en science des données. Dans mes précédentes affectations, c'était l'innovation méthodologique qui était la plus valorisée : par exemple, la proposition d'un nouvel algorithme, munie d'une preuve de convergence vers la solution optimale désirée, ainsi que de quelques expérimentations, si possibles très variées, afin de comparer la nouvelle méthode à celles de l'état de l'art. En revanche, l'application triviale (sans adaptation de cette nouvelle méthode) à un problème nouveau, permettant une nouvelle découverte en biologie par exemple, n'est que très peu valorisée. En revanche, en biologie, c'est le contraire. Ce qui se publie bien, c'est la découverte. Si celle-ci a justifié la mise en place d'un nouveau protocole, cela permettra éventuellement de le rendre populaire, sans pour autant qu'une publication méthodologique lui soit dédiée ; mais dans tous les cas, l'emphase reste sur la découverte proprement dite. À l'inverse, les innovations méthodologique, peuvent tout à fait être sous-traitées, y compris à des entreprises extérieures. C'est finalement assez naturel de constater que les "découvreurs" et les "méthodologistes" voient midi à leur porte, et cela peut sembler rassurant pour la possibilité de les voir collaborer. Cependant, ce n'est pas si simple et cela peut avoir des effets parfois délétères.

Il se trouve que les journaux ayant le plus haut facteur d'impact dans lesquels des travaux de protéomique peuvent apparaître traitent d'applications cliniques ou de biologie (le trio *Nature*, *Science* et *Cell* en tête), faisant la part-belle aux découvertes plutôt qu'à la méthodologie. Ces journaux ayant une influence standardisante importante, beaucoup de recherches en protéomique adoptent une présentation correspondant à des travaux de découverte, même si leur contenu est intrinsèquement méthodologique. Notamment, ces recherches sont invariablement publiées selon un plan de type *Introduction / Matériel & Méthodes / Résultats / Discussion*, alors qu'un plan de type *Introduction / Etat de l'art / Contributions / Evaluation comparative / Conclusions*, majoritaire dans les publications de *data science*, serait plus adapté.

Dès lors, un glissement pernicieux s'opère : La méthode ayant été développée sur un matériel particulier (un jeu de données auquel est associé un certain nombre de

questions biologiques), l'article présente en fait des découvertes mineures associées à ce matériel, en insistant sur la nécessité d'avoir utilisé une nouvelle méthode révolutionnaire pour cela ; et la méthode se retrouve ainsi validée de manière beaucoup moins solide qu'avec un article réellement méthodologique.

Ces publications à cheval entre méthodologie et découverte se repèrent facilement sur la base de leurs titres, qui se conforment au modèle suivant :

$$\{\text{Le nom d'un(e) méthode/outil}\} \quad \left\{ \begin{array}{c} \text{“allows”} \\ \text{“permits”} \\ \text{“provides”} \\ \text{“enables”} \end{array} \right\} \quad \{\text{Une nouvelle découverte}\}$$

Voici quelques exemples :

- Vaudel, M., Burkhart, J. M., Zahedi, R. P., Oveland, E., Berven, F. S., Sickmann, A., ... & Barsnes, H. (2015). PeptideShaker enables reanalysis of MS-derived proteomics data sets. Nature biotechnology, 33(1), 22-24.
- Cox, J., & Mann, M. (2008). MaxQuant enables high peptide identification rates, individualized ppb-range mass accuracies and proteome-wide protein quantification. Nature biotechnology, 26(12), 1367-1372.
- Deeb, S. J., D'Souza, R. C., Cox, J., Schmidt-Supprian, M., & Mann, M. (2012). Super-SILAC allows classification of diffuse large B-cell lymphoma subtypes by their protein expression profiles. Molecular & Cellular Proteomics, 11(5), 77-89.
- Gevaert, K., Ghesquière, B., Staes, A., Martens, L., Van Damme, J., Thomas, G. R., & Vandekerckhove, J. (2004). Reversible labeling of cysteine-containing peptides allows their specific chromatographic isolation for non-gel proteome studies. Proteomics, 4(4), 897-908..

Parmi ceux-ci, nous trouvons la célèbre suite logicielle Andromeda / MaxQuant / Perseus [74–76] ; qui est probablement la plus utilisée dans le monde de la protéomique. Celle-ci a été développée dans un des laboratoires de protéomique les plus influents ; cependant cet outil n'a jamais fait l'objet d'une publication méthodologique conforme aux standards de *data science*. Tout au plus, différentes parties des modules le constituant ont été publiées de manière succincte dans les “*Material & Methods*” de différentes publications de découvertes.

Cette situation est particulièrement problématique pour un *data scientist*. En effet, quand il doit comparer une de ses nouvelles méthodes à l'état de l'art, il n'a pas les moyens de le faire : la publication de référence étant centrée sur la découverte plutôt que sur la méthode, la description de cette dernière est incomplète, sans que cela ne pose problème puisque de toute manière, il ne s'agit pas du sujet principale, et que les *reviewers* ne s'y sont pas vraiment intéressés. C'est concrètement la difficulté majeure à laquelle nous sommes confrontés pour l'évaluation des travaux décrits dans le Chap. 3.

Au-delà des travaux décrit dans le suite de ce document, notons que cette situation quelque peu kafkaïenne pour l'ensemble de la communauté perdurera tant que le modèle de publication prédominant en protéomique sera celui de la publication orientée découverte, quand bien même il s'agit de rendre compte d'un travail méthodologique.

5 Constitution d'une équipe de recherche

5.1 Besoins en ingénierie

La différence entre recherche méthodologique et recherche de découverte impacte un laboratoire interdisciplinaire de manière concrète dans son mode de fonctionnement. Notamment, comment faire le trait d'union entre les deux ? On peut espérer que cette complémentarité de profils permette à chacun de trouver sa place. Cependant, dans le continuum de compétences à mettre en œuvre pour que tout le monde collabore, chaque chercheur a généralement une vision restreinte de ce qu'est la recherche (à savoir le petit morceau de méthodologie ou le petit morceau de découverte qui l'intéresse), et leur juxtaposition ne suffit pas à tout couvrir. Finalement, beaucoup de travail considéré comme de la "simple ingénierie" est absolument indispensable pour unifier et intégrer l'ensemble. Concrètement, pour rendre le travail d'un doctorant "méthodologiste" ayant produit un algorithme d'apprentissage automatique utilisable en routine par les protéomiciens d'une plateforme qui collaborent à un projet de "découverte", il va falloir beaucoup de travail d'ingénierie de qualité, permettant de régler des questions qui ne sont pas forcément plus faciles que celles adressées par le doctorant.

C'est notamment ce constat qui m'a amené à consacrer une part significative de ce manuscrit (le Chap. 3) à la supervision de travaux d'intégration et d'ingénierie de recherche : Je pense que ce type de travaux est devenu tout simplement indispensable dans un laboratoire interdisciplinaire jumelé à une plateforme de service, afin qu'il puisse assurer la continuité des compétences et le transfert des méthodes pour leur application à la recherche de découverte. Plus généralement, il me semble nécessaire de leur consacrer une place plus grande dans la valorisation des activités d'un laboratoire.

Ce constat m'amène aussi à l'heure actuelle à favoriser le recrutement d'ingénieurs de recherche (ou dans une certaine mesure, de post-doctorants) par rapport aux doctorants dans le groupe KDPD (*Knowledge Discovery from Proteomic Data*) que j'anime : Pour fonctionner, nous avons tout simplement besoin d'un ratio ingénieurs/chercheurs ou ingénieurs/doctorants élevé.

5.2 Le rôle des doctorants

Il y a cependant un autre élément qui renforce en moi cette conviction : Depuis quelques années, les bourses de thèse ne sont plus accordées sur de limpides critères de mérite, mais sur des sujets associés à des projets, quand elles ne sont pas tout simplement financées par un projet (ANR ou ERC). Dans ce contexte, un doctorant n'est plus un étudiant en thèse, qu'un enseignant (qui se trouve aussi être chercheur) accueille dans une structure de recherche afin de le tutorer dans la mise en place et le développement de son projet ; mais il devient un travailleur à part entière, recruté par un chercheur (qui éventuellement enseigne aussi) qui a besoin de ressources humaines supplémentaires pour mener un projet de recherche. Le doctorant se retrouve donc à travailler dans une logique projet, avec des délivrables, des échéances, et un pilotage dépassant le cadre de ses objectifs de formation. Finalement, il est demandé au jeune recruté, moins bien payé que dans le secteur privé, de préparer un

diplôme, de faire preuve d'autonomie afin qu'il démontre ses capacités de futur chercheur, d'apprendre les canons de la recherche académique (publication, évaluation, recherche de financement, etc.) tout en montrant en plus les capacités d'un ingénieur de recherche chevronné capable d'organiser son travail en fonction de contraintes extérieures fortes. Dans ce contexte, il ne me semble pas du tout étonnant que certains bons étudiants, conscients du piège[3], fuient la poursuite d'études doctorales ; et que les étudiants recrutés en thèse, quelles que soient leur qualités, soient souvent source de déception pour leurs encadrants. Ainsi, je pense qu'une fois titulaire de l'HDR, je chercherai en priorité à recruter des ingénieurs, et je ne profiterai pas de mon droit à encadrer plus de thésards pour en recruter davantage : je m'efforcerai d'en recruter peu, et de fournir à chacun un sujet et une position dans l'équipe de recherche qui garantissent le statut d'étudiant, en situation d'apprentissage du métier de la recherche.

5.3 La position des post-doctorants

Il reste désormais à traiter le cas intermédiaire des post-doctorants, qui de mon point de vue est le plus compliqué, alors que curieusement, leur encadrement ne nécessite pas l'HDR. La diversité des profils est en pratique telle, que le recruté peut dans certains cas être un chercheur parfaitement autonome et développant son propre projet. Dans d'autres cas, ce sera un ingénieur de recherche tout aussi autonome, au rôle essentiel dans le laboratoire. Malheureusement, il peut aussi s'agir d'une personne beaucoup moins autonome, recrutée par manque de candidats, et qui se comportera comme un jeune doctorant ou un ingénieur d'étude à encadrer de près. De plus, quel que soit le profil, la personne restera au laboratoire un temps particulièrement court (12 ou 18 mois, 24 mois au mieux), rendant la valorisation de son travail beaucoup plus difficile, et limitant ainsi le droit à l'erreur. Enfin, dans le cas particulier d'un laboratoire interdisciplinaire impliquant de la recherche en biologie, le problème du rôle du post-doctorant "traiteur de données" est manifeste.

En arrivant dans la communauté de la protéomique, j'ai été réellement surpris par la population postdoctorale que l'on rencontrait. Si on la compare à celle rencontrée dans un laboratoire de recherche d'informatique ou de statistique de rang équivalent, il apparaît que les post-doctorants en biologie sont : (1) plus âgés, (2) avec des listes de publications plus longues, (3) plus internationaux. Pour une part importante, cela s'explique par la possibilité de beaucoup de docteurs en science des données de rejoindre le monde de l'entreprise ; alors que ce n'est pas le cas pour les biologistes, ce qui augmente considérablement la concurrence lors d'un recrutement académique. Cependant, cela n'explique pas tout, notamment pour une catégorie bien particulière de post-doctorants que l'on pourrait décrire ainsi : (1) formation initiale en biologie, suivi d'un doctorat au cours duquel il ou elle a donné des cours de biostatistique ;

3. En fait, le piège est encore plus grossier : À l'heure actuelle, le chercheur a plus besoin de devenir encadrant de thèse, que l'étudiant de devenir docteur : Pour le premier, c'est un moyen indispensable pour augmenter les ressources humaines disponibles sur son sujet de recherche, et rester dans la compétition ; alors que pour le second, cela s'apparente de plus en plus à une première opportunité professionnelle tout à fait interchangeable avec une autre. On peut penser que quand une institution a plus besoin de fournir des diplômes pour survivre, que les personnes diplômées n'ont besoin de la reconnaissance que cela fourni, le diplôme perd mécaniquement de sa valeur.

(2) suite à un premier post-doctorat en analyse de données, la personne se spécialise en bioinformatique/biostatistique, et enchaîne les contrats sur ce type de profil, au grès des projets Européens, afin de venir apporter une compétence manquante et de plus en plus recherchée dans les laboratoires. De tels profils ont peu d'espoir d'obtenir un poste de chercheur : leurs encadrements successifs par des chercheurs n'ayant pas de compétences en *data sciences* ne leur permettent pas de compléter leur formation doctorale sur les aspects qu'ils mettent en valeur dans leur CV ; et leur positionnement peu adapté entre recherche méthodologique et recherche de découverte leur rend presque impossible des publications en premier auteur.

5.4 Direction de recherche

Finalement, l'animation d'une équipe de recherche dans un environnement interdisciplinaire mêlant recherche et activité de support ajoute une difficulté bien spécifique : Il faut éviter d'utiliser les jeunes chercheurs et les non-permanents comme autant de variables d'ajustement face aux difficultés intrinsèques de l'interdisciplinarité, face aux contraintes opérationnelles, ainsi que face aux changements des compétences recherchées (leur *turnover* étant beaucoup plus rapides que celui des personnels statutaires). C'est pourquoi, j'ai fait les choix suivants : je cherche à valoriser autant que possible le travail d'ingénierie dans la recherche académique interdisciplinaire, pour la simple raison qu'elle est nécessaire (tout en évitant de confondre celui-ci avec le travail de recherche). Je suis persuadé qu'un tel travail d'ingénierie ne peut en règle générale pas être correctement réalisé par un doctorant, et je pense qu'il serait souhaitable de moins confondre les rôles, contrairement à la tendance actuelle. En séparant bien le travail d'ingénierie de recherche et celui de doctorant, il devrait être possible de revaloriser les deux : rendre sa noblesse au premier, et rendre le mérite du diplôme au second. Quant aux post-doctorants, leur position me semble la plus délicate, tant leur rôle de variable d'ajustement semble généralisée (autant en France qu'à l'international), de sorte qu'ils me semblent devoir requérir la plus grande attention et une gestion au cas par cas. Cependant, je crois qu'en recrutant des post-doctorants avec un CV 100% *data science* plutôt qu'avec des compétences panachées (en raison d'une formation initiale en biologie ou chimie), il est possible de réduire l'inconfort de la situation. En effet, de tels profils peuvent plus facilement valoriser leurs développements méthodologiques indépendamment du projet biologique auxquels ils se rattachent, en publiant dans leur communauté d'origine des travaux qualifiés "d'appliqués". Par ailleurs, ces publications pourront avantageusement être complétées par des co-signatures dans des revues de protéomique, malgré une position dans la liste des auteurs moins valorisante. Cette stratégie a cependant un écueil : il est difficile d'attirer de tels profils, en raison de la faible compétitivité des salaires académiques en comparaison de ce qu'un *data scientist* peut espérer dans le privé ; mais aussi en raison du faible intérêt (voire de de l'ignorance) du candidat à l'égard du domaine d'application.

Deuxième partie

Travaux de recherche

Chapitre 3

Recherche pilotée

La place déterminante de l'ingénierie de recherche au sein d'un laboratoire interdisciplinaire n'est généralement pas prise en compte dans l'évaluation du travail scientifique, principalement centrée sur le nombre et la qualité des publications. C'est à mon avis un des éléments qui rend le travail interdisciplinaire moins attractif. Cependant, le laboratoire EDyP est hébergé sur le site du CEA de Grenoble, qui s'est justement fait une spécialité de permettre la rencontre de la recherche académique et du monde industriel. Ce contexte particulier permet à notre laboratoire de s'appuyer sur une réelle force d'ingénierie de recherche, et cela est un atout indéniable dans le contexte de la protéomique (où le travail de recherche proprement dit va de pair avec celui de plateforme technologique de support à la recherche). Dans cette partie, je décris d'abord mon activité de chef de projet d'ingénierie, puis les activités de recherche qui prennent racine sur ces projets. Au-delà de la présentation proprement dite de mes travaux, un des objectifs de ce chapitre est de montrer que cette activité d'ingénierie est tellement nécessaire à une certaine forme de recherche, que le rôle du chercheur (qu'il soit chargé ou directeur de recherche) devrait être élargi à la supervision de celle-ci.

1 Descriptif des outils développés

1.1 Contexte logiciel

Comme évoqué dans le chapitre précédent, la protéomique a beaucoup évolué en deux décennies. Initialement centrée sur la biologie et la biochimie (afin d'isoler manuellement des protéines d'intérêt), cette discipline s'est ouverte à des méthodes d'analyse physique (la spectrométrie) qui ont permis un changement d'échelle vertigineux. Cette production de données massives a été la principale cause de l'ouverture de la protéomique à l'informatique : il fallait stocker, trier, sécuriser, rendre accessible et traiter toutes les données produites. Ce n'est qu'ensuite que la dimension statistique est apparue, afin de "faire parler" ces entrepôts de données. C'est dans ce contexte que de nombreux laboratoires de protéomique ont positionné une partie de leur activité de recherche méthodologique sur la conception d'outils informatiques.

Historiquement, les premiers logiciels développés ont été ceux permettant d'identifier les spectres de peptides (selon l'une des trois approches décrites au Chap. 2 : *De Novo Sequencing*, *Library Search Engine*, ou *Database Search Engine*). Initiés

à l'aube des années 2000, ces outils sont maintenant matures, et l'offre est relativement stable, composée d'outils commerciaux comme académiques : citons par exemple Mascot dans la première catégorie [10], et Andromeda dans la seconde [74].

Autour de ces outils, d'autres se sont ensuite développés. Ceux que l'on trouve en amont des moteurs d'identification sont généralement fournis par les constructeurs de spectromètres, alors que pour ceux en aval, l'offre est plus diverse. Je ne propose pas ici d'état de l'art exhaustif de ces outils "aval", dans la mesure où le laboratoire à fait le choix de développer les siens avant mon arrivée.

Les premiers outils développés et distribués dans la suite **ePims** [77], permettent de récupérer les spectrogrammes produits par la plateforme instrumentale, de les stocker, de les archiver, de les indexer, et de les relier à la base de données correspondant aux projets biologiques. Il s'agit donc avant tout de gérer à bas niveau les flux de données. Grâce à cette suite logicielle, les spectres de peptides peuvent être transmis à un moteur d'identification ; à savoir l'outil commercial Mascot, pour lequel nous achetons une licence annuelle. La seconde génération, contenant les outils **IRMa** [78] et **hEIDI** [79], permet de récupérer les identifications de peptides, et d'effectuer l'ensemble des traitements permettant d'aboutir à une identification de protéines ou de groupes de protéines (dans le cas où plusieurs protéines peuvent expliquer les peptides observés). La troisième génération correspond aux travaux entrepris dans le cadre de l'infrastructure nationale ProFI, visant à fédérer les activités de certains laboratoires de protéomique français [80]. Le logiciel correspondant, **ProLine**, est encore en cours de développement. À terme, il a vocation à remplacer **IRMa** et **hEIDI** tout en y ajoutant de nombreuses fonctionnalités, telles qu'entre autres, la quantification relative des protéines identifiées. En attendant que cet outil soit disponible et utilisable en routine, le pipeline ePims-Mascot-IRMa-hEIDI est utilisé pour les projets ne nécessitant pas de quantification, voire une quantification approximative (avec la technique du "spectral count" qui utilise le nombre d'identifications de peptides comme estimateur de la quantité [81]) ; alors que pour les projets nécessitant une quantification plus fine, le logiciel MaxQuant [75] est utilisé. Comme celui-ci ne fonctionne pas avec Mascot, le pipeline devient : ePims-Andromeda-MaxQuant.

Quel que soit le pipeline bioinformatique utilisé lors d'un projet de quantification relative, il fournit en sortie de celui-ci un tableau de données similaire à celui de la Fig. 3.1 : chaque colonne (numérotée de 1 à N) représente un échantillon analysé, et chaque ligne (numérotée de 1 à P) représente une entité chimique qui a été identifiée dans l'un des échantillons (il peut s'agir d'un peptide, d'une protéine dont la présence est déduite des peptides identifiés, ou même d'un groupe de protéines, si les peptides identifiés sont partagés entre plusieurs protéines). Dès lors, le tableau résume l'ensemble des valeurs d'abondance pour chaque entité chimique et pour chaque échantillon (à des fins de quantification relative, cf. Chap. 2 Sec. 2.5). Par ailleurs, ce tableau peut être complété par des métadonnées, permettant de mieux caractériser les entités chimiques ou les échantillons.

1.2 Besoins d'outils pour l'analyse statistique

Une fois en possession d'un tel tableau, il faut mener ce que l'on appelle l'*analyse quantitative*. Son objectif ultime est d'extraire dudit tableau une liste réduite de

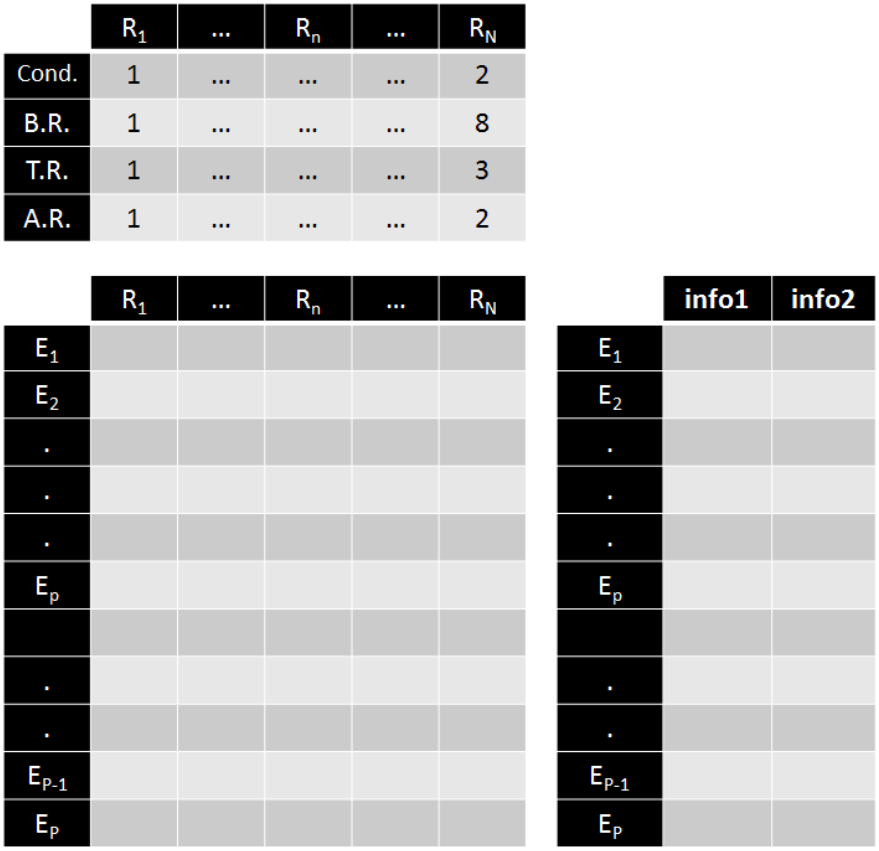

FIGURE 3.1 – *Les données de quantification relative peuvent être structurées en un ensemble de trois tableaux : le tableau central contient les valeurs d'abondance de chaque entité chimique* ***E*** *pour chaque réplicat (ou échantillon)* ***R****. Le tableau du haut contient des métadonnées associées aux échantillons (par exemple, afin d'indiquer quel échantillon appartient à quelle condition biologique à comparer aux autres, ou les niveaux de réplication des échantillons). Enfin, le tableau de droite fourni les métadonnées associées aux entités chimiques (par exemple, pour un peptide, la ou les protéines d'appartenance).*

protéines que l'on pense significativement différentiellement abondantes entre les conditions comparées, et d'associer à cette liste un certain nombre de garde-fous statistiques garantissant sa qualité. Naturellement, un telle analyse quantitative ne correspond ni aux missions, ni aux compétences d'un protéomicien. Il y a donc deux possibilités : la première consiste à dédier ce travail à un chargé d'études en statistiques ou en biostatistiques. Cette stratégie peut bien fonctionner mais risque de correspondre à un travail très répétitif et à faible valeur ajoutée pour le chargé d'étude, tout en rendant difficile la planification de son travail sur le très grand nombre de projets où il sera impliqué. La seconde possibilité est de faire porter les efforts sur le développement d'un outil qui permettra ensuite aux protéomiciens de conduire l'analyse quantitative par eux-mêmes, pour peu qu'ils aient suivi une formation minimale. Le risque de cette approche serait de finalement dépenser plus de temps en développement d'outils que le temps qu'aurait pris le traitement direct

des données. Finalement, nous avons pris le parti de développer des outils statistiques pour le pipeline le plus utilisé (la protéomique quantitative relative *label-free*), et de fournir une expertise de type "chargé d'études" pour les projets plus atypiques.

Le cahier des charges d'un outil permettant de mener une analyse quantitative exprime plusieurs des besoins utilisateurs : celui d'outils, d'interfaces graphiques et de scénarios d'usage, l'ensemble étant par ailleurs gouverné par des contraintes temporelles fortes :

Une collection d'outils statistiques : par ordre chronologique, c'est la première demande qui apparaît : beaucoup d'outils correspondant à tout ce qui a déjà été utilisé dans la littérature protéomique, afin de pouvoir tout tester, et choisir ce qui semble le mieux. À l'inverse, la nouveauté ou l'originalité méthodologique n'est pas demandée (que ce soit par pragmatisme, ou par crainte). Il est donc avant tout attendu une ré-implémentation de l'existant. Face à cette demande, les packages R sont une bénédiction : la collection demandée est là ; et elle est pléthorique.

Des interfaces graphiques : le second besoin apparaît au moment de prendre les algorithmes en main. Il n'est ni dans les compétences, ni dans les intérêts des protéomiciens de coder correctement un script R. Il faut donc développer des interfaces simplifiées leur permettant de travailler (des fonctions simplifiées, un interfaçage avec un tableur, ou des interfaces graphiques). En fonction du choix du type d'interface, le travail d'ingénierie demandé sera plus ou moins lourd, et permettra de plus ou moins autonomiser les utilisateurs.

Des scénarios d'utilisation : très rapidement, l'utilisateur se rend compte qu'il peut-être potentiellement noyé sous le choix des algorithmes à appliquer ; et qu'appliquer une succession d'essais/erreurs sur les données issues d'un projet de recherche (et non d'un *benchmark*) jusqu'à trouver la protéine espérée est dangereux. Dès lors, le chercheur en *data science* peut entamer un travail méthodologique réellement intéressant, visant à : (*i*) avoir un regard critique sur chaque algorithme afin d'en déterminer les cas d'usages adaptés ou la paramétrisation optimale ; (*ii*) la définition de pipelines d'analyse robustes et génériques à un grand nombre de situations (quels algorithmes ? appliqués dans quel ordre ?), ainsi que la justification des choix que cela induit ; (*iii*) le rappel de certaines bonnes pratiques statistiques qui peuvent être oubliées avec le temps. Concrètement, tout ce travail peut être valorisé par des publications de types "*review*", "*viewpoint*", "*tutorial*", "*technical brief*", voire "*guidelines*" et constitue, au-delà du travail de recherche, un élément que je pense être déterminant pour le bon fonctionnement d'une communauté interdisciplinaire. C'est notamment ainsi que j'ai pu faire valoriser les travaux du groupe KDPD sur différents sujets (qui seront détaillés dans la section suivante) :

- Réflexions sur l'imputation des valeurs manquantes [J'1] (*review*)
- Analyse critique d'un test statistique régulièrement mal utilisé en protéomique [J'2] (*viewpoint*)
- Réflexions sur le contrôle des fausses découvertes [J'3] (*technical brief*)

Par ailleurs, au-delà de leurs valorisations, ces réflexions ont permis d'améliorer la qualité des pipelines de traitement des données utilisés au laboratoire, et donc de participer à l'amélioration de la qualité des publications "de découvertes" associées.

Un processus de développement rapide : comme cela a déjà été évoqué (cf. p.40), les protéomiciens abordent simultanément un grand nombre de problèmes et de difficultés, de sorte qu'il est légitime pour eux de ne les prendre en compte que

quand ils se présentent effectivement. Ainsi, les besoins statistiques ne sont exprimés qu'une fois les données produites, ne laissant que peu de temps à la réflexion méthodologique. Cette difficulté est connue de tous les porteurs de projets interdisciplinaires, où finalement, les différents *work packages* ne peuvent être que faiblement intégrés, afin d'éviter les interblocages. Cela est d'autant plus vrai quand le versant applicatif est porté par une plateforme fournissant un service à la communauté, plutôt que par une équipe de recherche pouvant se contenter de prototypes. La plupart des choix technologiques (décrits dans le paragraphe suivant) ont donc été réalisés afin de permettre une réactivité maximale.

1.3 Réalisations

Officiellement lancés en Mars 2013, ces travaux, principalement réalisés par Samuel Wieczorek et Florence Combes, ont abouti 2 ans plus tard à un outil logiciel, toujours en cours d'amélioration et de valorisation [J'4]. Afin de tenir compte des contraintes de développement évoquées ci-dessus, cet outil est décomposé en plusieurs modules différents :

DAPAR (*Differential Analysis of Protein Abundances with R*) : il s'agit un package *wrapper*, c'est-à-dire dont le principal intérêt est de rassembler toutes les fonctions statistiques nécessaires sous une même interface de programmation. Les fonctions en question viennent soit de packages R préexistants, soit d'autres petits packages développés indépendamment au sein de KDPD, dans le cadre d'un travail connexe (cf. Sec. 2). Les fonctionnalités utilisables par DAPAR couvrent les sept étapes que nous avons définies comme essentielles pour une analyse quantitative de données *label-free* :

1. **Statistiques descriptives :** il s'agit de fournir des outils permettant une meilleure compréhension des données, qu'elles soient brutes, ou au contraire, qu'elles aient déjà subi quelques-uns des traitements correspondant aux autres étapes du pipeline.
2. **Filtrage des données** : il s'agit simplement de supprimer des lignes du tableau correspondant soit à des espèces contaminantes (par exemple la kératine présente dans les ongles ou les cheveux du biochimiste ayant préparé l'échantillon, des polluants en suspension dans l'air, etc.), soit à des identifications de type *decoy* (des spectres qui ont été identifiés comme des séquences d'acides aminés inexistantes d'après les bases de données de l'espèce), soit tout simplement des lignes contenant tellement de valeurs manquantes, qu'elles en deviennent inexploitables.
3. **Normalisation** : il s'agit de rééchelonner les valeurs des différentes colonnes du tableau, afin de tenir compte du fait qu'elles correspondent à des réplicats différents, n'ayant pas été analysés dans les mêmes conditions.
4. **Imputation des valeurs manquantes** : Comme nous le verrons plus en détails dans la Sec. 2.2, l'analyse LC-MS/MS produit de manière inhérente des jeux de données contenant un nombre de valeurs manquantes tellement important qu'il n'est pas possible de les ignorer. Alors qu'il est possible d'utiliser des modèles statistiques qui prennent en compte l'apparition de ces valeurs manquantes (sans pour autant les imputer), il est en pratique observé que de

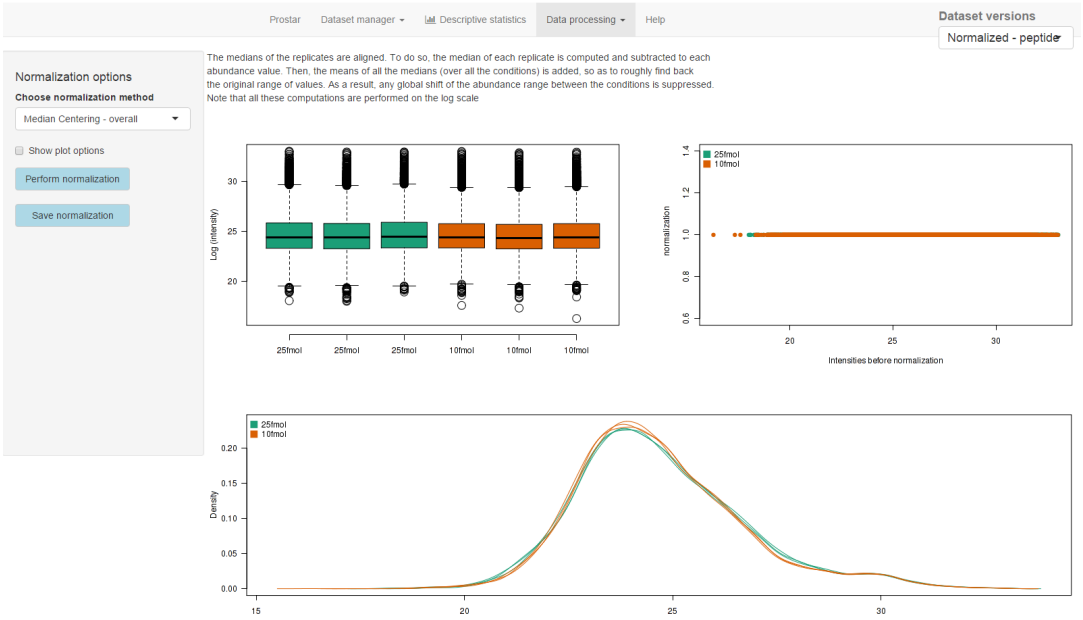

FIGURE 3.2 – *Capture d'écran de ProStaR, durant l'étape de normalisation.*

telles approches sont moins puissantes, et qu'elles amènent les protéomiciens à devoir déconsidérer des protéines pourtant essentielles du point de vue des questions biologiques soulevées (notamment dans le cas d'un faible nombre de réplicats par condition). C'est pourquoi, il est bien souvent préférable d'imputer les valeurs manquantes : si cela est fait avec suffisamment de prudence et de connaissance *a priori*, il est possible de maintenir la puissance de l'analyse statistique en minimisant le biais introduit.

5. **Agrégation** : si les mesures d'abondance concernent des peptides, il est nécessaire de les agréger en tenant compte du graphe de relation peptides-protéines, puisque le niveau protéique est le seul qui soit biologiquement informatif. En revanche, si les données de quantification sont déjà fournies au niveau protéiques (ce qui peut arriver, suivant les logiciels identifications et de quantification utilisés, même si, comme nous l'avons montré dans [J'1], cela n'est pas recommandable), cette étape n'a pas de raison d'être.
6. **Test d'hypothèse** : durant cette étape, chaque protéine est testée individuellement, afin de déterminer si elle est différentiellement abondante entre les conditions biologiques comparées, avec un seuil de significativité donné.
7. **Correction pour tests multiples** : enfin, il s'agit d'estimer la proportion des protéines qui ont été considérées à tort comme différentiellement abondantes.

ProStaR (*Proteomic STatistical Analysis with R*) est un package R basé sur la technologie Shiny [82], qui permet de générer des interfaces web à des fonctions R ; il s'agit de l'interface principale de DAPAR, grâce à laquelle le protéomicien peut accéder aux fonctionnalités de DAPAR sans avoir à coder. De plus, et c'est le principal intérêt de cette interface, elle permet de guider l'utilisateur tout au long d'un pipeline d'analyse global, dont la pertinence statistique est garantie ; tout en lui fournissant un environnement dédié, permettant d'annuler une étape, de revenir en arrière, de contrôler les résultats intermédiaires, et de garder un historique de son

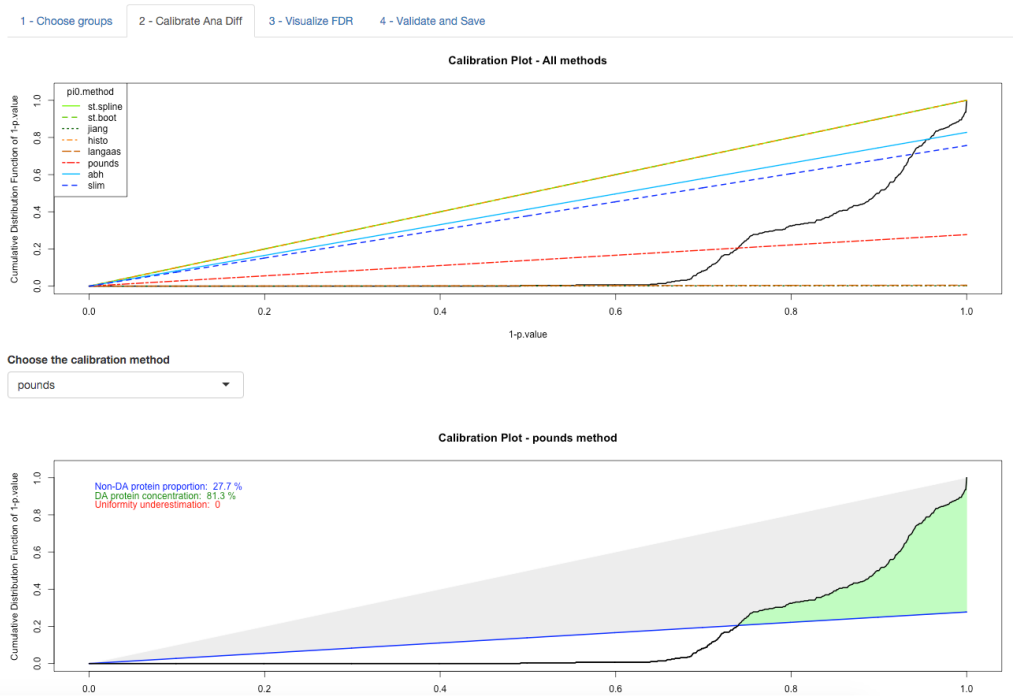

FIGURE 3.3 – *Capture d'écran de ProStaR représentant la fenêtre principale de visualisation durant l'étape de vérification de la bonne calibration des p-valeurs.*

travail (cf. Fig. 3.2 et 3.3, mais aussi la démonstration en ligne disponible à l'adresse www.http://prostar-proteomics.org/).

ProLine, ProteoRE, Bioconductor and Co. : au-delà de ProStaR, un certain nombre d'autres interfaces sont possibles pour les fonctionnalités de DAPAR, comme le logiciel ProLine (cf. Sec. 1.1), l'environnement GALAXIE [83], et bien sûr, toutes les interfaces disponibles dans l'environnement Bioconductor [84], dans lequel DAPAR et ProStaR s'inscrivent.

DAPARdata : il s'agit d'un package contenant simplement des jeux de données. Ceux-ci peuvent non seulement servir à illustrer l'utilisation de ProStaR, mais aussi aux développements méthodologiques, en tant que *benchmarks*, dans la mesure où ils sont étiquetés par une vérité-terrain (quelles sont les protéines différentiellement abondantes, et les autres). L'obtention de tels jeux de données est d'une réelle complexité en protéomique, et en posséder un certain nombre correctement étiquetés et structurés, permet de faciliter grandement la valorisation de nouveaux développements, tels que ceux décrits dans la partie suivante.

2 Une base pour des questions de recherche

Le but de cette section est d'illustrer le fait que posséder une plateforme logicielle suffisamment opérationnelle pour être utiliser en production (c'est-à-dire pour une activité de service en analyse protéomique) permet à la fois d'accélérer la valorisation de travaux méthodologiques, mais aussi d'amener de nouvelles questions de recherche. Ci-dessous sont présentés les principaux travaux (menés ou en cours), qui s'appuient avantageusement sur DAPAR/ProStaR.

2.1 Visualisation des relations peptides-protéines

À partir de la liste des peptides, soit identifiés, soit quantifiés dans un échantillon biologique, il faut remonter à l'identité des protéines constituées de ces peptides. Cette tâche n'est pas très compliquée algorithmiquement ; cependant elle correspond à un problème mal posé, pour plusieurs raisons : tout d'abord, plusieurs protéines peuvent n'être différenciées que par des modifications mineures de leur chaîne d'acides aminés correspondant à des peptides non visibles au spectromètre. Ainsi, la notion même de protéine ne peut pas être prise au pied de la lettre ; et à partir d'un ensemble de peptides, on ne peut identifier qu'un *protein group*, c'est à dire un ensemble d'isoformes très proches. Ensuite, il y a le fait que différents *protein groups* peuvent partager plus ou moins de peptides, en fonction de la granularité à laquelle sont définis les groupes ; par ailleurs, il n'est pas possible, pour un tel peptide partagé, de déterminer sans équivoques dans quelles proportions il vient de l'un ou de l'autre des *protein groups*. Finalement, au moment de l'inférence de l'identité des protéines à partir de l'identité et de la quantité des peptides, un certain nombre de choix arbitraires sont faits, avec le risque qu'autant de biais soient ajoutés. Par ailleurs, ces biais vont se propager tout au long de l'analyse quantitative.

Pour minimiser ces biais, nous proposons de faire intervenir le protéomicien. En effet, deux caractéristiques du problème nous permettent de nous appuyer efficacement sur l'expertise du propriétaire des données : tout d'abord, plusieurs échantillons similaires sont généralement analysés simultanément. Ceux-ci sont censés avoir la même liste de protéines, donc de peptides, et donc le même graphe de relation peptides-protéines (aux erreurs aléatoirement introduites près). Ensuite, le graphe est généralement composé d'un très grand nombre de petites *composantes connexes*[1] : la majorité d'entre elles ne contiennent qu'une seule protéine, et les composantes en contenant plus d'une dizaine ne sont que quelques-unes. Dès lors, en attendant de pouvoir mettre en place des outils automatiques d'alignement de graphes (comme par exemples [85, 86]), il est facile pour l'utilisateur de venir comparer les plus grandes composantes connexes entre chaque échantillon, et éventuellement corriger celles-ci à la marge pour améliorer leur correspondance.

Concrètement, nous avons donc mis en place un outil de visualisation des composantes connexes en collaboration avec Renaud Blanch, du Laboratoire d'Informatique de Grenoble, dans le cadre du projet Prospectom [87]. Contrairement à ce qu'on pourrait attendre, ses travaux antérieurs avaient montré que le graphe lui-même n'est pas la manière la plus adaptée pour représenter un graphe biparti constitué de plusieurs composantes connexes, et qu'il valait mieux se baser sur la matrice d'adjacence de celui-ci. En effet, la modification (ajout, suppression, ou changement de label) d'une arête ou d'un sommet a des conséquences fortes sur cette représentation matricielle. Nous avons donc co-encadré un stage de M2 (celui de Shivani Shah) sur le sujet. L'outil résultant (cf. Fig. 3.4) permet de naviguer

1. Une composante connexe est un sous-graphes dont aucune arête n'entre ou ne sort. Un graphe contenant plusieurs composantes connexes est donc en fait un ensemble d'autant de graphes indépendants, tel que représenté sur la Fig. 3.4. Dans le cas d'un graphe peptides-protéines, aucun des peptides d'une composante connexe n'appartient à une protéine qui n'est pas dans la composante connexe. Et inversement, deux protéines sont dans des composantes connexes différentes si elles ne partagent aucun peptides, que les protéines avec lesquelles elles partagent chacune des peptides n'ont elles-mêmes aucun peptides en commun, et ainsi de suite pour ses protéines et leurs peptides.

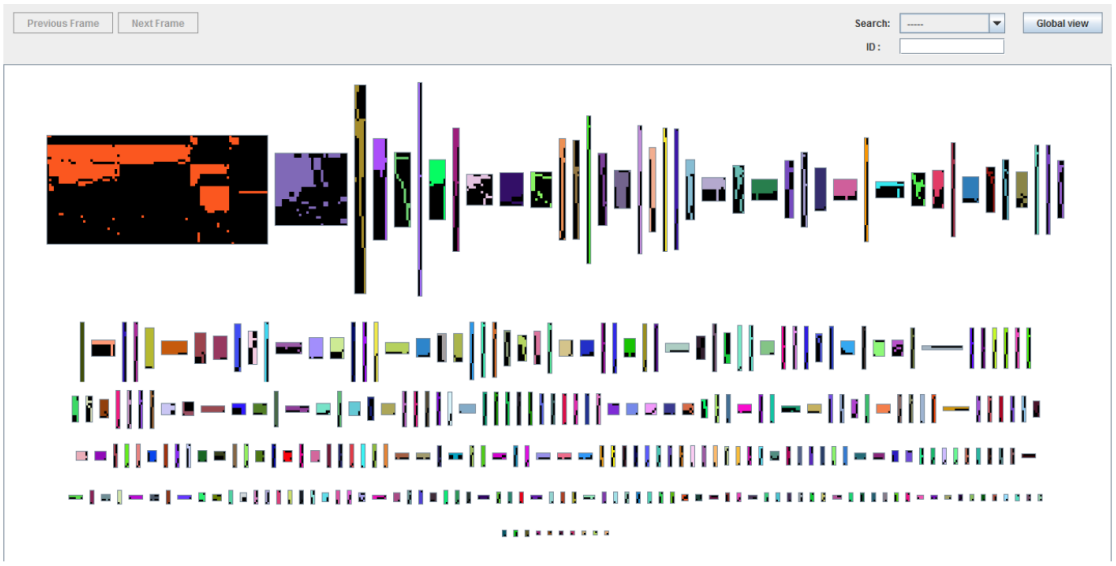

FIGURE 3.4 – *Interface de visualisation des relations peptides-protéines : chaque tableau coloré représente la matrice d'adjacence d'une composante connexe.*

facilement entre les différentes composantes connexes, et d'en aborder facilement la structure. À l'avenir, nous étendrons cet outil, afin qu'il permette la comparaison immédiate et intuitive de plusieurs composantes (en suivant les préconisations de [88]), ainsi que les routines de modifications interactives des graphes en question.

Bien que ce travail ne corresponde qu'à une simple implémentation de résultats connus de *visual analytics*, il me semble particulièrement prometteur. Tout d'abord, parce que la *visual analytics* est une discipline en plein essor ; mais aussi parce que les protéomiciens n'ont culturellement pas confiance en des algorithmes trop automatisés, et souhaitent pouvoir voir et vérifier régulièrement leurs données. Par ailleurs, les données biologiques sont mesurées à partir du monde réel, plutôt que produites, ce qui induit une forte complexification de leur taxonomie, et donc rend leur manipulation à haut niveau sémantique plus difficile. Ainsi, ces données constituent autant de défis intéressants pour la *visual analytics* (l'existence d'une conférence comme BioVis [89] tend à illustrer cela).

2.2 Imputation de valeurs manquantes

Comme indiqué plus haut, les valeurs manquantes sont très nombreuses sur les jeux de données quantitatifs *label-free*. Les raisons de leur présence sont innombrables, cependant, il est possible de les regrouper en deux catégories distinctes, chacune relative à un type de valeurs manquantes ayant un comportement statistique spécifique.

La première catégorie regroupe l'ensemble des processus de censure des valeurs d'intensité qui affectent plus ou moins uniformément l'ensemble des protéines ou des peptides : parmi ces valeurs manquantes, beaucoup sont dues au fait que le dispositif LC-MS/MS ne constitue pas un instrument de mesure exhaustif ; il laisse donc passer certaines entités chimiques "sans les voir" (c'est notamment pour pallier cela que le mode DIA a été imaginé - cf. Chap. 2, Sec. 3.5). Le reste de ces valeurs

manquantes provient d'erreurs de traitement réalisées durant le processus bioinformatique permettant de passer des sorties brutes des spectromètres aux tableaux de données sur lequel se joue l'analyse quantitative.

La seconde catégorie regroupe au contraire les processus de censure qui impactent principalement les entités chimiques d'abondance faible : en effet, comme tout instrument de mesure, le pipeline LC-MS/MS possède une limite de détection, et les peptides dont la concentration est inférieure à cette limite ne sont pas bien mesurés. Par ailleurs, il existe un phénomène supplémentaire qui limite la détection des espèces faiblement abondantes : à l'heure actuelle, les spectromètres de masse ont une gamme dynamique de 4 (voire 5 dans certains cas) ordres de grandeurs. Ainsi, quelle que soit l'abondance d'une espèce chimique, celle-ci sera indétectable si durant le même cycle, une autre espèce chimique 10 000 fois plus abondante est aussi mesurée. On peut faire l'analogie avec le système visuel humain : si l'on regarde face au soleil, la pupille se rétracte tellement que la rétine n'est plus suffisamment éclairée pour permettre d'observer des objets à l'ombre, dans le contre-jour.

Il existe une littérature très abondante sur les différents mécanismes permettant d'aboutir à des valeurs manquantes en protéomique [90–94]. Il existe une littérature tout aussi abondante en statistique, visant à décrire les lois gouvernant les différents types de valeurs manquantes [95–99]. Ce qu'il est important de retenir ici est que la première catégorie de valeurs manquantes peut être qualifiée de MCAR (*Missing Completely At Random*), et que la seconde peut être qualifiée de MNAR (*Missing Not At Random*), et plus précisément de *left-censored data*, puisque la censure impacte le côté gauche (valeurs faibles) de la distribution des données. Alors que les MCAR sont particulièrement étudiées en statistique, en raison du fait qu'on les retrouve dans de nombreux domaines d'étude, il n'y a pas grand-chose sur les MNAR, dont le traitement est toujours très "métier-dépendant".

Tout ce qui est décrit ci-dessus est connu depuis maintenant longtemps dans la communauté de la protéomique. Cependant, il est étonnant de constater que même dans les études les plus récentes (par exemple [100]) où tout ceci est répété en préambule, il n'a jamais été envisagé de traiter différentiellement les MCAR et les MNAR. Ainsi, les travaux de statistiques appliqués à la protéomique se bornent-ils soit à comparer toutes les méthodes existantes d'imputation afin de trouver expérimentalement celles qui fonctionnent le mieux (généralement, une imputation de type MCAR, puisque les algorithmes les plus élaborés s'intéressent à ce type de valeurs manquantes), soit à proposer un modèle prenant en compte les deux types de valeurs manquantes pour expliquer les données observées, sans réaliser d'imputation [101–105].

Dans ce contexte, avec Cosmin Lazar, un post-doctorant du groupe KDPD, et en collaboration avec Laurent Gatto, de l'université de Cambridge, nous avons été les premiers à faire une étude d'une grande simplicité, dans laquelle des jeux de données simulés contenant une part variable de MCAR et MNAR étaient imputés par différents algorithmes [J'1]. Il est naturellement apparu qu'il n'y a pas d'algorithme universellement meilleur pour l'imputation, et que la qualité de l'imputation dépend principalement de l'adéquation entre la nature des valeurs manquantes du jeu de données, et celle(s) que l'algorithme peut traiter efficacement. Ce résultat plaide pour un diagnostic de la nature des valeurs manquantes, et leur traitement différencié, en fonction de leur nature. Par ailleurs, nous avons montré qu'il était

plus pertinent d'imputer les valeurs manquantes au niveau peptidique, et d'agréger ensuite afin d'obtenir les abondances de protéines, plutôt que d'agréger d'abord des valeurs manquantes et observées, pour ensuite imputer au niveau protéique. Tout cela a orienté la suite des travaux de l'équipe, d'abord avec Cosmin, puis avec un autre post-doctorant, Quentin Giai Gianetto.

Commençons par formaliser le mélange des MCAR et MNAR dans un jeu de données : notons x_{ij} la mesure d'intensité du peptide i, $i \in [1, n]$, dans le réplicat j, $j \in [1, J]$. De plus, notons x_{ij}^{obs} si x_{ij} est observé, x_{ij}^{na} si x_{ij} est manquante, x_{ij}^{mcar} si x_{ij} est MCAR et x_{ij}^{mnar} si x_{ij} est MNAR.

Pour chaque réplicat $j \in [1, J]$, soit F_j la fonction de répartition des valeurs d'intensité complète et π_{na_j} la proportion de valeurs manquantes parmi les n intensités peptidiques. Dès lors :

$$F_j(x) = \pi_{na_j} F_j^{na}(x) + (1 - \pi_{na_j}) F_j^{obs}(x) \tag{3.1}$$

où F_j^{na} correspond à la fonction de répartition des intensités des valeurs manquantes x_{ij}^{na}, et F_j^{obs} à celles des valeurs observées x_{ij}^{obs}.

De plus, dans ce réplicat j, des MCAR et des MNAR coexistent en des proportions inconnues, notées π_{mcar_j} et $1 - \pi_{mcar_j}$. Comme les MCAR sont uniformément réparties dans l'intervale des valeurs d'intensité, leur distribution est la même que celle des valeurs d'intensité complètes, de sorte que F_j^{na} peut se réécrire :

$$F_j^{na}(x) = \pi_{mcar_j} F_j(x) + (1 - \pi_{mcar_j}) F_j^{mnar}(x) \tag{3.2}$$

où F_j^{mnar} est la fonction de répartition des MNAR. Par conséquent, l'Eq. 3.2 peut être reformulée de la manière suivante :

$$F_j^{na}(x) = \frac{\pi_{mcar_j}(1 - \pi_{na_j})}{1 - \pi_{na_j}\pi_{mcar_j}} F_j^{obs}(x) + \frac{1 - \pi_{mcar_j}}{1 - \pi_{na_j}\pi_{mcar_j}} F_j^{mnar}(x) \tag{3.3}$$

Dans l'Eq. 3.3, F_j^{obs} et π_{na_j} peuvent être directement estimées à partir du tableau de données (notons leur estimateur respectif $\hat{F}_j^{obs}$ et $\hat{\pi}_{na_j}$). Cependant, les estimations de π_{mcar_j}, F_j^{mnar} et F_j^{na} nécessitent quelques hypothèses supplémentaires.

Tout d'abord, autorisons-nous trois hypothèses générales quant à la distribution des données :

Hypothèse 1 (Absence de peptides non-quantifiés) *Pour chaque peptide identifié, il y a au moins une valeur d'intensité mesurée par condition biologique.*

L'Hyp. 1 provient du fait qu'il est impossible de conduire une analyse différentielle entre deux conditions parmi lesquelles l'une n'est jamais quantifiée. Ainsi, nous supposons que le protéomicien a filtré auparavant de tels peptides ; ou à la rigueur, qu'il les traite séparément à l'aide d'un pipeline dédié.

Hypothèse 2 (Independence intra-réplicat) *Les intensités complètes des peptides sont indépendamment distribuées dans chaque réplicat.*

À première vue, l'Hyp. 2 peut sembler inappropriée : une certaine corrélation est naturellement attendue entre les intensités de peptides venant d'une même protéine.

Cependant en pratique, il y a très peu de protéines avec beaucoup de peptides, et beaucoup de protéines avec peu de peptides. De plus, en raison de tous les artefacts de mesure qui sont décrits au Chap. 2, Sec. 3, plusieurs peptides avec une concentration réelle similaire peuvent en pratique conduire à des intensités MS différentes. À vrai dire, c'est d'ailleurs en raison de cela que la quantification *label-free* ne peut être utilisée que pour des comparaisons d'un même peptide entre plusieurs conditions, et non pour comparer les intensités de différents peptides dans un même réplicat. Finalement, pour toutes ces raisons, cette hypothèse d'indépendance n'a pas de conséquence directe.

Hypothèse 3 (Gaussianité des log-intensité) *F_j suit la fonction de répartition gaussienne.*

L'Hyp. 3 est très classique en protéomique. En fait, il est bien connu que les intensités MS sont distribuées selon une loi log-normale dans chaque réplicat, de sorte qu'une transformation logarithmique est classiquement appliquée comme prétraitement (cf. [92], [106]).

Ensuite, pour fournir un estimateur $\hat{\pi}_{mcar_j}$ de π_{mcar_j}, nous allons faire deux hypothèses temporaires sur F_j^{na} et F_j^{mnar}. Cependant, une fois π_{mcar_j} estimée, ces deux hypothèses seront oubliées pour déduire les estimateurs finaux $\hat{F}_j^{na}$ et $\hat{F}_j^{mnar}$ de F_j^{na} et F_j^{mnar}. Ces deux hypothèses temporaires sont les suivantes :

Hypothèse 4 (Distribution de Weibull des MNAR) *Les MNAR suivent une loi de Weibull translatée dont la fonction de répartition est :*

$$\forall x \in [l_j, u_j], \quad F_j^{mnar}(x) = 1 - \exp\left(-\frac{1}{\lambda^d}\left(\frac{x - l_j}{u_j - l_j}\right)^d\right) \tag{3.4}$$

où $d > 0$ est un paramètre de forme, $\lambda > 0$ est un paramètre d'échelle, $l_j = \min(\min_i(\tilde{x}_{ij}^{na}), \min_i(x_{ij}^{obs}))$ sert à approcher la valeur minimum des intensité complètes dans le réplicat j, et où u_j, qui est définit plus précisément dans l'Hyp. 5, est une approximation de la valeur d'intensité maximum.

Finalement, l'Hyp. 4 ne sert qu'à proposer un modèle paramétrique suffisamment flexible pour qu'il soit possible d'estimer la distribution des MNAR au moyen d'une régression sur lesdits paramètres (d et λ). C'est le résultat de cette régression qui permet de fournir l'estimateur final de π_{mcar_j}.

Hypothèse 5 (Fonction de répartition des valeurs manquantes) *$\exists M_j < u_j$ tel que $\forall x \geq M_j$, $F_j^{na}(x) \approx \tilde{F}_j^{na}(x)$, où $\tilde{F}_j^{na}$ est la fonction de répartition empirique de toutes les valeurs manquantes, après les avoir imputées avec un algorithme qui suppose qu'elles sont toutes MCAR (notons $\tilde{x}_{ij}^{na}$ de telles imputations), et où*

$$u_j = \min\left(\max_{i\in[1,n]}(\tilde{x}_{ij}^{na}), \max_{i\in[1,n]}(x_{ij}^{obs})\right) \tag{3.5}$$

Cette dernière hypothèse est moins intuitive. La logique générale de ce travail est de supposer qu'il est possible d'imputer dans un premier temps toutes les valeurs manquantes comme s'il s'agissait de valeurs MCAR (aboutissant aux valeurs notées $\tilde{x}_{ij}^{na}$, $\tilde{x}_{ij}^{mnar}$ ou $\tilde{x}_{ij}^{mcar}$ en fonction des cas), puis d'utiliser la fonction de répartition de ces valeurs imputées (ainsi que l'Hyp. 4) pour trouver un estimateur $\hat{\pi}_{mcar_j}$ sans biais pour π_{mcar_j}. Cependant, cela ne marche qu'à condition que cette fonction de répartition ne soit considérée que dans un intervalle de valeur d'intensité suffisamment élevé pour qu'il ne contienne pas de MNAR, et donc que l'hypothèse "100% MCAR" ne soit pas trop fantaisiste. C'est cette idée que formalise l'Hyp. 5. Concrètement, il est possible de montrer que si q_j^{mnar} désigne la fonction des quantiles des MNAR, alors l'intervalle $[q_j^{mnar}(100\%), u_j[$ est non-vide, et que c'est dans celui-ci que l'hypothèse "100% MCAR" est la plus réaliste ; de sorte que $q_j^{mnar}(100\%)$ est une borne M_j recevable.

Une fois que π_{mcar_j} est estimé, il est possible d'exprimer F_j en fonction des estimateurs de F_j^{obs} et π_{na_j}, ainsi que de F_j^{mnar} (et donc "d'oublier" les Hyp. 4 et 5, faites pour l'estimation de π_{mcar_j}) :

$$F_j(x) = \frac{1 - \hat{\pi}_{na_j}}{1 - \hat{\pi}_{na_j}\hat{\pi}_{mcar_j}} \hat{F}_j^{obs}(x) + \frac{\hat{\pi}_{na_j}(1 - \hat{\pi}_{mcar_j})}{1 - \hat{\pi}_{na_j}\hat{\pi}_{mcar_j}} F_j^{mnar}(x) \tag{3.6}$$

Dès lors, en utilisant l'Hyp. 3 (qui stipule que F_j suit une loi log-normale), il devient possible de fournir un estimateur $\hat{F}_j^{mnar}$ de F_j^{mnar}, ce qui nous amène à une spécification complète du modèle de valeurs manquantes [2].

À partir de $\hat{\pi}_{mcar_j}$ et $\hat{F}_j^{mnar}$, il devient possible d'estimer, par l'usage du théorème de Bayes, la probabilité qu'une valeur manquante spécifique soit MNAR (ou MCAR) conditionnellement à l'intervalle d'intensité dans laquelle elle se trouve :

$$\mathbb{P}(x_{ij}^{na} \text{ is MNAR} | a_i \leq x_{ij}^{na} \leq b_i) = (1 - \hat{\pi}_{mcar_j}) \frac{\hat{F}_j^{mnar}(b_i) - \hat{F}_j^{mnar}(a_i)}{\hat{F}_j^{na}(b_i) - \hat{F}_j^{na}(a_i)} \tag{3.7}$$

où les bornes de l'intervalle $[a_i, b_i]$ peuvent être estimées à partir des données observées (dans tous les réplicats d'une même condition biologique) pour chaque peptide i à partir de toutes les données observées.

Dès lors, il devient possible de classifier les valeur manquantes en fonction de leur probabilité d'être MNAR ou non, et ensuite, d'imputer chaque classe de valeurs manquantes séparément, avec un algorithme adapté ; notamment, l'algorithme SLSA (*Structured Least Squares Algorithm*, [J'5]) pour les MCAR et l'algorithme IGCDA (*Imputation under a Gaussian Complete Data Assumption*, [J'5]) pour les MNAR : SLSA est simplement une amélioration de l'algorithme LSA [107], qui permet de prendre en compte différents plans d'expérience.

De son coté, IGCDA propose d'imputer les MNAR selon une distribution telle que son inclusion dans un modèle de mélange avec la distribution observée donne une gaussienne, conformément à l'Hyp. 3. Il est possible de déterminer les paramètres de

2. Malheureusement, les estimateurs $\hat{\pi}_{mcar_j}$ et $\hat{F}_j^{mnar}$ que nous avons proposés n'admettent pas de forme analytique puisqu'ils sont le résultat de diverses régressions. Il ne m'a pas donc semblé pertinent de détailler plus leur obtention ; cependant, le lecteur intéressé pourra les retrouver en détails dans [O'1, J'5].

Algorithm 2: Imputation multiple des MNAR et MCAR.

Choisir N le nombre d'imputation à combiner.
Calculer $p_{ij} = \mathbb{P}(x_{ij}^{na} \text{ is MNAR} | a_i \leq x_{ij}^{na} \leq b_i),\ \forall x_{ij}^{na}$.
for $\ell \in [1, N]$ **do**
    for $i \in [1, n]$ **do**
        Générer des tirages de Bernouilli de paramètre $p_{hj}\ \forall (x_{hj}^{na})_{h \neq i}$.
        Imputer les $(x_{hj}^{na})_{h \neq i}$ associées à un tirage négatif avec SLSA.
        Imputer les $(x_{hj}^{na})_{h \neq i}$ associées à un tirage positif avec IGCDA.
        Utiliser ces données complétées pour imputer $(x_{ij}^{na})_j$ avec SLSA.
Calculer la moyenne des N imputations $\forall (x_{ij}^{na})_{ij}$.

cette gaussienne par régression des quantiles les plus élevées des valeurs observées (puisque les MNAR sont absentes des intensités les plus fortes). Ainsi, nous pouvons utiliser le modèle suivant :

$$q_j^{obs+mcar}(l) = m + s \times q_{\mathcal{N}(0,1)}((1 - \gamma_j) \times l + \gamma_j) + \epsilon(l) \tag{3.8}$$

où $q_j^{obs+mcar}$ est la fonction des quantiles des valeurs observées et des MCAR, où $l \in [\eta, 1[$ avec $\eta \in]0, 1[$ tel que $F_j^{mnar}(q_j^{obs+mcar}(\eta)) = 1$, où $\epsilon(l)$ est un bruit blanc gaussien, où $\gamma_j = \pi_{na_j}(1 - \pi_{mcar_j})$, et où m et s correspondent à l'espérance et l'écart-type de la distribution gaussienne des valeurs complètes. On peut montrer qu'un choix judicieux pour η est

$$\hat{\eta} = \hat{F}_j^{obs+mcar}(\max\{x \in \mathbb{R} \text{ tel que } \tilde{F}_j^{mnar}(x) < 1\}) \tag{3.9}$$

où $\tilde{F}_j^{mnar}$ est la fonction de répartition des MNAR après imputation par un algorithme dédié au MCAR, tel que SLSA. Une fois $\hat{m}$ et $\hat{s}$ obtenus, il devient possible, par une sorte de "soustraction de distributions" [3], de récupérer la distribution des MNAR ; et donc d'effectuer un tirage sous celle-ci, selon l'esprit d'IGCDA.

Une telle approche, reposant sur l'imputation des valeurs manquantes étiquetées comme MCAR par SLSA et de celles étiquetées MNAR par IGCDA a cependant deux inconvénients : tout d'abord, les corrélations inter-réplicats résultant de plans d'expériences complexes ne sont pas prises en compte par IGDCA. Ensuite, la sensibilité aux erreurs de catégorisation entre les deux types de valeurs manquantes est importante. Pour contrebalancer ces deux inconvénients, nous avons mis en place un scénario d'imputation multiple décrit dans l'Alg. 2, dont l'objectif est de calculer la valeur moyenne de N imputations réalisées avec différentes catégorisations aléatoires des valeurs manquantes. Comme cela apparaît dans [J'5], cette stratégie permet de réduire de manière significative le biais et la variance des valeurs imputées, quelle que soit la proportion de chaque type de valeurs manquantes.

3. Comme en pratique, nous n'avons que les fonctions de répartition observées à disposition, la "soustraction de distribution" est un peu plus compliquée : il est nécessaire d'appliquer un algorithme PAV [108] sur $(\Phi_{\mathcal{N}(\hat{m},\hat{s})}(x) - (1 - \gamma_j)\hat{F}_j^{obs+mcar}(x))/\gamma_j$.

2.3 Test d'hypothèse sur objets structurés

Ce travail résulte d'une collaboration (non financée) avec Laurent Jacob, chargé de recherche CNRS en biostatistiques au Laboratoire de Biométrie et Biologie Évolutive, à Lyon. Il est motivé par l'interrogation suivante : comment utiliser l'information quantitative fournie par un peptide partagé entre une protéine différentiellement abondante et une protéine qui ne l'est pas ? Intuitivement, pour estimer correctement les abondances des deux protéines, la quantité du peptide en question doit être partagée en deux : celle provenant de la protéine dont la quantité est stable, et celle provenant de l'autre protéine. De même, la variation de la quantité du peptide entre les deux conditions doit être décomposée en deux : la variation de l'abondance de la protéine différentiellement abondante, et celle résultant de la variabilité inter-réplicats. Il en résulte que l'agrégation des intensités peptidiques au niveau protéique d'une part, et le test des abondances des protéines d'autre part, ne peuvent être considérés séparément ; du moins, si l'on souhaite tenir compte des peptides partagés.

Ce problème est bien connu de l'état de l'art (voir par exemple [109–111]), et une manière élégante de l'aborder consiste à : (1) proposer un modèle liant l'intensité des peptides observés (partagés comme protéines-spécifiques) aux abondances inconnues de protéines ; (2) effectuer une première régression sur ce modèle en supposant que les abondances protéiques sont les mêmes pour toutes les conditions ; (3) effectuer une seconde régression en supposant cette fois que les abondances protéiques sont différentiellement abondantes. Ensuite, pour chaque protéine, il s'agit de comparer le pouvoir explicatif des deux régressions : si celui de la seconde est supérieur, alors la protéine est considérée comme différentiellement abondante.

En pratique, ce cadre général va être raffiné par différent choix. Notamment le choix du modèle, la manière dont les deux régressions de celui-ci sont calculées, et enfin, la manière dont le pouvoir explicatif des régressions sont comparés. Dans ce contexte, les propositions de Laurent ont été les suivantes :

Choix du modèle : afin de faciliter sa résolution (voir plus loin), seuls des effets fixes sont considérés :

$$\ln(y_k)|X,\theta,\alpha \sim \mathcal{N}\left(\sum_{j=1}^{p} x_{kj}\theta_j + \alpha_k, \sigma^2\right) \tag{3.10}$$

où y_k représente l'intensité du peptide k, X est la matrice d'adjacence du graphe peptides-protéines, de terme générique x_{kj}, α représente le vecteur des effets peptides (la capacité de chaque peptide à être observé durant l'expérience), avec $\alpha_k \sim \mathcal{N}(0, \sigma_k^2)$, et enfin, où θ_j est le logarithme de l'abondance de la protéine j.

Méthode de régression : la solution du maximum de vraisemblance s'obtient par la méthode des moindres carrés :

$$\left(\hat{\theta}, \hat{\alpha}\right) = \arg\min_{\theta,\alpha} \|y - X\theta - W\alpha\|^2, \quad \hat{\sigma}^2 = n^{-1}\|y - X\hat{\theta} - W\hat{\alpha}\|^2, \tag{3.11}$$

où y est un vecteur de taille nq, avec n égale au nombre de réplicats, et q au nombre de peptides, et où $W \in \{0,1\}^{nq\times q}$ relie les nq mesures aux q peptides. Cependant, afin de régulariser la variance empirique (afin de tenir compte du faible nombre de réplicat, comme dans [112]), un prior de loi inverse-gamma est imposé ($\sigma^2 \sim$

$inv\mathcal{G}(-1, \beta)$). Dès lors, l'estimateur MAP (*maximum a posteriori*) de la variance devient :

$$\hat{\sigma}^2 = n^{-1}\|y - X\hat{\theta} - W\hat{\alpha}\|^2 + s_0, \text{ avec } s_0 = 2\beta n^{-1}. \tag{3.12}$$

Comparaison : l'explication des observations par un modèle supposant l'égalité des abondances protéiques dans les deux conditions correspond typiquement à l'hypothèse nulle d'un test statistique, dont le rejet indique qu'un modèle acceptant une différence d'abondance est à privilégier. Par ailleurs, Le lemme de Neyman-Pearson [113] nous indique que le ratio des vraisemblances des deux modèles

$$\Lambda(y) = \frac{L_{H_0}(y)}{L_{H_1}(y)} \tag{3.13}$$

permet de construire un test d'hypothèse puisque d'après le théorème Wilk [114], sous l'hypothèse nulle, $-2\log\Lambda(y)$ converge en loi vers la distribution du χ^2 à 1 degré de liberté, quand $nq \to \infty$. Par ailleurs, ce test de ratio de vraisemblance est démontré être de puissance maximale. Malheureusement, la convergence proposée par Wilk ne tient plus pour l'estimateur MAP de l'Eq. 3.12. Cependant, il est possible de montrer [J'6] que dans le cas d'effets fixes, l'expression de $-2\log\Lambda(y)$ se simplifie en

$$nq\left(\ln(\hat{\sigma}^2_{H_1} + s_0) - \ln(\hat{\sigma}^2_{H_0} + s_0)\right) \tag{3.14}$$

et que pour $s_0 \geq 0$, il suffit de multiplier le ratio de vraisemblances par le facteur correctif suivant $\frac{\sigma^2+s_0}{\sigma^2}$ pour de nouveau converger en loi vers la distribution du χ^2 à 1 degré de liberté :

$$\frac{\sigma^2 + s_0}{\sigma^2} nq\left(\ln(\hat{\sigma}^2_{H_1} + s_0) - \ln(\hat{\sigma}^2_{H_0} + s_0)\right) \to \chi^2_1. \tag{3.15}$$

Ce qui permet d'aboutir à une p-valeur pour chaque protéine testée.

Ce travail est encore en cours [J'6], mais les évaluations préliminaires montrent qu'un tel test est d'autant plus intéressant que le nombre de peptides partagés est élevé, ce qui nous amène à continuer dans cette voie, et à espérer, à terme, généraliser la méthode afin de tenir compte de différents plans d'expérience. Par ailleurs, bien que plus lourd calculatoirement que l'enchaînement d'une étape d'agrégation et d'un test statistique, le temps de calcul reste acceptable, contrairement à d'autres méthodes basées sur le même principe, comme par exemple [111], mais fondées sur un modèle plus compliqué que celui de l'Eq. 3.10.

2.4 Contrôle qualité et fausses découvertes

Considérons une expérience de protéomique quantitative où les abondances de milliers de protéines sont comparées entre plusieurs réplicats qui se divisent en deux conditions biologiques (par exemple mutant *vs. wild-type*). Rappelons les définitions introduites Chap. 1, Sec. 2.3, et précisons-les un peu (cf. Fig. 3.5) :

Une découverte potentielle : toute protéine quantifiée de l'expérience ;

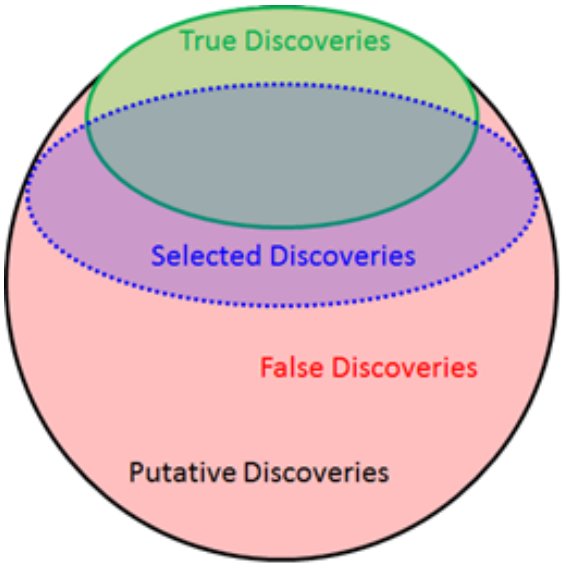

FIGURE 3.5 – *Diagramme de Venn représentant les vraies, fausses découvertes, ainsi que les découvertes potentielles et sélectionnées.*

Une vraie découverte : une protéine qui est différentiellement abondante entre les conditions biologiques, et qui est recherchée pour son intérêt biologique potentiel ;

Une fausse découverte : une protéine qui est non différentiellement abondante entre les conditions biologiques, et qui est donc dépourvue de tout intérêt biologique dans le cadre de cette expérience ;

Une découverte sélectionnée : une protéine qui a passé un certain seuil statistique défini par l'utilisateur, et qui devrait être biologiquement pertinente, même si, dans la pratique, on ne sait pas si elle l'est (ce serait une véritable découverte alors) ou pas (une fausse découverte).

Comme nous l'avons vu aussi Chap. 1, Sec. 2.3, ce seuil statistique est défini par rapport à la p-valeur, cette valeur chiffrée qui est fournie par le statisticien comme une réponse à une question différente de celle que le praticien se pose ; et dont les difficultés d'interprétation ont été illustrés par l'Ex. 2, p. 13, où 3000 découvertes potentielles étaient considérées simultanément. En effet, la subtilité d'interprétation de la p-valeur est plus facile à appréhender dans un tel contexte, et c'est d'ailleurs pour cela qu'il est possible de tirer parti de la coexistence d'un très grand nombre de p-valeurs (ce que l'on appelle les tests multiples, cf. p. 14) pour réaliser une estimation de la quantité suivante, nommée Proportion de Fausses Découvertes (ou FDP pour *False Discovery Proportion*) :

$$\text{FDP} = \frac{\#\{\{\text{Fausse Découvertes}\} \cap \{\text{Découvertes Sélectionnées}\}\}}{\#\{\text{Découvertes Sélectionnées}\}} \tag{3.16}$$

Bien que cela ne réponde vraiment pas à la question initiale (Pour chaque protéine, quelle est la probabilité d'être une fausse découverte ?), la FDP donne néanmoins une indication intéressante, car elle répond à une question connexe : parmi toutes les protéines sélectionnées, combien sont des fausses découvertes ?

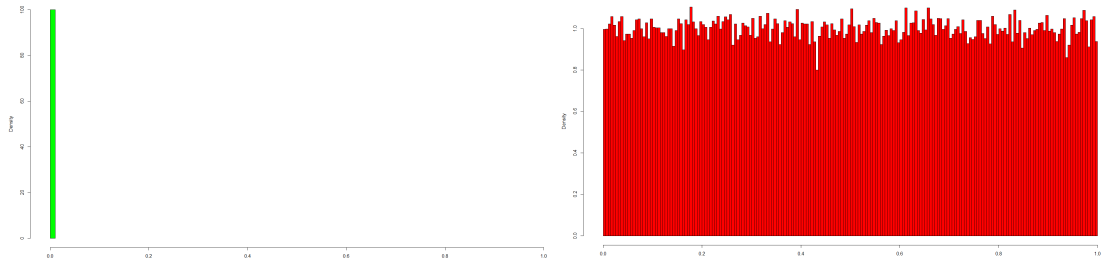

FIGURE 3.6 – *Histogramme des p-valeurs des vraies découvertes (à gauche), et des fausses découvertes (à droite) que l'on doit théoriquement observer à la suite d'un test d'hypothèse.*

Commençons par fournir une intuition sur la façon dont il est possible d'estimer cette FDP. Pour cela nous allons nous appuyer sur le comportement statistique connu des p-valeurs :

- Les p-valeurs de vraies découvertes sont assez simples à caractériser : elles sont petites. Ainsi, au lieu d'être distribuées sur tout l'intervalle $[0, 1]$, elles se concentrent sur une petite région à proximité 0, comme illustré sur l'histogramme de gauche de la Fig. 3.6, qui représente un jeu de données simulé où toutes les protéines seraient clairement différentiellement abondantes.
- À l'inverse, les p-valeurs de fausses découvertes ont un comportement étrange : on pourrait penser que, contrairement à celles des vraies découvertes, elles se distribuent dans le haut de l'intervalle $[0, 1]$, de sorte qu'il serait possible de trouver un seuil de discrimination entre les p-valeurs faibles (vraies découvertes) et p-valeurs élevées (fausses découvertes). Cependant, ce n'est pas le cas. En fait, si le test statistique est bien choisi et fournit des p-valeurs bien calibrées [J'3], les dernières sont censées être réparties uniformément dans $[0, 1]$. Autrement dit, il y a autant de petites p-valeurs que de grandes (ou d'intermédiaires) parmi les fausses découvertes. L'histogramme de droite de la Fig. 3.6 illustre cela avec un jeu de données simulé où 100% des protéines sont non différentiellement abondantes. Bien que contre-intuitif, ce résultat peut être démontré mathématiquement, et reste toujours vrai tant que les fausses découvertes suivent bien la distribution nulle du test (ce qui dans la vie réelle n'est pas toujours garanti).

En conséquence, si un jeu de données de protéomique quantitative contient une proportion π_0 de fausses découvertes et $1 - \pi_0$ de vraies découvertes, l'histogramme devrait ressembler à celui de la Fig. 3.7. Si l'on zoome sur le côté gauche de cet histogramme, deux observations peuvent être faites : tout d'abord, il faut ajuster le seuil de sélection de telle sorte que toutes les vraies découvertes soient sélectionnées, tout en minimisant le nombre de fausses découvertes, comme illustré par la ligne verticale en pointillés sur la Fig. 3.7. Ensuite, il est possible d'avoir une première estimation grossière de la FDP en calculant le ratio qui est représentée par les cases colorées. Pour calculer ce ratio, nous allons introduire quelques notations supplémentaires. Soient :

- α le seuil choisi par l'utilisateur sur la p-valeur (la ligne verticale en pointillés sur la Fig. 3.7) ;
- m le nombre total de découvertes potentielles dans l'ensemble de données ;

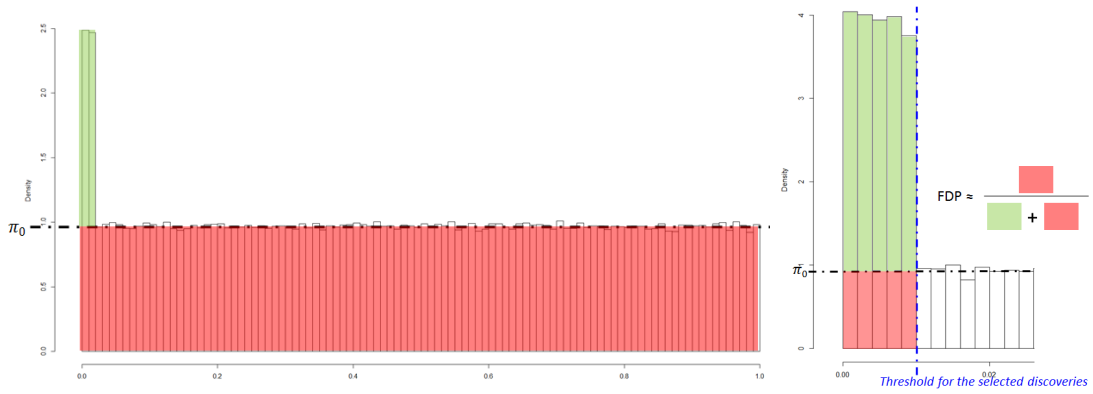

FIGURE 3.7 – *Histogramme représentant la distribution des p-valeurs dans un jeu de données contenant à la fois des vraies découvertes (vert) et des fausses (rouge).*

– k le nombre de découvertes sélectionnées.

À partir de là, il est assez simple de déduire un premier estimateur naïf : le nombre total de découvertes sélectionnées (C'est-à-dire le dénominateur de la formule de la Fig. 3.7) est égal à k. Ensuite, le numérateur est égal à une proportion α du nombre total de fausses découvertes, ce dernier étant égal à $m \times \pi_0$. Finalement, on en déduit :

$$\text{FDP} \approx \frac{\alpha \times m \times \pi_0}{k} \tag{3.17}$$

Bien sûr, cet estimateur de la FDP n'est pas très fin (sans compter le fait que pour l'instant, on n'a aucune idée de la valeur de π_0). Cependant, cela illustre bien qu'il est possible de tirer parti de la grande quantité de découvertes potentielles pour extraire des informations supplémentaires qui peuvent être utiles aux protéomiciens.

Tournons-nous maintenant vers le Taux de Fausses découvertes, (ou FDR pour *False Discovery Rate* ; cf. [17] pour une revue du sujet). Essentiellement, un FDR est un estimateur de la FDP qui est doté de certaines propriétés statistiques importantes. Un FDR doit être :

Conservateur : cela signifie qu'il ne faut pas sous-estimer le véritable FDP, ou tout du moins, cela ne doit arriver qu'avec une très faible probabilité. Ceci est d'une importance capitale pour s'assurer qu'il n'y a pas plus de fausses découvertes qu'annoncées. En d'autres termes, cette conservativité est essentielle pour le contrôle de la qualité des conclusions biologiques.

Asymptotiquement convergent : intuitivement, cela signifie que la moyenne d'un très grand nombre de FDR calculés sur des jeux de données avec les mêmes distributions statistiques doit être égale à leur véritable FDP.

Naturellement, les statisticiens ont beaucoup travaillé sur ces propriétés, et ils ne sont pas nécessairement d'accord sur leur mise en œuvre précise. Par exemple, les définitions mathématiques précises de ces deux propriétés diffèrent légèrement, selon que nous utilisons la définition de Benjamini et Hochberg du FDR (BH, [115–118]), ou celle de Storey et de Tibshirani (ST, [112, 119–122]). Cependant, du point de vue de la protéomique, ces détails techniques peuvent être survolés : en fait, les familles BH et ST se ressemblent plus qu'elles ne diffèrent :

Premièrement, les deux familles sont plus ou moins liées à l'estimateur naïf de la FDP décrit plus haut. Dans ce contexte, les FDR des deux familles nécessitent d'être capable d'estimer π_0.

Deuxièmement, les deux familles ont le même type de faible sensibilité à l'égard de l'estimateur de π_0 : c'est pourquoi, la connaissance précise de π_0 n'est pas obligatoire, même si elle est bien sûr un bonus. Ce qui importe vraiment ici est d'éviter toute sous-estimation. Voilà pourquoi, dans le travail original de Benjamini et Hochberg, on retrouve $\pi_0 = 1$ [115]. Alors qu'une estimation extrêmement précise serait nécessaire pour déterminer de manière fiable la probabilité qu'une découverte potentielle soit fausse (la question du protéomicien à laquelle le statisticien ne répond pas), une estimation moins précise, mais conservative, est suffisante pour avoir un bon FDR (qu'il soit de type BH ou ST).

Troisièmement, les FDR des deux familles reposent sur une procédure algorithmique similaire pour être calculés. Cette procédure, consiste tout d'abord à trier de manière ascendante les p-valeurs,

$$p_{\min} = p_{(1)} \leq p_{(2)} \leq \ldots \leq p_{(m-1)} \leq p_{(m)} = p_{\max} \tag{3.18}$$

puis à définir le FDR de la liste des i protéines correspondant aux p-valeurs $\{p_{(1)}, \ldots p_{(i)}\}$ de la manière suivante :

$$FDR_{(1,\ldots,i)} \sim f\left(\frac{p_{(i)} \times m \times \pi_0}{k}\right) \tag{3.19}$$

$$\sim g\left(p_{(i)}\right) \tag{3.20}$$

Dans la première formulation, il s'agit d'appliquer une fonction f à notre estimateur naïf de la FDP (Eq. 2.4) avec $\alpha = p_{(i)}$; le rôle de f est de rendre cet estimateur conservateur et asymptotiquement convergent. Dans la seconde formulation, le produit par $\frac{m \times \pi_0}{k}$ est inclus dans f pour définir g ; cela permet de faire apparaître que le FDR des i premières protéines peut être calculé par une transformation dépendant de la p-valeur $p_{(i)}$; ce qui a motivé les dénominations *p-valeurs ajustées* (dans la famille BH), ou *q-valeurs* (dans la famille ST).

Toutes ces notions faisant largement partie de l'état de l'art, j'aurais en toute logique dû les introduire dans la section dédiée aux statistiques, Chap. 1. Cependant, celles-ci ont cela de particulier qu'elles ne sont que très rarement maîtrisées dans la communauté protéomique, en dépit de leur rôle essentiel dans son quotidien. Cet écart entre besoin élevé et compréhension faible est flagrant, et incomparablement plus important que pour toute autre notion de *data science*, quelle qu'elle soit. Mon propos ne doit pas être perçu comme suffisant vis-à-vis de la communauté protéomique : l'échec est clairement à mettre au compte de la communauté statistique, qui cumule un choix de vocabulaire trompeur (dans le langage courant, "taux" et "proportion" sont synonymes, de sorte que FDR et FDP devraient l'être), et une difficulté à communiquer déjà discutée au chapitre précédent (Sec. 4.3) ; difficulté bien connue dans d'autres domaines d'application, comme les sciences sociales, où la discussion est même portée par les journaux du domaine [13–16].

C'est pour cela qu'une partie importante de mes publications ou communications traitant du FDR comporte une dimension pédagogique, ou encore, se concentre sur la description d'outils permettant d'introduire dans la communauté de meilleures pratiques quant à son usage. Voici deux exemples concrets :

Exemple 3 (Filtrage ou estimation des fausses découvertes) *La confusion entre FDP et FDR est lourde de conséquences, puisqu'elle gomme la notion d'estimateur. Ainsi, alors que pour un statisticien, "contrôler le FDR" signifie trouver un estimateur de la borne supérieure de la FDP, pour un protéomicien, cela signifie bien souvent "faire la chasse aux fausses découvertes afin d'avoir une FDP plus petite", en mettant en place une série de filtres* ad-hoc*. L'ensemble de ces pratiques a donné lieu à toute une littérature du FDR en protéomique, qui reste indépendante de celle développée en statistique, comme souligné dans [123]). De mon point de vue, cette littérature "indépendante" sur le FDR en protéomique est pertinente, mais là où l'on ne l'attend pas. En effet, filtrer les fausses découvertes, comme le fait le protéomicien, est aussi important que d'estimer combien il en reste (après filtrage), tel que propose de le faire le statisticien. Le filtrage permet d'améliorer la qualité du travail, et l'estimation du FDR de donner une valeur numérique à cette qualité. Les deux sont donc très complémentaires. Malheureusement, ils ne sont pas utilisés comme tels, pour la raison suivante : d'après moi, l'un des travers de la biologie à grande échelle est de vouloir réduire l'analyse et le traitement des données à sa dimension statistique, afin d'automatiser celle-ci et de réduire la part d'erreur relevant de la subjectivité du praticien. Ce travers s'exprime notamment par le fait qu'un protéomicien justifiera plus facilement un traitement qu'il a appliqué à ses données si celui-ci relève d'une "méthode statistique" plutôt que de son expertise. Ainsi, afin que celle-ci puisse s'exprimer malgré tout, elle est souvent déguisée en "méthode statistique". C'est finalement cela qui amène à la publication de méthodes de contrôle du FDR douteuses du point de vue de la méthodologie statistique : des protéomiciens ont habillé leur expertise d'un vernis mathématique, afin croyaient-ils de les rendre plus présentables ou plus rigoureuses (à titre d'exemple, considérons [124], qui est une justification statistique* post-hoc *des choix réalisés dans [125], largement critiqués, notamment dans [123] précédemment cité). Cependant, ce faisant, l'expertise métier rentre* de facto *en compétition avec les méthodes issues de la communauté statistique. De mon point de vue, il serait bon de supprimer cette compétition, en rendant aux protéomiciens la pertinence et le bien-fondé de leur expertise, qui ne pourra jamais être remplacée par des calculs statistiques. Cela permettrait de juger la littérature sur le FDR en protéomique sur la base de ce qu'elle contient vraiment (des outils de filtrage), mais aussi d'amener les protéomiciens à comprendre la pertinence des outils issus des statistiques. C'est en tout cas le message que j'ai essayé de faire passer dans mes* viewpoints, reviews *ou tutoriels précédemment mentionnés.*

Exemple 4 (Détournement du *fudge factor*) *En 2001, Tusher, Tibshirani et Chu ont proposé une nouvelle méthode de contrôle de la FDP adaptée aux données issues de puces à ADN [112]. Pour cela, ils proposent entre autre de modifier le test de Student pour tenir compte de la difficulté à estimer la variabilité de telles données. Ainsi, au lieu de la statistique classique, où pour chaque gène i, l'écart entre les moyennes des deux groupes A et B est pondéré par l'écart-type observé s sur des données, à savoir :*

$$T_i = \frac{\mu_i^A - \mu_i^B}{s_i}; \tag{3.21}$$

ils proposent de considérer la statistique suivante,

$$T_i^* = \frac{\mu_i^A - \mu_i^B}{s_i + s_0}, \tag{3.22}$$

où s_0, appelé fudge factor*, est fixé à une valeur très faible, afin notamment d'éviter, dans le cas d'observations identiques, une variance nulle et une statistique indéfinie. Cette variation du test de Student s'est révélée particulièrement pertinente en transcriptomique, de sorte qu'elle a été adaptée à la protéomique. Historiquement, en protéomique, le test de Student n'était pas utilisé, et seule la différence des moyennes (après transformation logarithmique), à savoir $\mu_i^A - \mu_i^B$, était considérée pour la recherche de protéines différentiellement abondantes. Cependant, sous l'influence des guides de bonnes pratiques statistiques édités par les journaux, des tests statistiques ont été adoptés, afin de prendre en compte la variabilité des observations. Notamment, le test de Tusher, Tibshirani et Chu s'est révélé particulièrement intéressant, puisqu'il suffisait de régler s_0 sur une valeur 10 à 100 fois supérieure à celle préconisée dans [112] pour conduire à des résultats similaires à ceux obtenus avec la différence des moyennes. En effet, si s_0 est grand devant tous les s_i, on a :*

$$T_i^* = \frac{\mu_i^A - \mu_i^B}{s_i + s_0} \approx \frac{\mu_i^A - \mu_i^B}{s_0} \propto \mu_i^A - \mu_i^B \tag{3.23}$$

Ce contournement des guides édités par les journaux est très répandu : en effet, par défaut, Perseus [76], l'outil d'analyse statistique de la suite logicielle de référence Maxquant [75] propose de faire ainsi. Cette dérive est même involontairement plébiscitée par de nombreuses publications de haut niveau [126–131], voire même par la définition d'un protocole spécifique [132]. Afin de modestement lutter contre cette dérive, nous avons publié une explication détaillée [J'2], mais aussi incorporé dans ProStaR des outils alternatifs. C'est notamment pour permettre cela, qu'avec Quentin, nous avons ajouté une étape de vérification visuelle de la proportion π_0 de fausses découvertes [J'3] au cours de la procédure de tests multiples et de contrôle du FDR.

Finalement, en comparaison des travaux présentés juste avant (imputation de valeurs manquantes ou tests d'hypothèses structurés), ce travail sur le FDR apparaît comme reposant sur un apport méthodologique particulièrement faible. Concrètement, il s'agit surtout d'un mélange de revue de l'état de l'art, d'ingénierie et de pédagogie dans lequel l'innovation *stricto sensu* est minoritaire. Cependant, ce type de travail est essentiel pour la communauté, comme pour les chercheurs qui le conduisent : pour la communauté, parce qu'il permet d'améliorer les pratiques et la qualités des analyses protéomiques, ainsi que de renforcer sa culture interdisciplinaire sur des aspects qui lui font défaut ; et pour les chercheurs, parce que cela permet de renforcer leur visibilité à peu de frais, par le seul fait que le type de publications qui résultent d'un tel travail touche une large audience. Enfin, à titre personnel, je pense que ce type de travail permet de décloisonner les expertises et de rompre avec la surspécialisation des chercheurs, tout en renforçant le rôle historique du chercheur académique, dont une partie concerne la diffusion de connaissances et de bonnes pratiques (rigueur, contrôle qualité, etc.).

3 Questions futures

Au-delà des questions de recherche de la section précédente, qui s'appuient clairement sur l'infrastructure fourni par DAPAR et ProStaR, j'espère que de nombreuses autres apparaîtront à l'avenir. Parmi celles-ci, une part viendra de demandes très opérationnelles des utilisateurs protéomiciens, alors que d'autres seront plutôt guidées par les opportunités scientifiques.

3.1 Questions guidées par les besoins

Ces dernières années, j'ai entendu de nombreux protéomiciens se plaindre des relations qu'ils pouvaient avoir avec certains biologistes, qui réduisaient tellement la complexité de leur métier, qu'ils en venaient à nier *de facto* l'expertise que cela nécessitait : il suffirait d'injecter l'échantillon dans le spectromètre, puis d'appuyer sur un bouton pour récupérer en sortie une liste exhaustive de protéines, quantifiées précisément. Quand une vision aussi réductrice est de mise (ce qui n'est heureusement pas systématique), cela entrave sérieusement la collaboration entre biologistes et protéomiciens. En effet, dans de tels cas, ces derniers sont perçus comme de simples sous-traitants techniques devant forcement être capable de fournir un résultat immédiat, exhaustif et parfaitement reproductible ; alors qu'au contraire, il serait plutôt nécessaire de les considérer comme des partenaires dont l'expertise va au-delà du traitement de l'échantillon (ne serait-ce que pour déterminer les biais dans les résultats qui découlent du protocole d'analyse utilisé).

La situation est exactement la même entre protéomiciens et *data scientists* : les premiers peuvent avoir tendance à minimiser l'expertise des seconds, et à réduire celle-ci, soit à du développement logiciel, soit à du "pousse-bouton" (afin de déclencher une analyse statistique puis d'en transmettre les résultats bruts). Le rejet par un *data scientist* des demandes de protéomiciens relevant d'un tel schéma d'interactions n'est pourtant pas souhaitable : en l'absence de réponse à leurs préoccupations quotidiennes, nos interlocuteurs se convaincraient de notre inutilité. Il faut donc lutter contre progressivement cela, afin de petit à petit mettre en place une collaboration plus équilibrée. Finalement, celle-ci peut apparaître pour le bénéfice mutuel des protéomiciens et des *data scientists*, et aboutir à des questions que le groupe KDPD va explorer dans les mois à venir. Voici trois exemples de telles questions :

Prise en compte de données temporelles : au-delà des comparaisons binaires (par exemple sains *vs.* malades, cf. Chap. 2, Sec. 2.5) de nombreuses questions biologiques nécessitent d'analyser l'évolution du protéome au cours du temps. Par exemple au cours d'une infection ou d'une guérison ; aux différentes étapes d'un processus cellulaire ; etc. Pour ce genre d'expériences, des échantillons sont prélevés à différents instants, gelés (afin de stopper l'évolution des processus biochimiques), puis analysés classiquement, en LC-MS/MS. Ensuite, il est attendu ou espéré du praticien qu'il puisse étudier ces données dites "longitudinales". Cela pose néanmoins de nombreuses difficultés : la première tient à l'échantillonnage (presque tout le temps moins de 10 points, souvent moins de 5) qui rend les outils classiques de traitement du signal ou des séries chronologiques inapplicables. La seconde difficulté tient au nombre de questions biologiques différentes que les praticiens regroupent sous le terme "analyse longitudinale" ; souvent parce que la question est encore trop difficile

à préciser compte-tenu des incertitudes liées aux données. Enfin, il y a le nombre des outils et la combinatoire des organisations possibles en pipelines qui rend nécessaire un choix un peu arbitraire. En raison de tout cela, il ne sera possible d'intégrer dans DAPAR et ProStaR qu'un ensemble restreint d'outils, correspondant à un plus grand dénominateur commun. Finalement, le travail de standardisation des questions, de sélection des outils et d'organisation en pipelines relativement génériques reste la principale difficulté pour ce type de données.

Tandem Affinity Purification : Le principe de la *Tandem Affinity Purification* (ou TAP) est celui de la pêche à la ligne : l'idée est d'extraire d'un échantillon complexe, l'ensemble des protéines ayant une forte interaction avec une protéine-appât. De telles expériences sont indispensables à une caractérisation protéomique fine, capable d'aller au-delà de la liste (et de la quantité) des protéines, pour décrire son réseau d'interactions (ce que l'on appelle l'interactomique, cf. Chap. 2, Sec. 2). Au cours d'une TAP, deux conditions de réplicats sont comparées : Celle contenant des réplicats dans lesquels la protéine-appât n'est pas présente "au bout de la ligne" (la condition "contrôle"), et celles contenant les réplicats où elle l'est (la condition "test"). Ainsi, dans les grandes lignes, la manière de traiter les données sera très proche de celle appliquée dans le cas de la comparaison de deux conditions quelconques (comme KO *vs.* WT). Cependant, dans de nombreuses expériences, le praticien souhaite réaliser une pêche pour plusieurs conditions biologiques à étudier, puis les comparer, afin de décrire l'évolution du réseau d'interactions d'une condition biologique à l'autre. Ainsi, on se retrouve classiquement avec plusieurs conditions telles que KO-test, KO-contrôle, WT-test et WT-contrôle, avec comme objectif de les comparer afin de déterminer les variations qui relèvent seulement de la différence entre KO et WT. Il est déjà connu [133] que ce type de questions se traitent avec une ANOVA (***an****alysis* ***of*** ***va****riance* [134]) ; cependant, il reste encore quelques détails à considérer : *(i)* la grande quantité de valeurs manquantes et l'absence de référence permettant une normalisation facile rendent l'analyse statistique particulièrement sensible aux choix des prétraitements (validité des hypothèses comme significativité du résultat) ; *(ii)* l'ANOVA nécessite aussi la mise en place de tests *post-hoc* adaptés, dont l'absence risque d'induire une surinterprétation non-souhaitable des résultats. La mise en place d'un tel pipeline est une priorité pour les utilisateurs de ProStaR, mais nécessite donc encore un peu de réflexion.

Bioanalyse : ce terme désigne l'ensemble des processus à mettre en œuvre pour rapprocher les résultats obtenus via des méthodes issues de la *data science* des projets biologiques dont ils sont issus. Dans le cas de la protéomique, cela signifie être capable de contextualiser une liste de protéines différentiellement abondantes afin de remonter à une sémantique biologiquement pertinente. Cela peut se faire de bien des manières, et il n'y a, à ma connaissance, pas beaucoup de réflexions méthodologiques autour de celles-ci. Néanmoins, je pense qu'il est possible de faire la distinction entre (1) l'enrichissement des données à partir de bases de données extérieures, de la littérature, ou d'expertises ; (2) l'usage d'outils spécifiques permettant de représenter ou d'analyser de manière interactive ces données enrichies, voire de mettre en place des inférences automatiques sur celles-ci. En d'autres termes, et en référence à la pyramide DIK (cf. Fig 1.1), la première étape consiste en la création de contextes supplémentaires ; puis la seconde étape en l'utilisation de ces contextes pour élever le niveau sémantique. Je ne suis pas persuadé que l'intégration d'outils dédiés à la

seconde étape (l'inférence) soit pertinente dans ProStaR, pour des raisons similaires à celles évoquées à propos des données temporelles : il y a trop d'outils dont l'usage est spécifique pour qu'il soit rentable de pousser l'intégration de chacun d'eux. En revanche, la première étape (l'enrichissement des données) doit pouvoir être réalisée dans ProStaR, dans un cadre relativement générique, mais néanmoins suffisamment souple pour tenir compte de la diversité des projets de protéomique. Ce sont des questions que nous commençons tout juste à aborder avec Florence Combes.

3.2 Convergence vers l'IA et la fusion de données

Finalement, les données temporelles, le traitement des TAP et les questions de bioanalyse sont autant d'éléments préfigurant les futures évolutions de ProStaR. Par ailleurs, en complément de ces besoins très pragmatiquement émis par les protéomiciens du laboratoire, il m'est possible d'orienter les développements prenant racines sur DAPAR et ProStaR vers des questions de recherche ouvertes, qui ne se posent pas de manière explicite pour l'instant, mais dont à terme, la résolution sera utile à la discipline. Une telle démarche se fonde sur un choix délibéré de faire converger des connaissances théoriques et un sujet d'application, sans objectif préconçu autre que d'espérer à terme une amélioration quelconque de l'état de l'art (mais avec le risque assumé que cela n'apporte rien à l'utilisateur). Dans ce contexte, j'ai particulièrement à cœur de pouvoir trouver un débouché applicatif aux travaux en IA que j'ai mené en tant que maître de conférences entre 2008 et 2011.

Tous les sujets évoqués plus haut, qu'il s'agisse du contrôle des fausses découvertes, ou encore du traitement de jeux de données contenant des valeurs manquantes, correspondent parfaitement au corpus de problèmes que prétend adresser l'IA, les théories de l'incertain, et plus particulièrement les fonctions de croyance, auxquelles j'ai consacré une part significative de mes travaux jusqu'à présent. Cependant, à l'heure actuelle, il m'est vraiment difficile de faire converger l'expérience que j'ai acquise dans ce domaine, avec des questions de protéomique.

C'est à l'origine durant ma thèse, que j'ai profité du sillage de chercheurs de mon laboratoire d'accueil pour commencer à m'intéresser aux fonctions de croyance. Avec le recul, j'identifie maintenant plusieurs raisons à la naissance de cet intérêt :

- **Raison 1 :** l'approche "sciences humaines" que l'on retrouve en IA plus facilement que dans les autres branches de la science des données, (cf. Introduction), correspondaient à l'idée que je me faisais à l'époque du travail de recherche et de la production d'une "Thèse" au sens étymologique du terme.
- **Raison 2 :** l'objectif de la communauté des fonctions de croyance est de modéliser un processus de décision en s'appuyant sur une description fine de toutes les incertitudes associées aux informations sur lesquelles baser cette décision. Une telle approche me semblait plus intéressante que de considérer aveuglement une décision comme un problème d'optimisation en termes de risques et de gains, tel que rencontré souvent en apprentissage automatique.
- **Raison 3 :** j'ai collaboré durant mon travail doctoral avec une doctorante stambouliote, qui objectivement, était beaucoup plus compétente que moi en traitement d'images et apprentissage automatique. Cela m'a encouragé à renforcer, par complémentarité, mon expertise dans l'usage que l'on pouvait faire des fonctions de croyance au bénéfice de notre sujet d'étude commun.

– **Raison 4 :** avec un financement CIFRE, j'ai passé la moitié de mon doctorat en entreprise, loin de la course à la publication à laquelle doit se livrer un thésard ayant des objectifs de carrière académique. Malgré tout, je devais atteindre un niveau de contribution scientifique suffisant pour l'obtention d'une qualification de Maître de Conférences. Ma stratégie consista donc à m'intéresser à des théories plus confidentielles, et à les appliquer de manière originale à des problèmes préexistants, afin d'en montrer l'apport.

Finalement, je me suis penché sur les fonctions de croyance, autant par intérêt scientifique que par opportunité stratégique. Néanmoins, durant cette période, même si une bonne proportion de mes publications y faisait référence, aucune d'entre elles n'étaient destinée à la communauté des fonctions de croyance proprement dite.

La période où j'ai travaillé à l'université de Bretagne Sud (entre 2008 et 2011) fut une période charnière, pendant laquelle j'ai été particulièrement libre de mes choix de recherche, de leurs positionnements théoriques, comme de leurs domaines d'application. Mon intérêt pour la modélisation du processus de décision n'était pas assouvi ; et la thématique des fonctions croyance gardait cet intérêt stratégique que je lui avais découvert en thèse, et qui me permettait de contrebalancer mon isolement thématique. J'ai donc explicitement cherché à publier dans les journaux et conférences de cette communauté. Finalement, sur mes 21 communications de cette période de 3 ans, on en trouve 3 sur des sujets divers [M'1, O'2, O'3], correspondant à des opportunités de collaborations, 4 relatives au travail de Willy Allègre [O'4, O'5, P'1, O'6], dont je co-encadrais la thèse, 2 en apprentissage statistique [O'7, P'2], et 12 traitant de près ou de loin des fonctions de croyance [S'1–S'3, J'7, P'3, O'8, P'4, O'9, P'5, O'10, P'6, P'7]. Parmi ces dernières, certaines étaient théoriques, d'autres avaient pour sujet d'application la reconnaissance d'écriture ou la combinaison de classifieurs, et donc affichaient un lien avec le domaine de l'apprentissage automatique, sans prétendre être des contributions pures dans ce dernier domaine.

Ainsi, au moment de passer le concours du CNRS, même si mon dossier avait une coloration forte en apprentissage automatique et reconnaissance de formes, la tonalité dominante restait celle de la théorie de Dempster-Shafer. C'est donc avec beaucoup de sincérité que j'ai présenté un projet de recherche équilibré entre ces deux domaines de compétences, avec pour objectif de les mettre au service de la protéomique. Si l'usage de l'apprentissage automatique semblait assez naturel et peu novateur, celui des fonctions de croyance était plus original ; mais je ne doutais pas du bien-fondé de la démarche que je proposais.

Cinq ans plus tard, le constat est clair : parmi tous les cadres théoriques de la *data science* que je maîtrise, celui des fonctions de croyance est celui que j'arrive le moins à appliquer à la protéomique, et celui pour lequel mes travaux précédents se trouvent le moins réutilisables... Et ce n'est pas faute d'essayer. Visiblement, ce constat dépasse ma seule situation : il n'y a pour l'instant pas de domaines d'application posant des problèmes pour lesquels ce cadre méthodologique fournit des solutions qui dominent clairement l'état de l'art. Les publications appliquées du domaine sont donc souvent présentées dans les mêmes média (journaux ou conférences) que les travaux centrés sur la théorie elle-même ; et les fonctions de croyance n'ont pas colonisé de média principalement centré sur une application, comme cela est arrivé avec les outils d'apprentissage automatique, qui font maintenant référence en vision par ordinateur, en analyse de séquences génomiques, ou en reconnaissance de la parole.

Je pense que la communauté des fonctions de croyance se focalise sur des problèmes inadaptés, ou du moins que son bagage méthodologique n'est plus en phase avec les nouveaux problèmes par rapport auxquels elle prétend se positionner : à l'heure du *big data* et de toutes les applications commerciales qui en font le succès, beaucoup de ses nouveaux problèmes ont une connotation *machine learning* forte, qui, à mon sens ne cadre pas avec l'orientation actuelle de la théorie. Cet écart a même tendance à s'accentuer : alors que la théorie avait initialement un cadre statistique et probabiliste fort, ce dernier a été partiellement abandonné au profit du courant logique de l'IA, alors que la communauté du *machine learning* a suivi le chemin inverse, s'éloignant de l'IA pour se rapprocher des mathématiques appliqués. C'est en tout cas ce que j'ai essayé de défendre sans succès dans [M'2]. Cependant, cette vision est partagée. Ainsi Eyke Hüllermeier, éditeur associé du journal *Fuzzy Sets and Systems*, se permet une prise de position similaire dans un article intitulé "*Does machine learning need fuzzy logic ?*" publié fin 2015 [135]. Malgré la forme interrogative, son jugement est particulièrement sévère : tout en s'adressant à une communauté légèrement différente de celle des fonctions de croyance (bien que les deux partagent une culture IA forte), il fait le constat de l'inutilité de ces théories de l'incertain pour les problèmes d'apprentissage, et pointe même la création d'une rente scientifique similaire à celle que j'avais commencé à exploiter durant ma thèse : celle-ci consiste pour un chercheur, à se positionner avec des méthodes originales (les théories de l'incertain les plus confidentielles) sur des problèmes en vogues (*machine learning* et *big data*), indépendamment de l'utilité pour l'application finale et de la validité de l'approche de "fuzzification systématique" des algorithmes.

3.3 Quelques pistes pour cette convergences

Tout cela ne veut pas dire que les fonctions de croyance "ne servent à rien" : selon moi, elles ne peuvent simplement pas être utilisées aveuglément sur tous les problèmes, y compris ceux d'apprentissage en vogue de nos jours. En revanche, dans une logique similaire à celle présentée en Introduction avec la pyramide DIK, elles doivent prendre leur place dans la science des données, et être avantageusement utilisées sur d'autres problèmes correspondant à leur philosophie, et qui ne manquent pas d'apparaître en protéomique. La suite de ce chapitre contient trois exemples de tels problèmes, sur lesquels j'espère pouvoir me pencher à l'avenir.

Méta-scoring ou combinaison de classifieurs : le problème de l'identification de peptides est le premier de ces problèmes. De nombreux moteurs d'identification (DBSE pour *DataBase Search Engine*) sont disponibles pour la communauté, qu'ils soient commerciaux ou non, accessibles en open-sources ou non. Cependant, ces DBSE sont particulièrement complexes : ils sont soumis à de nombreux paramètres, fonctionnent sur des principes différents, et peuvent prendre en entrée différentes bases de données (contenant des séquences génomiques). Finalement, ils sont souvent utilisés comme des "boites noires" par les protéomiciens. Cependant, ceux-ci ont remarqué que les différents DBSE ne donnent pas toujours les mêmes résultats, ou ne permettent pas exactement les mêmes identifications, en fonction de leurs forces et faiblesses respectives, ouvrant la voie à leur combinaison [136, 137], afin d'obtenir une meilleure couverture du protéome. Le principal intérêt des fonctions de croyance réside dans la modélisation fine de plusieurs agents (humains

ou machines) ayant des points de vue partiels, incertains et incomplets ; et dont la combinaison permet un gain informationnel, et donc, une meilleure décision. Ainsi, en considérant que chaque DBSE est un agent, et en encodant sous la forme d'une fonction de croyance le résultat d'identification qu'il fournit, il devrait être possible de fournir un cadre élégant et général au problème de combinaison de DBSE. Il y a cependant deux difficultés : la première est que la puissance de ce formalisme ne s'exprime que quand le nombre de sources est important. Or, à l'heure actuelle, peu de laboratoires utilisent plus de deux DBSE, principalement en raison de leur complexité. La seconde est qu'une fonction de croyance est définie sur l'ensemble des parties de l'univers, et non sur l'univers lui-même. Il en résulte une explosion combinatoire peu compatible avec le nombre de peptides, déjà gigantesque. Cependant, la plupart des DBSE fournissent moins de 10 identifications candidates possibles, et il devrait être possible de travailler sur un univers réduits, limitant la combinatoire.

Tests sur des données imprécises : comme nous l'avons vu plus haut, les valeurs manquantes sont un réel problème en protéomique. Les ignorer n'est simplement pas possible compte-tenu de leur nombre. À l'inverse, les imputer risque de corrompre les données. Une solution intermédiaire serait particulièrement intéressante : plutôt que de considérer qu'il n'y a aucune information disponible quant à la valeur non mesurée (ou de manière équivalente, qu'il n'est pas possible de raffiner son intervalle d'appartenance $[0, +\infty]$), ou au contraire, plutôt que de sur-préciser l'information disponible (en réduisant l'intervalle d'appartenance à une valeur unique, la valeur imputée), il serait logique, facile et plus fiable de définir un intervalle de valeurs plausibles ou vraisemblables, voire même une distribution de probabilité de cette valeur manquante. Cependant, les traitements de données ultérieurs en seraient complexifiés, notamment la réalisation d'un test statistique. De tels tests existent déjà dans certains cas, non-paramétriques [138]. En revanche, dans le cas de tests paramétriques basés sur la loi de Student, très courants en protéomique, la généralisation semble particulièrement difficile. Le sujet n'en reste pas moins intéressant du point de vue théorique, et important du point de vue de la protéomique.

Réinterprétation de la probabilité critique : comme déjà mainte fois discuté, l'interprétation de la p-valeur est contre-intuitif : là où l'on espère connaître $\mathbb{P}(H_0 = 0|S \geq s)$ afin de prendre une décision, le test statistique ne fournit que $\mathbb{P}(S \geq s|H_0 = 1)$; et le passage de celui-ci à celui-là nécessiterait des informations supplémentaires inconnues. En l'absence de celles-ci l'approche bayésienne consistant à les fixer *a priori* n'est pas réellement satisfaisante. En revanche, dans le cadre des fonctions de croyance, il est possible d'utiliser un *prior* non-informatif, et de majorer la probabilité d'intérêt par une plausibilité, permettant une décision à la fois robuste et facilement interprétable. Cette approche, particulièrement récente, que l'on doit à Thierry Denœux [139, 140], est encore difficile à mettre en place en raison de contraintes calculatoires fortes sur l'estimation de la fonction de vraisemblance sous-jacente, qui seront vraisemblablement surmontées dans les années à venir.

4 Conclusion du chapitre

J'espère avoir montré dans ce chapitre que parmi les différentes manières de superviser de l'ingénierie, certaines sont compatibles avec le métier de chercheur, et

que cela peut faire partie de ses attributions. Le principal avantage d'un tel positionnement est qu'il permet de participer activement au décloisonnement disciplinaire.

Les principales difficultés d'un tel positionnement concernent finalement l'évolution de carrière du chercheur qui embrasse un tel choix. Elles sont au nombre de deux : d'abord, il faut affronter la condescendance des théoriciens pour une telle activité. Ainsi, la Section 7 du CNRS a refusé de co-évaluer mon dossier au motif qu'il ne "permettait pas de juger d'une activité relevant de la section" (malgré plusieurs articles publiés dans des journaux canoniques de la section pendant la période évaluée). J'interprète cela comme une décision politique, visant surtout à montrer que les frontières disciplinaires ne sont pas poreuses à souhait ; ce qui est aussi tout à fait entendable (tout le monde n'étant pas expert en tout). Ensuite, il y a la difficulté de valorisation d'une telle activité de supervision d'ingénierie : elle est très consommatrice de temps au regard des publications qu'elle peut générer.

Néanmoins, comme esquissé plus haut, il y a plusieurs types de publications possibles, et avoir conscience de chacun d'entre eux peut rendre viable l'activité :

- Il y a tout d'abord les articles présentant ce travail d'ingénierie : de simples *application notes* pour des packages R à des grosses infrastructures bioinformatiques pouvant être publiées dans les meilleurs journaux par de très gros consortiums, il y a tout un éventail de cibles possibles. Ces papiers à faible valeur ajoutée du point de vue de la recherche *stricto sensu*, peuvent néanmoins générer un nombre important de citations.
- Ensuite, il y a toutes les publications de collaborations, basées sur une utilisation un peu raffinée et un peu intégrée d'un outil pour une problématique biologique dédiée (comme par exemple [J'8]). Cela correspond typiquement aux publications que l'on trouve sur les CV des ingénieurs de recherche, et elles ne peuvent compenser un travail de recherche en propre. Cependant, elles peuvent constituer un bon complément permettant d'entretenir son réseau de collaborateurs.
- Enfin, il y a toutes les publications réellement méthodologiques dont la production est accélérée par l'infrastructure que constituent les outils développés. Il peut s'agir de valoriser un nouvel algorithme précis, dont la diffusion et la validation se trouvent facilitées par l'infrastructure logicielle qui le porte ; ou le fruit de réflexions méthodologiques sur l'usage des outils qu'en fait une communauté, comme évoqué p. 64 et Sec. 2.

Finalement, en jouant sur tous ces tableaux, il devient possible de joindre le "collectivement utile" (le travail interdisciplinaire) à l'individuellement utile (la valorisation efficace).

Chapitre 4

Recherche opportuniste

Ce chapitre décrit un projet de recherche qui relève d'une démarche inverse de celle mise en place pour le projet ProStaR/DAPAR. Cette fois-ci, il s'agit de prendre comme point de départ un domaine de la science des données bien maîtrisé car déjà investigué par le passé, au travers un certain nombre de travaux visant à répondre soit à des questions théoriques, soit à des questions venant d'autres domaines applicatifs ; puis il s'agit, au moyen d'une reformulation adaptée, de faire rentrer un problème de protéomique dans ce cadre méthodologique, afin de proposer une solution nouvelle et originale.

Une telle approche est relativement accessible aux protéomiciens ayant une culture physico-chimique ; mais elle est beaucoup moins naturelle pour ceux qui sont biologistes de formation. En effet, pour ces derniers, une telle démarche semble souvent peu fondée et relever du raisonnement "puisque j'ai un marteau, tous les problèmes sont des clous". C'est effectivement un risque immédiat de la généralisation de cette démarche, il faut le reconnaître. En revanche, appliquée à bon escient, elle a un intérêt réel. Notamment, elle peut permettre la mise en place de méthodologies en rupture avec l'état de l'art.

1 Contexte

Ces travaux trouvent leurs origines dans une collaboration avec Nicolas Courty, Maître de Conférences à l'université de Bretagne Sud dans un laboratoire voisin du mien. À l'époque, Nicolas effectuait ses recherches en animation d'images de synthèse, et mettait en avant une approche très mathématique, fondée sur la géométrie riemannienne (en gros, il s'agit de la géométrie des espaces courbes, par opposition à la géométrie euclidienne, se rapportant au plan). La géométrie riemannienne est particulièrement importante en animation d'images de synthèse : dans l'espace, l'ensemble des positions que peut occuper une extrémité du corps humain (par exemple l'index, la main ou le pied) est caractérisé par une surface courbe régulière nommée *variété riemannienne*.

Comme nous l'avons vu dans la Sec. 5.2 du Chap. 1, la géométrie riemannienne est aussi récemment apparue comme un outil fondamental de la science des données. En effet, quand une population statistique de petite taille est immergée dans un espace de très grande dimensionnalité, il y a de fortes chances pour que celle-ci ne peuple pas tout l'espace, mais plutôt que les individus se concentrent dans certaines

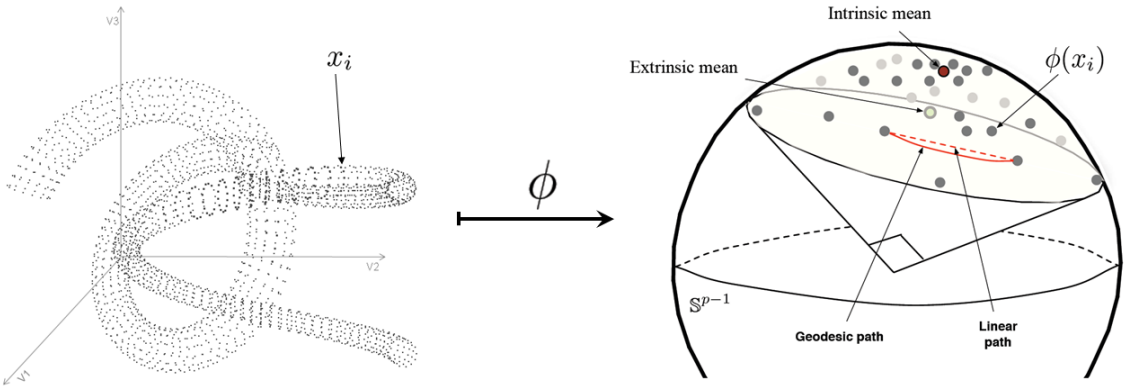

FIGURE 4.1 – *Schéma de l'idée sous-jacente à ces premiers travaux : à gauche, une population statistique est distribuée sur une variété complexe, de sorte que les outils euclidiens ne sont pas adaptés alors que les outils riemanniens sont d'une complexité limitante. En utilisant le* kernel trick*, le problème devient abordable par les outils de la géométrie euclidienne, même si ceux-là ne sont pas parfaitement adaptés. En revanche, si le noyau est bien choisi, les données sont projetées sur une variété suffisamment régulière pour qu'il soit possible d'appliquer les outils riemanniens.*

zones, formant ce que l'on appelle une variété statistique [54], telle qu'illustrée sur la Fig. 1.7 représentant un "S".

Nos premières collaborations nous ont amenés à proposer de nouveaux algorithmes fondés sur une vision riemannienne des problèmes classiques d'apprentissage automatique. Leur point commun était de proposer de travailler sur la variété portant les données d'apprentissage, non pas dans l'espace ambiant comme cela se fait classiquement, mais dans l'espace de Hilbert à noyau Gaussien. La justification est la suivante : ce noyau a la particularité de projeter les données sur une hypersphère (une sphère dans un espace de grandes dimensions), c'est-à-dire une surface tellement régulière qu'il est presque aussi facile de travailler dessus que sur un plan. Ainsi, plutôt que de travailler avec des outils statistiques dérivés de la géométrie euclidienne dans cet espace de Hilbert, comme cela se fait classiquement avec l'astuce du noyau (cf. Chap. 1, Sec. 4.2), nous pouvions utiliser des outils issus de la géométrie hypersphérique (cf. Fig. 4.1). Dans ce cadre, nous avons abordé plusieurs problèmes :

Classification [P'2] : l'algorithme que nous avons proposé est en fait facilement réinterprétable en termes de *kernel subspace classifier* [141]. Chaque classe est représenté dans l'espace de Hilbert à noyau reproduisant par un sous-espace propre, et la distance entre un individu à classer et chaque classe est déterminée par sa distance à son projeté sur l'espace propre en question.

Partitionnement [P'8] : nous avons proposé de reformuler l'algorithme des k-moyennes sur l'hypersphère induite par le noyau gaussien, en utilisant des distances géodésiques, plutôt qu'euclidiennes ; de sorte que les centroïdes des classes appartiennent aussi à l'hypersphère, et que le calcul de leur pré-images soit possible.

Réduction de dimensionnalité [P'9] : en nous inspirant du lien connu [142] entre l'algorithme des k-moyennes et celui de l'ACP, nous avons transposé nos travaux sur le partitionnement à la réduction de dimensionnalité, afin de proposer une version de l'ACP opérant sur l'hypersphère gaussienne, et permettant d'extraire la

sous-hypersphère minimisant l'erreur de reconstruction des données.

Echantillonnage adaptatif [P'9] : il est aussi possible de partir d'un algorithme de partitionnement pour réaliser l'échantillonnage d'une population, en recherchant les représentants de chaque *cluster*, plutôt que les *clusters* eux-mêmes. Un tel algorithme permet ainsi de minimiser l'erreur de représentation de l'échantillonnage. À l'inverse, il peut aussi être intéressant de maximiser la diversité de l'échantillon, ce qui revient à maximiser la variance des individus sélectionnés [1] (ce que l'on appelle l'analyse archétypale [143, 144]). Nous avons proposé des algorithmes permettant de tels échantillonnages sur l'hypersphère gaussienne, plutôt que sur la variété originale des données. Une fois re-projeté dans l'espace d'origine, la qualité de l'échantillonnage était supérieure, tout étant plus rapide à calculer.

Les résultats obtenus sur tous ces travaux étaient tout à fait satisfaisants, et nous ont permis de publier dans des conférences reconnues en apprentissage automatique, comme ECML-PKDD, où nos idées ont essaimé [145, 146]. Cependant, il faut reconnaître que l'incrément de l'état de l'art n'était pas suffisant pour définir un nouveau standard, indépendamment de l'application et de la nature du jeu de données. Ainsi, l'application brutale et systématique de tous nos algorithmes à la protéomique, bien que possible (et d'ailleurs initiée à l'occasion d'une autre collaboration [J'9]) peut sembler un peu artificielle, au sens où elle ne change pas fondamentalement le regard du protéomicien sur ces données. C'est notamment pour cela que je n'ai pas poursuivi plus loin la rédaction de publications de ce type.

À la suite de cela, nous avons abordé le problème de la factorisation de grandes matrices. Celui-ci est central en apprentissage automatique, pour la simple raison qu'il est le fondement algébrique d'une bonne partie des problèmes précédemment mentionnés. Il y a en effet une connexion très forte entre ce problème et l'apprentissage de descripteurs, le partitionnement, la réduction de dimensionnalité, l'échantillonnage, etc. Nous nous sommes donc intéressés à la mise en place d'une méthode de factorisation pouvant être opérée conjointement avec l'astuce du noyau, sous contrainte de parcimonie, ainsi que d'une complexité linéaire par rapport à la taille des données. Ce n'est qu'ensuite, une fois arrivé à EDyP, que j'ai réalisé que notre algorithme fournissait une solution élégante au problème de démultiplexage des spectres de fragmentation obtenus en mode DIA (cf. Chap. 2, Sec. 3.5).

Pour rappel de ce que nous avons vu au Chap. 2, le principe de la DIA est de sélectionner plusieurs ions précurseurs (éventuellement tous), afin de les fragmenter simultanément. L'avantage est de pouvoir fragmenter un plus grand nombre d'ions précurseurs, et donc d'augmenter la profondeur de l'analyse spectrométrique. L'inconvénient est que les spectres de fragmentation des différentes espèces se retrouvent superposés (ou mélangés), empêchant une identification directe par un moteur classique comme Mascot.

2 Etat de l'art : interprétation de spectres DIA

Il existe pour l'instant plusieurs alternatives pour le traitement de signaux issus de DIA, mais, en fait, celles-ci ne règlent pas complétement le problème. Une pre-

1. Il se trouve que ce second critère est particulièrement proche de celui mis en œuvre pour la première étape de l'algorithme de factorisation de matrice qui est décrit plus loin dans ce chapitre.

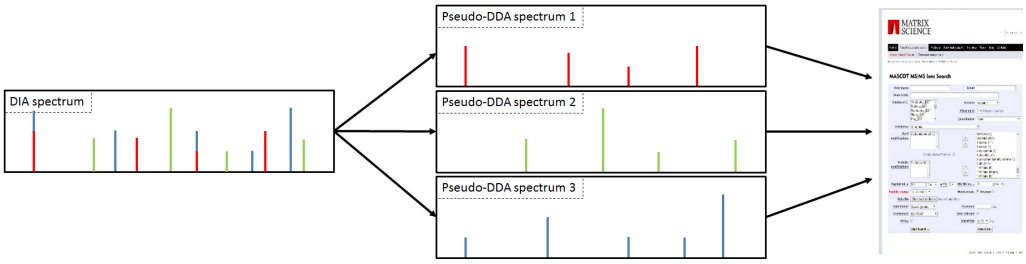

FIGURE 4.2 – *Principe du démultiplexage de spectres MS/MS : à partir d'un spectre contenant les raies de plusieurs ions précurseurs (représentés de couleurs différentes), on cherche à reconstruire les spectres que l'on aurait obtenu en DDA, à savoir des spectres ne contenant qu'un et un seul précurseur, de sorte que l'identification par un moteur classique soit efficace.*

mière solution consiste à ne pas utiliser d'approche DIA pour l'identification initiale des peptides (importante en protéomique de découverte), mais seulement pour leur ré-identification et leur quantification (dans le cas d'expériences de protéomique ciblée). C'est notamment ce que propose la méthode SWATH [147], ainsi qu'un certain nombre d'autres méthodes moins répandues. La seconde approche consiste à trouver un moyen de séparer les spectres qui ont été mélangés, tel que cela est illustré sur la Fig. 4.2. Cependant, cette séparation est rendue difficile par le fait que, même si le résultat du mélange des signaux (souvent appelé *spectre chimérique*) est connu, ceux-ci ne sont pas connus individuellement : nous verrons donc dans la Sec. 3 qu'il faut réaliser une *séparation aveugle des signaux* [148], pour laquelle justement la factorisation matricielle constitue un cadre de travail intéressant.

2.1 Méthodes analytiques

Dans les premiers travaux sur l'interprétation de spectres DIA, il a été envisagé de contourner la nature "aveugle" du multiplexage par des raffinements analytiques [149]. Grâce à ceux-ci, il est devenu possible d'accéder à des informations supplémentaires quant aux signaux à séparer : l'ensemble des spectres simples qui ont été multiplexés, ou encore une description plus fine du processus de multiplexage.

La plus populaire de ces méthodes est SWATH [147]. Elle a d'abord été introduite comme une méthode de suivi d'un grand nombre de transitions dans une approche de type PRM (*Parallel Reaction Monitoring* [150]), puisque toutes les deux sont adaptées à la quantification des protéines précédemment identifiées. En effet, avec SWATH, l'information supplémentaire utilisée est contenue dans des bibliothèques spectrales, c'est-à-dire des ensembles d'identifications précédemment réalisées et stockées. C'est la raison même pour laquelle ce protocole ne peut pas être utilisé en protéomique de découverte.

Formellement, la technologie SWATH est limitée aux instruments du constructeur AB Sciex, puisque les outils bioinformatiques correspondants (ProteinPilots et Peakview) sont limités aux formats de fichiers de ces instruments. Cependant, il est possible d'imiter ce pipeline sur d'autres instruments grâce à différentes technologies ou logiciels : Spectronaut [151] (qui repose en grande partie sur des peptides de référence incorporés dans les échantillons, afin d'aligner les temps d'élution entre

bibliothèques acquisitions), OpenSWATH [152], Skyline [153], etc.

Finalement, la philosophie derrière toutes les méthodes "*SWATH-like*" est la même. Elles ne cherchent pas à démultiplexer les spectres à proprement parler, mais plutôt à faciliter l'exploitation de leur contenu en rendant accessible le contenu des signaux multiplexés : concrètement cela se fait grâce à l'acquisition préalable de bibliothèques spectrales, ainsi qu'en supposant que les spectres multiplexés sont chacun une somme de plusieurs spectres DDA connus et issus de ces bibliothèques.

Une autre catégorie de méthodes est basée sur une philosophie différente : aucune bibliothèque spectrale n'est impliquée, de sorte que l'identification dans un contexte de protéomique de découverte devient possible. En contrepartie, le pipeline d'analyse est modifié de sorte que la complexité du processus de multiplexage est maîtrisée ; ce qui rend le démultiplexage possible.

Ainsi, PAcIFIC [154] propose d'examiner de très petites fenêtres MS (2,5 Th), de sorte que, même si plusieurs précurseurs sont co-analysés, leur nombre est susceptible de rester assez petit pour induire un "multiplexage faible" qui ne devrait pas être trop difficile à traiter. Le principal inconvénient de cette approche est que plus les fenêtres sont petites, plus le nombre de cycles nécessaires à la couverture de la gamme m/z est important, jusqu'à nécessiter d'augmenter le nombre d'injections MS dans des proportions incompatibles avec les contraintes d'une plateforme protéomique (dans [154], une analyse de 5 jours et 67 injections est rapportée).

Pour contourner cette limitation, il a été proposé avec MSX [155] de fusionner plusieurs petites fenêtres d'isolation (de l'ordre de 4 Th, ce [2] qui est à peu près du même ordre de grandeur qu'avec PAcIFIC) en de grandes fenêtres d'isolation (de l'ordre de 20 Th, soit du même ordre de grandeur qu'avec SWATH) pour accélérer la couverture de la gamme m/z. Cependant, à chaque fois, les 5 petites fenêtres à fusionner ensemble sont choisies de façon aléatoire, de manière à pouvoir espérer que toutes les combinaisons apparaissent, rendant par la même occasion le démultiplexage beaucoup plus simple : à titre d'illustration, si l'on fusionne d'abord les fenêtres {W1, W2} puis les fenêtres {W2, W3}, et finalement les fenêtres {W1, W3}, il est possible de récupérer les spectres de la fenêtre W2 par l'opération suivante :

$$(\{\text{W1, W2}\} + \{\text{W2, W3}\} - \{\text{W1, W3}\})/2$$

Fondamentalement, l'idée est donc simplement de multiplexer de façon intelligente les petites fenêtres MS pour accélérer les analyses et ainsi compenser les inconvénients de PAcIFIC. Il est intéressant de noter que, d'un point de vue mathématique, un tel multiplexage correspond à une transformation de Hadamard, à savoir un produit de matrice avec un motif de convolution prédéterminé et connu. En d'autres termes, les mathématiques derrière le démultiplexage de MSX correspondent à une forme faible de celles proposées ici. Dans le cas de MSX, cette forme plus faible de démultiplexage est suffisante, puisqu'en forçant la stratégie de multiplexage à un motif spécifique, le problème devient plus facile à résoudre. Cependant, cela a des conséquences sur les résultats d'analyse : d'abord, la définition des fenêtres d'isolation est limitée par les capacités du codage d'Hadamard plutôt que les contraintes analytiques ; deuxièmement, la répartition aléatoire réduit le nombre d'analyses pour chacune des fenêtres MS, de sorte que les spectres sont de moins bonne qualité (par

2. Le Thomson est l'unité de mesure des rapports masse-sur-charge.

exemple, il n'y a pas nécessairement d'acquisition à l'instant précis de l'apex du chromatogramme).

Finalement, toutes ces méthodes reposent sur des modifications du pipeline d'analyse permettant de rendre le problème de démultiplexage assez simple pour être résolu par une méthode naïve. S'il était possible de résoudre efficacement le problème de démultiplexage aveugle uniquement avec des outils informatiques, l'expertise analytique pourrait être focalisée sur ce qu'elle permet de meilleur : l'amélioration du pipeline d'acquisition afin d'augmenter la couverture du protéome, à la fois en termes d'identification (le niveau de qualité des spectres) et de quantification (reproductibilité des mesures). Heureusement, des efforts ont été menés en ce sens.

2.2 Méthodes computationelles

Parmi les différentes approches formelles visant à résoudre le problème de démultiplexage aveugle, une première catégorie d'algorithmes propose de boucler entre les deux étapes suivantes : (1) identifier quelques fragments dans les spectres multiplexés (2) soustraire ceux-ci des spectres multiplexés. L'intuition sous-jacente est que plus le nombre de fragments enlevés est important, plus il devient facile d'identifier ceux qui restent ; de sorte qu'il est possible d'espérer que cette stratégie itérative conduise finalement à l'identification de tous les fragments dans les spectres multiplexés. Concrètement, non seulement le score d'identification est utilisé pour déterminer quels fragments peuvent être identifiés et soustraits, mais la corrélation entre l'intensité des précurseurs et celle des fragments[3] l'est aussi ; cela fait sens, dans la mesure où la plupart des outils d'identification reposent en grande partie sur la masse du précurseur pour guider l'identification. Dans cette catégorie, on trouve FT-ARM [156], ProIDTree [157] (qui a été conçu pour traiter les spectres multiplexés indépendamment de leur protocole d'acquisition) et PLGS (logiciel spécifique pour la méthode d'acquisition MS^E [158]). De plus, il existe quelques outils logiciels initialement conçus pour traiter les spectres accidentellement multiplexés en DDA (lorsque deux ions précurseurs ne peuvent être séparés en raison de ratios m/z trop proches) plutôt que de véritables spectres DIA ; mais qui peuvent être utilisés car basés sur des algorithmes similaires : MixGF [159], Physikron [160] (dont l'usage est restreint à un couplage avec le logiciel Mascot) et Complementary Finder [161] (qui tente de distinguer des fragments de différents précurseurs en associant chaque fragment à son fragment complémentaire).

Le principal problème de ces méthodes est que l'identification et le démultiplexage ne sont pas séparés en étapes distinctes, de sorte que l'on est en droit de questionner la validité statistique de l'usage consécutif d'une approche *target-decoy* [162, 163] pour estimer le taux de fausses découvertes, alors que cela est pourtant classique. Plus généralement, ces algorithmes ne cherchent pas à démultiplexer d'abord, puis à identifier ensuite les pseudo-spectres DDA ainsi générés comme on identifierait des spectres DDA classiques. Par conséquent, il n'est pas vraiment possible de tirer parti de l'intégralité des pipelines bioinformatiques qui sont couramment utilisés en DDA, et qui ont prouvé leur efficacité par une décennie d'utilisation.

3. La section suivante explique plus en détail pourquoi la chromatographie liquide introduit une importante corrélation temporelle entre les signaux MS et MS/MS.

Voilà pourquoi, d'autres algorithmes suivent une stratégie légèrement différente, où des fragments sont itérativement soustraits à partir des spectres multiplexés et regroupés en pseudo-spectres DDA, avant toute identification. Pour ce faire, les scores d'identification peuvent là encore être utiles, mais il est également possible de donner plus d'importance à la corrélation entre les fragments ou entre un fragment et un précurseur. En effet, en raison du protocole d'acquisition, tous les sommets du chromatogramme associé à un peptide donné (soit en MS1 ou MS2) devraient co-éluer. AIF [164] (réservé aux instruments ThermoFisher Scientific, ainsi qu'à une identification avec Andromeda) et XDIA [165] suivent cette stratégie. Cette idée est poussée plus loin dans un dernier groupe de méthodes, où la soustraction itérative du fragment n'est plus considérée pour générer des pseudo-spectres DDA. La raison en est que, en fonction de l'ordre de traitement des fragments, on peut se retrouver avec des pseudo-spectres DDA différents. C'est pourquoi, il est plutôt proposé d'appliquer un algorithme de *clustering* sur les fragments (la mesure de similarité utilisée pour la classification étant liée à la corrélation entre les profils d'élution), et de construire les pseudo-spectres DDA après, sur la base des *clusters*. Le premier outil à proposer cette stratégie était DEMUX [166]. L'outil le plus récent de l'état de l'art, DIA-Umpire [167] suit aussi cette stratégie.

3 Reformulation du problème

3.1 Discussion sur l'état de l'art

Dans notre laboratoire, une comparaison a récemment été menée entre les capacités d'identification d'un pipeline classique DDA, et d'un pipe-line DIA ne s'appuyant pas sur une bibliothèque de spectres, en adéquation avec les contraintes de la protéomique de découverte. Plus concrètement, la comparaison a porté sur le pipeline classiquement utilisé au laboratoire d'une part, et sur l'outil le plus récent, DIA-Umpire, d'autre part. Même si ce dernier semble surpasser les performances annoncées par d'autres outils propres à la DIA, les résultats qu'il fournissait était systématiquement moins bons que ceux du pipeline de référence (qui fournissaient entre 12 % et 84 % des identifications supplémentaires). Cela semble indiquer que pour l'instant, la seule façon d'espérer que les méthodes DIA permettent une amélioration de la couverture des protéomes (ce qui est l'argument ayant motivé cette direction de recherche), est de compter sur les bibliothèques spectrales précédemment acquises par des approches DDA.

Selon moi, voici la raison de cet échec relatif : le démultiplexage aveugle de signaux complexes est un problème difficile. Toutefois, ce problème n'est pas inconnu. Il a même beaucoup été étudié, car il s'agit d'une question récurrente en ingénierie des télécommunications (comment peut-on séparer le bruit du signal de la voix dans une conversation téléphonique ? Comment peuvent être supprimés les échos - lorsque les ondes radio se reflètent sur les murs en béton ? etc.). Voilà pourquoi, une théorie complète de la séparation aveugle de sources (ou séparation aveugle de signaux) a été construite par la communauté de traitement du signal depuis la fin des années 80 et au début des années 90, sur la base des travaux initialement menés par Hérault et Jutten [148] . Cette théorie était bien stabilisée autour des années 2000, avant

de subir un vif regain d'intérêt il y a dix ans. En effet, l'un des principaux outils mathématiques sur lesquels elle se fonde, à savoir la factorisation de matrices [168], s'est avéré être d'une importance primordiale en vision par ordinateur et en indexation de contenus web [44]. Depuis, cette théorie de la séparation de sources a été profondément renouvelée, tout en convergeant vers l'apprentissage statistique, les mathématiques appliquées, et le traitement de gros volumes de données. Assez logiquement, les travaux les plus récents autour du traitement de données protéomiques n'ont pas encore eu le temps d'être irrigués par ces évolutions de l'état de l'art en *data science*. Voilà pourquoi, au lieu de traiter le démultiplexage des spectres DIA via le formalisme de la séparation aveugle de signaux, de nombreuses heuristiques ont été utilisées pour contourner le problème et le refondre dans des cadres plus classiques, tels que le *clustering*, ou ces algorithmes itératifs "d'identification et soustraction". Cependant, tous ces efforts ne doivent pas non plus être décriés. L'intuition de se fonder sur l'algèbre matricielle pour résoudre une version plus simple du problème (dans des algorithmes tels que DEMUX ou MSX) montre déjà que leurs auteurs ont su par leur propres moyens redécouvrir en partie des résultats de la séparation aveugle de sources. De même, un élément de base de cette théorie est que le démultiplexage ne peut être résolu que si certaines hypothèses (notamment de régularité) peuvent être faites sur les signaux à séparer. Cette idée apparaît également dans plusieurs algorithmes (par exemple MSE ou DIA-Umpire) à travers l'hypothèse selon laquelle tous les pics de fragments et l'ion précurseur devraient avoir des profils chromatographiques fortement corrélés, voire même partager une forme spécifique.

3.2 Corrélation temporelle résultant de la LC

Pour pouvoir séparer des signaux inconnus et multiplexés, il est nécessaire d'avoir accès à une autre information révélant, au moins en partie, la structure des signaux à séparer. Comme cela est brièvement évoqué dans la section précédente, en protéomique, cette structure est naturellement véhiculée par la cohérence temporelle de la chromatographie liquide. En effet, dans une analyse LC-MS/MS, les différentes copies d'un peptide éluent plus ou moins à la même vitesse, de sorte que l'abondance de l'espèce peptidique en question varie au court du temps : de nulle, celle-ci augmente progressivement jusqu'à atteindre un maximum, puis décroît avant de finalement disparaître, une fois que toutes les copies sont sorties de la colonne. Cette cohérence temporelle apparaît aussi évidemment au niveau des spectres de fragmentation, de sorte que tous les pics du spectre d'un même peptide apparaissent, atteignent le maximum et disparaissent en même temps, tel que cela est illustré sur la Fig. 4.3. Grâce à cela, il va nous être possible d'utiliser un algorithme de factorisation de matrice pour réaliser cette séparation des signaux.

3.3 Introduction du formalisme matricielle

Pour cela, commençons par définir les matrices sur lesquelles travailler. Le plus naturel reste de considérer une matrice contenant des spectres (en lignes) et des chromatogrammes (en colonnes), sur le modèle de la Fig. 4.3. Nous appellerons une telle matrice un *chromatospectre*.

Maintenant, revenons au fonctionnement du spectromètre, afin de comprendre

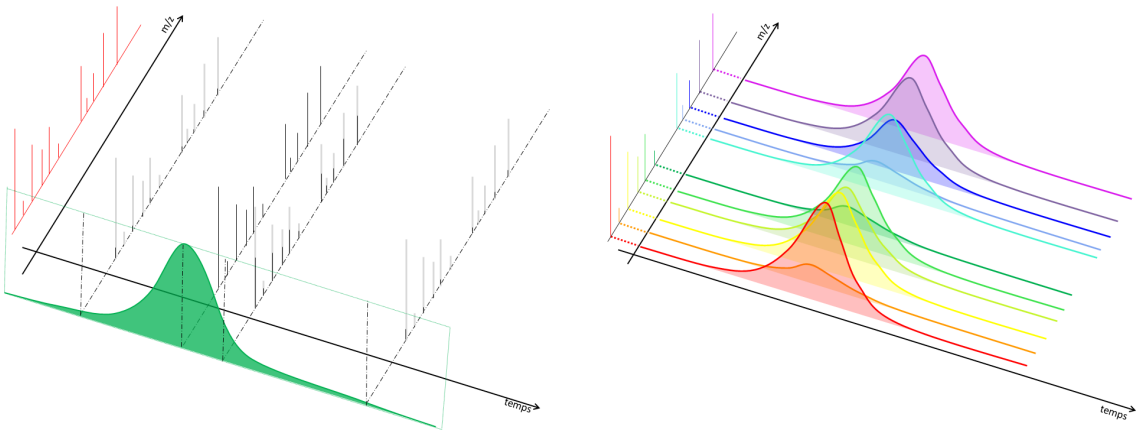

FIGURE 4.3 – *Illustration du décours temporel d'un spectre de fragmentation au fur et à mesure de son élution par chromatographie. À Gauche, le signal est découpé en tranches temporelles, faisant apparaître un spectre de fragmentation d'intensité différente à chaque instant ; à droite, le signal est découpé en tranches de masse, faisant apparaître des chromatogrammes identiques, mais d'intensité correspondant à celle de chaque pic.*

comment se constituent les spectres chimériques. Nous supposerons dans un premier temps que deux peptides avec des chromatogrammes différents se retrouvent co-fragmentés. Pour chacun de ces peptides, nous avons un chromatospectre spécifique, tel que cela est représenté par les deux matrices de gauche sur la partie supérieure de la Fig. 4.4 : le chromatogramme et le spectre du premier peptide sont représentés en jaune et rouge respectivement. Une fois discrétisé, le décours temporel du spectre peut être représenté par le chromatospectre dont l'intensité des valeurs est représentée par le diamètre des disques oranges. De manière similaire, pour le second peptide, le chromatogramme, le spectre et le chromatospectre sont représentés respectivement en vert, en bleu et en turquoise. Si les deux peptides co-éluent partiellement et se retrouvent à un moment tous les deux dans la chambre de fragmentation, les fragments ayant la même valeur de m/z vont participer à l'intensité d'un même pic. Plus généralement chaque pic aura une intensité correspondant à la somme des intensités du pic en question dans les deux chromatospectres. Finalement, le chromatospectre chimérique correspondant s'obtient par l'addition (au sens matricielle du terme) des deux chromatospectres. C'est ce qui est représenté sur la matrice de droite de la première ligne de la Fig. 4.4.

Pour chacune de ces trois matrices, le chromatogramme et le spectre sont respectivement représentés à gauche et au-dessus, car il s'agit des *marginales* des chromatospectres. En effet, si l'on somme (ou que l'on moyenne, ici, à peu de choses près, cela revient au même) l'ensemble des colonnes d'un chromatospectre, nous obtiendrons une colonne correspondant au chromatogramme. De même, si l'on somme ou moyenne l'ensemble des lignes d'un chromatospectre, nous obtenons une ligne qui correspond au spectre. La relation qui lie les marginales à la matrice n'est pas la même pour les 2 peptides (matrices à gauche et au centre) et pour le mélange (à droite). En effet, dans la mesure où, pour chaque peptide, son information d'élution (le chromatogramme) et de fragmentation (le spectre) sont indépendantes (au sens où l'information de l'un ne contraint en aucun cas l'information de l'autre),

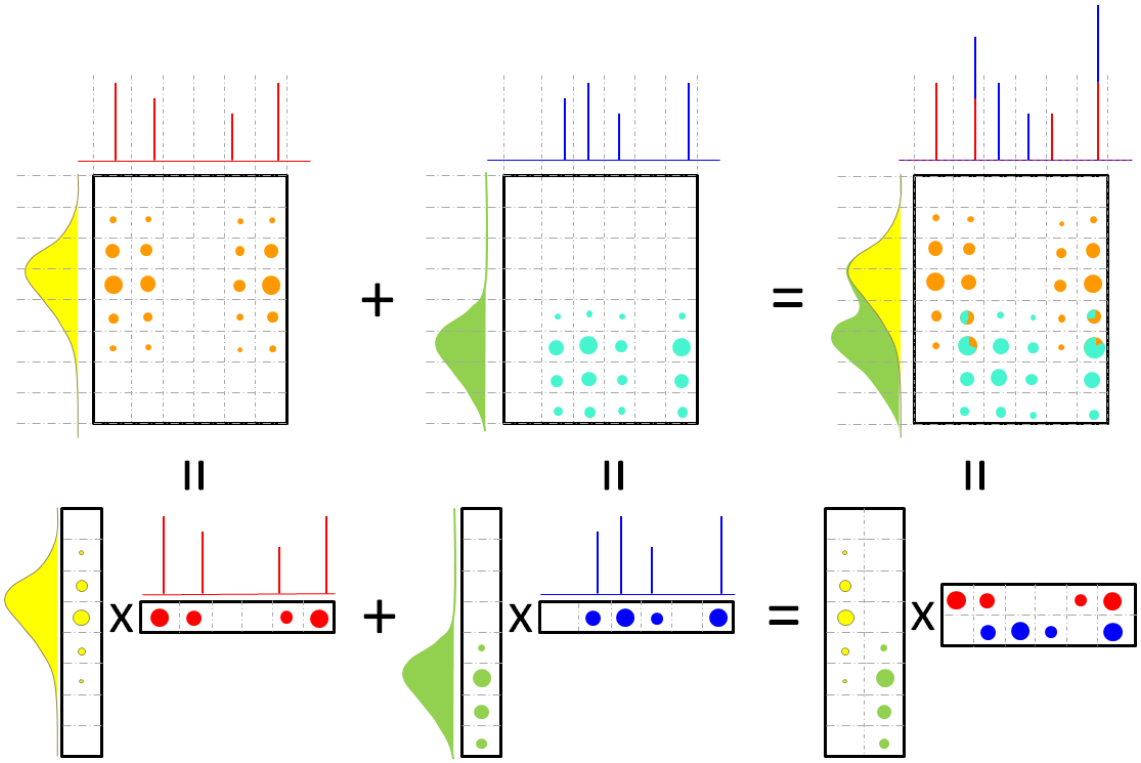

FIGURE 4.4 – *Représentation schématique d'un chromatospectre multiplexé : il peut être vu comme l'addition de deux chromatospectres, ou comme un produit de matrices contenant respectivement des chromatogrammes et des spectrogrammes.*

chaque chromatospectre a la propriété de pouvoir être reconstruit intégralement à partir du produit de ces deux marginales. Ainsi, sur la seconde ligne, il apparaît bien que le produit du vecteur-colonne (chromatogramme) jaune (ou vert) par le vecteur-ligne (spectre) rouge (ou bleu) permet de reconstruire la matrice (chromatospectre) orange (ou turquoise). En revanche, ce n'est pas le cas pour le mélange des deux peptides, comme cela apparaît sur la Fig. 4.5 : un chromatospectre chimérique ne peut en aucun cas être reconstruit à partir du chromatogramme cumulé et du spectrogramme cumulé des deux peptides.

En revanche, les règles du calcul matriciel nous disent que la somme de deux produits de deux vecteurs chacun (respectivement ligne et colonne) est égale au produit de deux matrices correspondant chacune aux vecteurs lignes et colonnes concaténés. Autrement dit :

$$(V_1 \times V_2^\top) + (V_3 \times V_4^\top) = [V_1; V_3] \times [V_2; V_4]^\top$$

avec V_i un vecteur-colonne, $V_i^\top$ un vecteur-ligne et " ;" représentant la concaténation de deux vecteurs-colonnes en une matrice à deux colonnes. Cette équation est aussi représentée de manière plus imagée sur la seconde ligne de la Fig. 4.4. Finalement, il apparaît que si nous sommes capables d'écrire un chromatospectre chimérique comme un produit de deux matrices, alors, les colonnes de la première correspondront à des chromatogrammes, et les lignes de la seconde à des spectres, tels que cela est illustré sur la Fig. 4.6. Dès lors, il apparaît complétement justifié de chercher à factoriser un chromatospectre : cela permet de démultiplexer des spectres chimé-

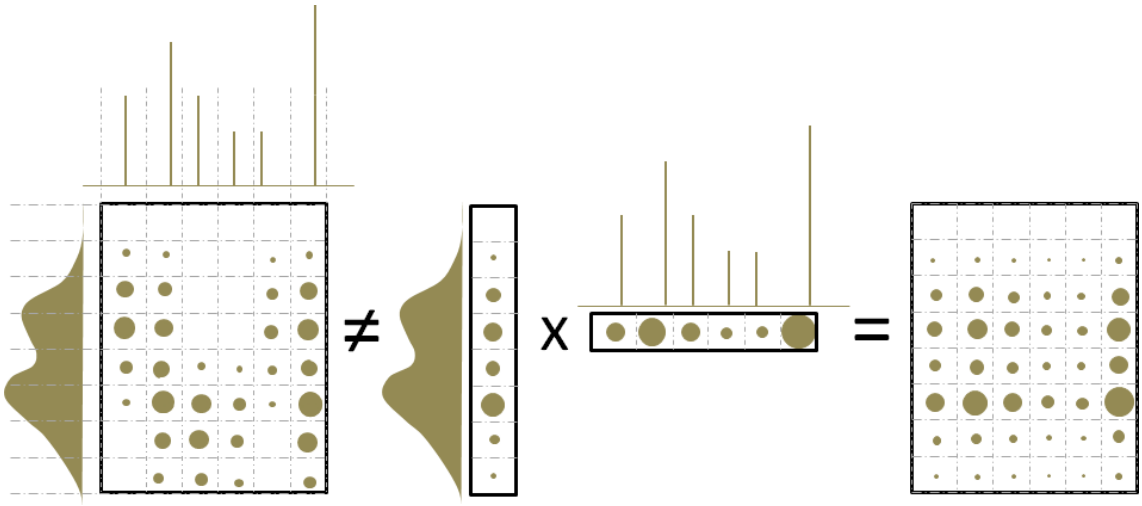

FIGURE 4.5 – *Un chromatospectre multiplexé ne peut être reconstruit sur la base de ses marginales.*

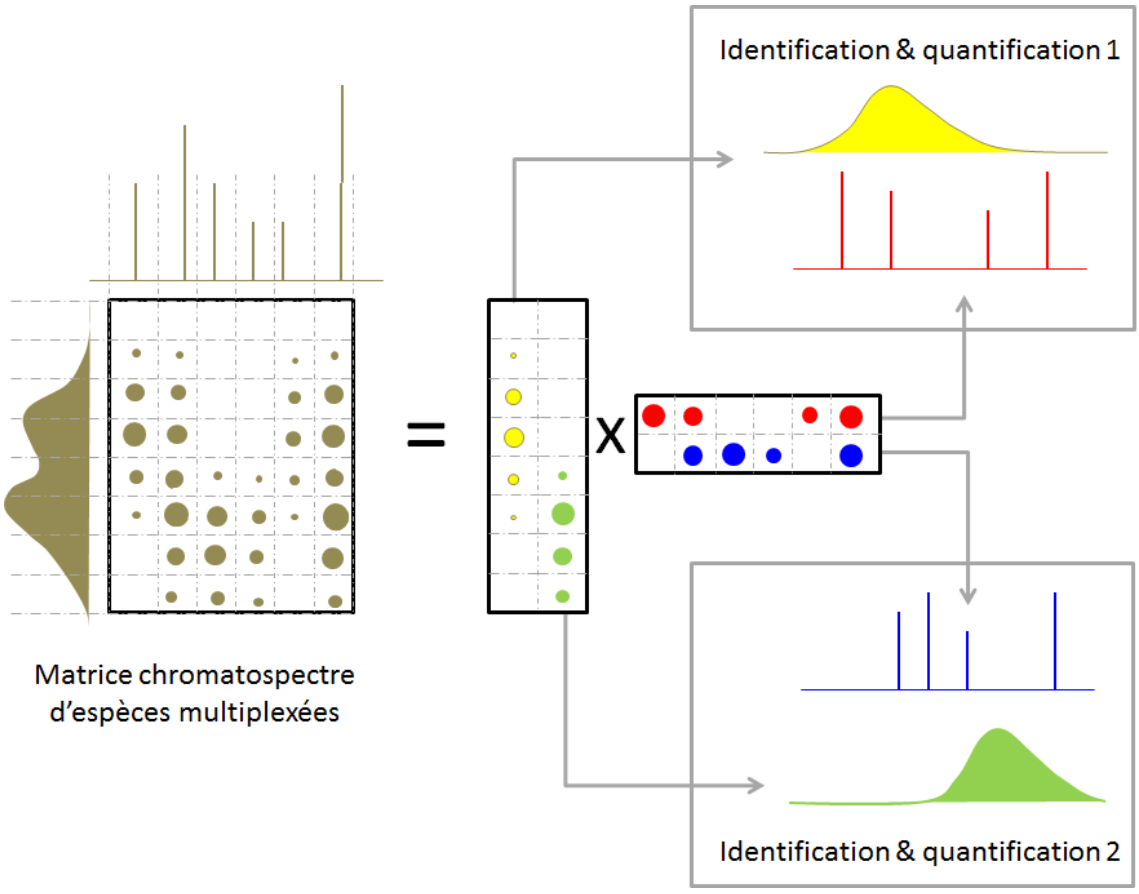

FIGURE 4.6 – *Si l'on arrive à factoriser un chromatospectre, il est possible d'interpréter les matrices résultantes en termes de chromatogrammes et de spectrogrammes.*

riques. Dans le cas général, où N ions précurseurs sont multiplexés, les matrices facteurs doivent contenir respectivement N colonnes, et N lignes.

4 Algorithme SAGA

Nous venons de voir que sur le principe, une factorisation de matrice nous permet de résoudre le problème de démultiplexage de spectres DIA. Nous allons maintenant découvrir comment réaliser en pratique cette factorisation. Dans notre cas, la difficulté vient principalement de la taille de la matrice : alors que l'usage de méthodes de factorisation pour la séparation de signaux est connu depuis quelques décennies maintenant (cf. Sec. 3.1), il était jusqu'à il y a peu, inenvisageable d'appliquer ces méthodes à des données aussi complexes et volumineuses que celles produites en spectrométrie de masse. En effet, si l'on souhaite une discrétisation suffisamment fine pour que les détails ayant une pertinence chimique ne soient pas perdus, la matrice à considérer peut contenir jusqu'à quelques milliards de valeurs : par exemple, si les valeurs de m/z sont arrondies à 3 décimales après la virgule, et si l'on souhaite couvrir la gamme allant de 100 à 1700 Th, alors la matrice aura 1 600 000 colonnes ; quant au nombre de lignes, il dépend de la longueur de la colonne d'élution, mais il y peut y avoir facilement plusieurs milliers de cycles MS/MS. Ce ne sont que le récent renouveau de la discipline, précédemment mentionné, couplé à l'augmentation de la puissance de calcul, qui permettent de commencer à considérer des matrices aussi grandes que nécessaires dans un contexte d'application protéomique.

4.1 Les origines de la factorisation de matrices

Historiquement, il y avait deux manières de décomposer une matrice $\mathbf{M}$ en un produit de deux matrices $\mathbf{A}$ et $\mathbf{B}$: la décomposition en valeurs singulières et la programmation alternée.

Décomposition en valeurs singulières : cette première méthode est aussi vieille que l'algèbre, puisqu'il s'agit à peu de choses près d'une diagonalisation, que nous avons déjà mentionnée au Chap. 1, Sec. 5.1. Grâce à cette procédure, il est possible d'écrire $\mathbf{M} = \mathbf{U} \cdot \mathbf{D} \cdot \mathbf{V}$ où $\mathbf{D}$ est une matrice diagonale. Ensuite, il suffit de poser $\mathbf{A} = \mathbf{U} \cdot \mathbf{D}$ (ou $\mathbf{A} = \mathbf{U}$) et $\mathbf{B} = \mathbf{V}$ (ou $\mathbf{B} = \mathbf{D} \cdot \mathbf{V}$) pour obtenir la factorisation désirée. Cette méthode a deux inconvénients : tout d'abord, quelle que soit la procédure de calcul des valeurs propres (il en existe de nombreuses), la complexité est toujours plus ou moins la même, à savoir cubique. Cela signifie que si l'on multiplie la taille des données par 10, alors, les ressources de calculs (temps et mémoire) sont multipliées par 10 au cube, à savoir 1000. Cela rend la procédure trop coûteuse pour des gros jeux de données. Ensuite, les matrices $\mathbf{A}$ et $\mathbf{B}$ sont le résultat d'opérations algébriques qui ne peuvent pas forcément être reliées à la réalité physique des données. Ainsi, si $\mathbf{M}$ contient des valeurs d'intensités MS/MS, nous nous attendons à des matrices $\mathbf{A}$ et $\mathbf{B}$ qui représentent des signaux de même nature, et donc il est légitime d'attendre de ces matrices qu'elles ne contiennent pas de valeurs strictement négatives ; ce que ne peut pas garantir la décomposition en valeurs singulières.

Programmation alternée : cette seconde méthode est beaucoup plus récente, puisque les travaux préliminaires datent du milieu des années 90 [169], et que la publication de référence montrant que le système visuel humain réalise un traitement similaire à une factorisation de matrice, justifiant ainsi l'usage de cette technique en vision par ordinateur, date de 1999 [168]. Le principe est d'appliquer l'Alg. 3,

Algorithm 3: Factorisation de matrice par programmation alternée

```
Input : Matrix M ; Parameter : Error ϵ
Initialization : r ← ∞ ; A, B ← RandomInitialization()
repeat
    A ← Update(A, B)
    B ← Update(B, A)
    r ← Norm(M − A · B)
until r ≤ ϵ;
```

qui consiste, à partir d'un couple de matrices ($\mathbf{A}$, $\mathbf{B}$) choisi aléatoirement, à effectuer des modifications successives de sorte que la matrice résiduelle $[\mathbf{M} - \mathbf{A} \cdot \mathbf{B}]$ contienne des coefficients de plus en plus petits à chaque itération. Pour peu que la modification itérative de $\mathbf{A}$ et $\mathbf{B}$ soit faite correctement, il est prouvé que cet algorithme converge vers une solution, tout en préservant l'éventuelle non-négativité des coefficients (contrairement à la méthode précédente). Cependant le coût calculatoire reste très important, et la vitesse de convergence est difficile à contrôler, de sorte que cet algorithme n'est pas non plus adapté à des matrices aussi grandes que celles rencontrées en protéomique.

4.2 Les méthodes "modernes"

La première communauté scientifique à avoir cherché à dépasser les limitations des méthodes précédemment décrites est celle de l'imagerie hyperspectrale [4] (cf. [170] pour une revue de la question), avant d'être rejointe dans cet effort par les communautés de l'apprentissage automatique [171–173], de la vision par ordinateur [44] et de l'acquisition comprimée [41, 174]. À l'heure actuelle, il existe de nombreuses méthodes de factorisation, mais la plupart sont basées sur un principe similaire, constitué de deux étapes.

La première étape consiste à trouver un sous-ensemble des colonnes de $\mathbf{M}$ qui encodent des signaux élémentaires, et qui peuvent être utilisés pour reconstruire des signaux plus complexes. En d'autres termes, on cherche à isoler un dictionnaire (cf. Chap. 1 Sec. 4). Ce dictionnaire est souvent appelé CSS (*Column Selection Subset*), mais d'autres noms apparaissent dans la littérature (*Riesz basis*, *endmembers* ou encore *anchorpoints*). L'ensemble des éléments vecteurs-colonnes qui forment le CSS sont rassemblés en une matrice pour définir $\mathbf{A}$.

La seconde étape consiste à reconstruire toutes les colonnes de $\mathbf{M}$ qui ne sont pas dans le CSS à partir des éléments du CSS. D'un point de vue mathématique, cela revient à projeter les colonnes de $\mathbf{M}$ sur une région spécifique (un cône, un simplex, ou l'espace entier) de l'espace engendré par le CSS. Le résultat de ces projections forme alors la matrice $\mathbf{B}$.

Il a été effectivement montré que l'enchaînement de ces deux étapes, à savoir la sélection intelligente d'un dictionnaire, suivi d'une série de projections permet de réaliser la factorisation d'une matrice de manière très efficace. En théorie, il est même

4. L'analyse d'images acquises par satellites, avec plusieurs dizaines de couleurs primaires, au lieu des trois que peuvent percevoir nos yeux.

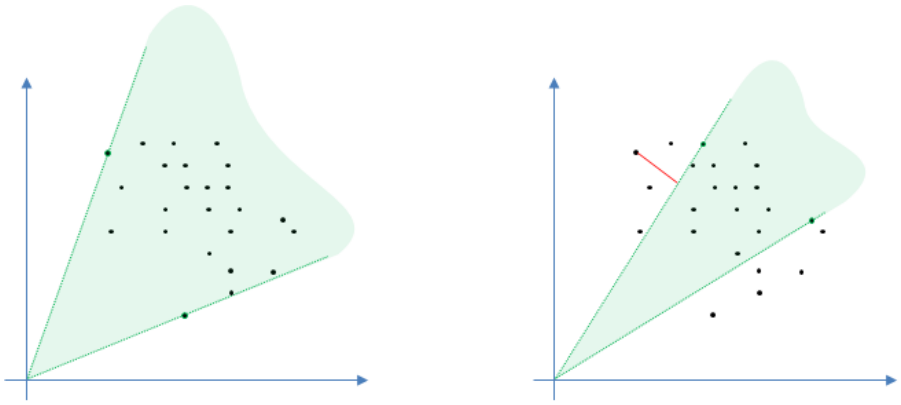

FIGURE 4.7 – *À gauche, les deux points sélectionnés se trouvent sur l'enveloppe du nuage de points : le volume du cône est maximal, est l'ensemble des points se trouve à l'intérieur de celui-ci. À l'inverse, à droite, le cône est supporté par deux points à l'intérieur du nuage : le volume est plus petit est les points à l'extérieur ne pourront pas être intégralement reconstruits suite à la projection (en rouge apparaît l'erreur de projection).*

possible d'atteindre des complexités linéaires, c'est-à-dire de définir un algorithme dont le coût calculatoire n'augmente pas beaucoup plus vite que la taille des données. Il s'agit notamment du cas de l'algorithme SAGA, que nous avons développé avec Nicolas Courty. La section suivante donne quelques précisions supplémentaires quant à la manière dont la sélection et la projection sont spécifiquement réalisées dans SAGA [J'10].

4.3 Les spécificités de SAGA

Column Subset Selection : Le fait que les colonnes de $\mathbf{M}$ s'expriment comme une combinaison linéaire non-négative des éléments du CSS correspond à une réalité géométrique : si les colonnes de $\mathbf{M}$ forment un nuage de points, alors, les éléments du CSS doivent appartenir à la bordure de ce nuage. Si c'est en effet le cas, la région qu'ils engendrent sur laquelle on va projeter durant la seconde étape (qu'il s'agisse d'un cône ou d'un simplex) contiendra la plupart des points, et la projection n'impliquera aucune perte d'information (voir Fig. 4.7-gauche). À l'inverse, si le CSS n'échantillonne pas correctement la bordure du nuage, certains points à l'extérieur seront mal reconstruits, et la factorisation induira une perte d'information (voir Fig. 4.7-droite). De manière assez intuitive, l'ensemble de points qui approxime le mieux la bordure du nuage de points (qu'on appelle aussi son enveloppe convexe) est celui délimitant le volume maximum parmi tous les ensembles de points possibles. C'est donc pour cela que les questions de factorisation de matrices rejoignent celles de l'analyse archétypale ou de l'échantillonnage, comme évoqué plus haut.

Par ailleurs, il existe un théorème ([175], Th. 1) qui stipule que si on augmente la taille du CSS, la qualité de la factorisation ne peut être qu'améliorée. À partir de là, il devient parfaitement justifié d'aborder la construction du CCS de manière itérative : à partir d'un ensemble de n éléments, choisir pour $(n+1)$-ième élément celui qui maximise l'augmentation du volume, et ce jusqu'à arriver à un CSS dont l'aug-

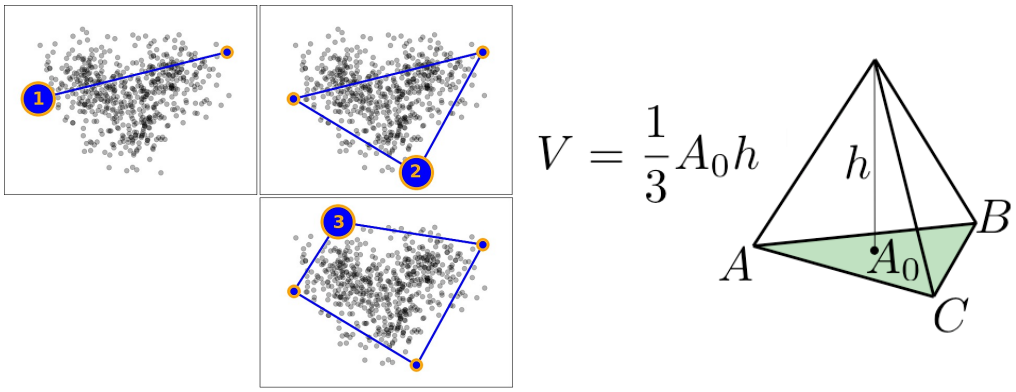

FIGURE 4.8 – *Gauche : parcours itératif d'un nuage de points afin de trouver des éléments permettant de décrire son enveloppe convexe. Droite : calcul du volume d'un simplex en dimension 3 (un tétraèdre) - tirés de [176, 177].*

mentation du volume est marginale (ou simplement insuffisante pour être justifiée), tel qu'illustré de manière schématique sur la Fig. 4.8-gauche.

Les méthodes de factorisation proposant de définir ainsi le CSS sont nombreuses. Dans notre cas, la particularité vient du fait que nous souhaitions opérer dans un espace de Hilbert à noyau reproduisant, afin de prendre en compte la géométrie de la variété contenant les données, comme expliqué au début de chapitre. Si cela complique un peu le travail (la définition du CSS ne peut être réalisée qu'en se basant sur des produits scalaires), cela présente aussi des avantages : notamment si le noyau est bien choisi, les données sont toujours projetées dans un espace de dimensionnalité égale à la taille du jeu de données. Dans un tel contexte, les éléments du CSS définiront un volume particulier, dénommé simplex, dont on peut facilement calculer le volume : en effet, un simplex est la généralisation à un espace de dimension n quelconque du segment, (dimensionnalité 1), du triangle (dimensionnalité 2), et du tétraèdre (dimensionnalité 3). De même que l'aire d'un triangle est donnée par la multiplication de la longueur d'un segment (la base du triangle) par la distance entre le 3$^{\text{ième}}$ point et le segment (la hauteur du triangle) divisé par la dimensionnalité (2), le volume du tétraèdre est donné par le produit de la surface du triangle de base par la hauteur, divisé par la dimensionnalité (3) ; et ainsi de suite (Fig. 4.8-droite). Dans le cas général, si δ^p représente un p-simplex et si Δ_i^{p+1} représente un $(p+1)$-simplex constitué de Δ^p et du point $\mathbf{x}_i$, alors on a la relation suivante entre leur volume :

$$\text{Vol}(\Delta_i^{p+1}) = \frac{\text{Vol}(\Delta^p) \times \text{hauteur}(\phi(\mathbf{x}_i), \Delta^p)}{p} \tag{4.1}$$

où $\phi(\mathbf{x}_i)$ est l'image de $\mathbf{x}_i$ dans l'espace à noyau reproduisant, et où hauteur(a,b) est une notation abusive pour désigner la distance entre un point a et son projeté orthogonal sur l'espace engendré par les éléments de b (ce qui correspond au segment h sur la Fig. 4.8-droite). En pratique, nous savons que [P'2] :

$$\text{hauteur}(\phi(\mathbf{x}_i), \Delta^p) = 1 - \left(\mathbf{k}_{\mathbf{x}_i}^\top \cdot \mathbf{K}_p^{-1} \cdot \mathbf{k}_{\mathbf{x}_i}\right) \tag{4.2}$$

où $\mathbf{K}_p^{-1}$ est la matrice de Gram des éléments du CSS, et où $\mathbf{k}_{\mathbf{x}_i}$ est tel que $\mathbf{k}_{\mathbf{x}_i} = [k(\mathbf{x}_j, \mathbf{x}_i)]_{\mathbf{x}_j \in \mathbf{X}\mathbf{W}^{(p)}}$. Dès lors, à partir d'un CSS de taille n formant un simplex, pour

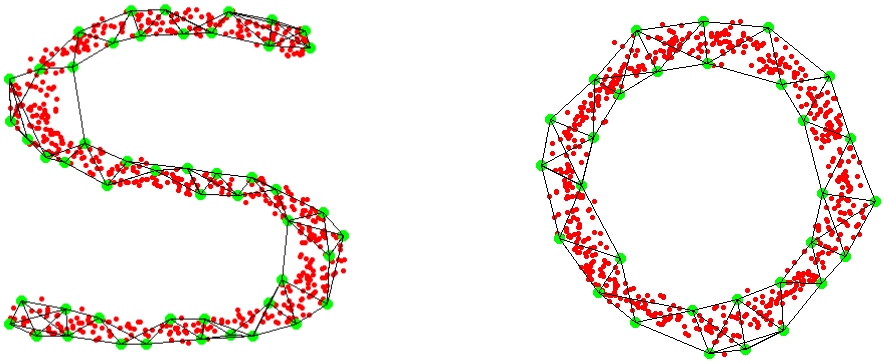

FIGURE 4.9 – *Illustration de la capacité de SAGA à extraire un CSS qui décrit l'enveloppe de la variété des données : les points rouges correspondent aux vecteurs colonnes de la matrice à factoriser, et les points verts à ceux faisant partie du CSS (les arêtes reliant les points verts n'ont été rajoutées que pour faciliter la visualisation) : ces derniers constituent clairement un sous-échantillonnage régulier des points définissant la bordure du "S" et de l'anneau, respectivement.*

trouver le simplex de taille $n+1$ qui soit de volume maximum, il suffit de trouver le point dont la distance aux simplex de taille n (sa hauteur en quelque sorte) est la plus grande, ou de manière équivalent le point minimisant $\mathbf{k}_{\mathbf{x}_i}^\top \cdot \mathbf{K}_p^{-1} \cdot \mathbf{k}_{\mathbf{x}_i}$ dans la formule ci-dessus. Autrement dit, la p-ième itération de la construction du CSS se fait par la sélection de la colonne i de la matrice $\mathbf{M}$ telle que :

$$i = \arg\max_q \mathrm{Vol}(\Delta_q^{p+1}) = \arg\min_q \left[\mathbf{k}_{\mathbf{x}_q}^\top \cdot \mathbf{K}_p^{-1} \cdot \mathbf{k}_{\mathbf{x}_q}\right] \tag{4.3}$$

Concrètement, il s'agit donc à chaque itération p, d'inverser une matrice $p \times p$, de sorte que si le CSS contient ℓ éléments, la complexité est d'ordre $\mathcal{O}(\ell^4)$. Heureusement, le CSS représente une faible portion des colonnes de la matrice de sorte que ℓ est petit devant le nombre de colonnes de $\mathbf{M}$. Par ailleurs, il est en pratique possible de mettre en place des stratégies calculatoires beaucoup moins lourdes, telle que celle dont les détails techniques sont résumées en Sec. 4.4.

Terminons en mentionnant que le fait de travailler dans un espace de Hilbert à noyau reproduisant a une autre conséquence, attendue (et même recherchée) : le simplex dans cet espace de Hilbert correspond, dans l'espace d'origine des données, non pas à l'enveloppe convexe de volume maximum, mais à l'enveloppe de la variété : elle caractérise précisément la géométrie du jeu de données. Cela est illustrée sur la Fig. 4.9, où des données simulées sont réparties sur des variétés simples (un anneau, un "S"), et où il apparaît clairement que le CSS ainsi sélectionné caractérise bien le contour de l'objet dessiné par la variété. En pratique, cela permet d'améliorer de manière considérable la qualité et la compacité de la redescription, puisque celle-ci "colle" à la topologie des données.

Projection : Alors que nos travaux initiaux sur SAGA nécessitaient systématiquement une projection sur un simplex, il apparait que pour le problème du démultiplexage de spectre, une projection sur un cône suffit. Or cette projection est plus

simple (car elle correspond plus souvent à une projection orthogonale) et peut être calculée avec la même méthode que celle mise en place pour SAGA (il s'agit d'une généralisation aux espaces de Hilbert à noyau reproduisant du *greedy selector and simplex projector*, proposé dans [178]). Concrètement, ce projecteur permet d'ajouter une contrainte de parcimonie durant la factorisation sans surcoût calculatoire : il suffit d'imposer une contrainte sur la dimensionnalité de l'espace de projection. Ainsi, au lieu de projeter sur le simplex ou le cône entier, de dimensionnalité ℓ, il est possible de projeter sur l'hyperface de dimension $\lambda < \ell$ fournissant la meilleure projection. Grâce à cela, l'algorithme SAGA (dont on peut maintenant détailler l'acronyme : *Sparse And Geometry Aware matrix factorisation*) reste d'une complexité linéaire par rapport à la taille des données.

Néanmoins, cette astuce permettant d'ajouter une contrainte de parcimonie sans surcoût calculatoire a un inconvénient : suivant le type de noyau utilisé, la projection sur un cône peut devenir difficile à calculer. En effet, si celui-ci déforme trop fortement l'espace d'entrée, les méthodes numériques utilisées pour calculer le projeté risquent de ne pas converger, notamment sous la contrainte de parcimonie, qui peut amener l'algorithme à changer de face sur laquelle projeter au fur et à mesure des itérations. Afin de garantir que la projection soit numériquement stable, il convient de se placer dans le cadre de la *Restricted Isometry Property* (ou RIP [179]) : cette propriété stipule que la déformation de l'espace d'origine induite par le noyau est suffisamment faible pour que l'on puisse considérer l'espace reproduit comme "localement linéaire", garantissant la convergence de la projection.

4.4 Implémentation du CSS

Dans cette section sont décrits les détails de la sélection incrémentale des ℓ vecteurs constituant le CSS. Ceux-ci n'ont d'intérêt que pour le lecteur cherchant une compréhension fine de ce qui permet une implémentation efficace du choix de le CSS. Les autres lecteurs peuvent se rendre directement à la Sec. 5.

Par suite de l'Eq. 4.3, la p-ième itération de la construction du CSS se fait par la sélection de la colonne i de la matrice $\mathbf{M}$ telle que :

$$i = \arg\min_q \left[\mathbf{k}_{\mathbf{x}_q}^\top \cdot \mathbf{K}_p^{-1} \cdot \mathbf{k}_{\mathbf{x}_q}\right] \tag{4.4}$$

Durant cette p-ième itération, il faut calculer la quantité $\left[\mathbf{k}_{\mathbf{x}_q}^\top \cdot \mathbf{K}_p^{-1} \cdot \mathbf{k}_{\mathbf{x}_q}\right]$ pour chaque colonne q parmi les $m - p$ colonnes de la matrice qui ne sont pas encore dans le CSS. La stratégie présentée ci-dessous permet de faire cela avec une complexité en $\mathcal{O}(p^2 + mp)$, de sorte que la complexité totale pour les ℓ itérations est en $\mathcal{O}(\ell^3 + m\ell^2)$. En pratique, comme un faible pourcentage des colonnes de la matrice fait partie du CSS, la complexité est en fait en $\mathcal{O}(m\ell^2)$, c'est à dire qu'elle est linéaire par rapport au nombre de colonnes de la matrice (dans notre cas, le nombre de valeur m/z correspondant à la gamme de masse du spectromètre, de l'ordre du million), et qu'elle est quadratique par rapport à la taille du CSS (le nombre de peptides dans l'échantillon, de l'ordre de 10000). Pour cela, la stratégie proposée par Thomas Fortin, membre du groupe KDPD, s'appuie sur 3 éléments :

– La décomposition de Cholesky de $\mathbf{K}_p$ en un produit $\mathbf{L}_p\mathbf{L}_p^\top$ où $\mathbf{L}_p$ est une matrice triangulaire inférieure ;

– Un schéma de mise à jour de rang un, qui permet d'obtenir $\mathbf{L}_{p+1}$ à partir de $\mathbf{L}_p$;
– L'évitement du calcul explicite de l'inverse du noyau $\mathbf{K}_p^{-1} = \mathbf{L}_p^{-\top}\mathbf{L}_p^{-1}$ en profitant du fait que :

$$\mathbf{k}_{\mathbf{x}_q}^\top \cdot \mathbf{K}_p^{-1} \cdot \mathbf{k}_{\mathbf{x}_q} = \left\langle \mathbf{L}_p^{-1} \cdot \mathbf{k}_{\mathbf{x}_q}, \mathbf{L}_p^{-1} \cdot \mathbf{k}_{\mathbf{x}_q} \right\rangle . \tag{4.5}$$

Supposons qu'à l'itération p, l'application de l'Eq. 4.4 nous amène à sélectionner la colonne $\mathbf{x}_i$, en s'appuyant entre autre sur les calculs intermédiaires donnant $\mathbf{k}_{\mathbf{x}_i}$, $\mathbf{L}_p$ et $\varepsilon_{\mathbf{x}_i} = \mathbf{k}_{\mathbf{x}_i}^\top \cdot \mathbf{K}_p^{-1} \cdot \mathbf{k}_{\mathbf{x}_i}$. Par définition,

$$\mathbf{K}_{p+1} = \begin{pmatrix} \mathbf{K}_p & \mathbf{k}_{\mathbf{x}_i} \\ \mathbf{k}_{\mathbf{x}_i}^\top & 1 \end{pmatrix} = \mathbf{L}_{p+1}\mathbf{L}_{p+1}^\top \tag{4.6}$$

avec :

$$\mathbf{L}_{p+1} = \begin{pmatrix} \mathbf{L}_p & \mathbf{0}_p \\ \mathbf{k}_{\mathbf{x}_i}^\top \mathbf{L}_p^{-\top} & \sqrt{1-\varepsilon_{\mathbf{x}_i}} \end{pmatrix} \tag{4.7}$$

(cf. les résultats sur les mises à jour de rang 1 d'une décomposition de Cholesky [180]). Il nous faut ensuite calculer $\mathbf{L}_{n+1}^{-1}$ dont la forme est :

$$\mathbf{L}_{p+1}^{-1} = \begin{pmatrix} \mathbf{L}_p^{-1} & \mathbf{0}_p \\ \mathbf{z}_p^\top & \zeta_p \end{pmatrix} \tag{4.8}$$

où $\mathbf{z}_p^\top$ et ζ_p sont respectivement un vecteur ligne et un scalaire, tous deux à préciser. Pour cela, nous savons que $\mathbf{L}_{p+1}\mathbf{L}_{p+1}^{-1} = \mathbf{I}_{p+1}$,dont la résolution nous amène à :

$$\mathbf{L}_{p+1}^{-1} = \begin{pmatrix} \mathbf{L}_p^{-1} & \mathbf{0}_n \\ (1-\varepsilon_{\mathbf{x}_i})^{-\frac{1}{2}} \mathbf{k}_{\mathbf{x}_i}^\top \mathbf{K}_n^{-1} & (1-\varepsilon_{\mathbf{x}_i})^{-\frac{1}{2}} \end{pmatrix} \tag{4.9}$$

qui peut se calculer en $\mathcal{O}(p^2)$ opérations. Enfin, il faut calculer $\left\langle \mathbf{L}_p^{-1} \cdot \mathbf{k}_{\mathbf{x}_q}, \mathbf{L}_p^{-1} \cdot \mathbf{k}_{\mathbf{x}_q} \right\rangle$ pour chaque colonne $\mathbf{x}_q$ restante (au nombre de $m-p-1$). Pour chacun d'eux, il suffit de mettre à jour le calcul fait à l'itération précédente en rajoutant le calcul de la dernière ligne, qui prend $\mathcal{O}(p)$ opérations. Cela nous amène bien à une complexité totale pour la p-ième itération de $\mathcal{O}(p^2 + mp)$.

5 Le projet Reveal-MS

5.1 Objectifs

L'écart entre le développement d'un outil de factorisation de matrice comme SAGA et son application à des problèmes concrets est en fait très variable. Dans les cas d'application où la séparation de sources est déjà bien entrée dans les mœurs, ou dans les cas où les données se traitent naturellement via le formalisme matriciel (comme par exemple les images numériques), le transfert au domaine d'application peut-être immédiat ; voire même naturel quand celui-ci fournit des *benchmarks* pour l'évaluation de l'outil. En revanche, dans les autres cas, le portage vers une application spécifique est aussi long et complexe que le développement proprement dit de l'outil. C'est notamment le cas pour le démultiplexage de spectres DIA. Il y a plusieurs raisons à cela :

- Les données manipulées sont de forme complexe : *(i)* elles ne sont naturellement pas faites pour être formatées en un chromatospectre ; *(ii)* les informations issues des deux types de spectres (MS1 et MS2) doivent pouvoir être reliées pour être pleinement exploitées ; *(iii)* les signaux récupérés en sortie des spectromètres ont déjà subi un certain nombre de traitements, constructeurs-dépendants et inconnus.
- L'interprétation et l'évaluation du démultiplexage ne peut se faire à la main, et nécessite l'intégration complète de SAGA dans le pipeline bioinformatique d'identification de spectres. Il est donc difficile d'isoler ce module pour en faciliter le test et l'amélioration.
- Comme déjà évoqué dans le cadre du projet DAPAR/ProStaR, il est difficile de produire des jeux de données qui soient suffisamment réalistes vis-à-vis de la complexité des informations qu'ils contiennent, tout en étant étiquetés par une vérité-terrain nous permettant de définir un *benchmark*. Cette difficulté est aussi présente dans le cas du démultiplexage de spectres ; et y pallier nécessite un important travail de la part de la plateforme du laboratoire.
- Enfin, il faut souligner l'aspect culturel : les utilisateurs et testeurs finaux ont, sur la base de leurs habitudes de travail et de leur formation académique, une idée de comment le développement de l'outil devrait être conduit ; et cette vision peut significativement diverger de ce que les *data scientists* mettent en place. C'est un point que je détaillerai un peu plus en conclusion de ce chapitre.

Finalement, pour toutes ces raisons, l'application de SAGA à la protéomique est un travail ambitieux et de longue haleine, aussi bien pour le groupe KDPD que pour le laboratoire EDyP. Nous avons néanmoins décidé de nous y confronter dans le cadre d'un projet dénommé *Reveal-MS*. Concrètement, l'objectif de ce projet est de mettre en place une brique logicielle venant s'interfacer entre la sortie du spectromètre et le moteur d'identification de peptides et qui permettra : (*i*) de récupérer en entrée les spectres acquis en mode DIA ; (*ii*) de fournir en sortie des pseudo-spectres DDA, c'est-à-dire pouvant être traités comme si le spectromètre avait réalisé l'acquisition en mode DDA ; et donc (*iii*) de garder inchangée l'intégralité du pipeline bioinformatique permettant leur traitement.

5.2 Un projet interdisciplinaire

Les compétences à mettre en œuvre afin de réaliser le projet Reveal-MS sont multiples : il y a évidemment un fort besoin en génie logiciel, afin de transformer le travail de recherche en un outil robuste, facilement installable dans un pipeline d'identification de peptides, et facilement utilisable par les ingénieurs protéomiciens travaillant sur une plateforme instrumentale. À ce propos, notons que les contraintes des utilisateurs sont relativement fortes, puisqu'en plus d'orienter le travail de développement logiciel, tel que les interfaces graphiques (de paramétrisation comme de visualisation), les contraintes de la plateforme impactent directement le projet : notamment, malgré la complexité linéaire de l'algorithme, les chromatospectres correspondant à des données réelles sont suffisamment grands pour nécessiter un temps de calcul important. Afin de limiter celui-ci, il est nécessaire de considérer dès le début du projet, des implémentations qui puissent être facilement parallélisées, mais aussi des schémas numériques pour lesquels la précision et la stabilité soient contrô-

lées. Ces questions sont déjà en cours d'étude par Thomas Fortin.

Le second volet de compétences disciplinaires concerne la spectrométrie de masse et la protéomique. En effet, il sera nécessaire : (*i*) de mettre en place des jeux de données *benchmarks*, afin tout d'abord, de pouvoir tester l'adéquation de SAGA avec des données de protéomiques, puis ensuite d'entrer dans un cycle itératif de tests et d'améliorations ; (*ii*) d'optimiser la partie instrumentale du pipeline (LC et MS/MS), afin de produire des données de la meilleure qualité possible, et ayant la meilleure adéquation possible avec les spécificités de SAGA : par exemple, il sera nécessaire de déterminer la largeur des fenêtre MS1 et l'alternance des cycles MS1 et MS2, mais aussi de vérifier dans quelle mesure ces choix ainsi que les paramètres de SAGA sont généralisables à plusieurs types d'instruments, ou au contraire, si nous devons nous concentrer sur un modèle d'instrument spécifique. Enfin, (*iii*) dans une optique finale de valorisation, il sera nécessaire de comparer Reveal-MS aux autres méthodes de l'état de l'art, et donc d'acquérir la maîtrise des outils correspondants, puis de définir des protocoles de comparaison adéquats.

Enfin, un certain nombre de questions de *data science* restent à aborder, notamment afin d'instancier ou de raffiner SAGA en vue d'être le plus en adéquation possible avec les données de protéomique. Cela inclut notamment, de nombreuses questions relevant de l'apprentissage automatique, dont voici une liste des plus importantes : tout d'abord, nous devons affiner la définition d'un noyau permettant de capturer la sémantique des chromatogrammes, afin que la recherche de ceux qui sont les plus représentatifs des espèces chimiques coéluantes corresponde bien à la recherche d'un simplex dans l'espace de Hilbert correspondant. Par ailleurs, cette définition du noyau doit se faire tout en garantissant le respect de la RIP énoncée plus haut. Ensuite, il sera nécessaire d'optimiser le calcul des projections, car en l'état actuel de SAGA, cette étape est soit trop lourde d'un point de vue calculatoire, soit trop instable : cependant, la définition d'une matrice de préconditionnement adaptée, ainsi qu'un choix plus fin de l'initialisation de l'algorithme de descente de gradient, devraient être suffisants pour résoudre le problème. Enfin, nous devrons mettre en place un certain nombre d'heuristiques, afin de quantifier le niveau de factorisation nécessaire (en fonction du contenu informationnel de la matrice des résidus, en fonction du nombre d'ions précurseurs attendus dans l'expérience, etc.) et afin d'associer la masse des ions précurseurs (résultant de la factorisation des MS1) aux spectres de fragmentations (résultant de la factorisation des MS2).

Enfin des questions à cheval entre *data science* et traitement des signaux de spectrométrie se posent : en effet, la réussite du projet est notamment liée à la possibilité de rendre les données compatibles avec deux hypothèses absolument nécessaires pour la bonne application de SAGA. Jusqu'à présent, je n'ai pas discuté ces deux hypothèses, de sorte qu'elles ont pu passer inaperçues. Dans la section suivante, je me focaliserai donc sur celles-ci, et je détaillerai les questions qu'elles soulèvent.

5.3 Deux hypothèses fortes

La première hypothèse, qui a été faite au moment de l'introduction du formalisme matriciel, est qu'il est possible (et facile) de passer d'un fichier contenant les sorties brutes du spectromètre à une structure matricielle, le chromatospectre. En pratique, ce n'est pas si simple. En effet, l'échantillonnage du signal, aussi bien en temps qu'en

m/z, n'a aucune raison d'être régulier : les pas de temps suivent les alternances de cycles MS1 et MS2, mais ceux-ci peuvent en plus être décalés par la physique du spectromètre. Ainsi, dans le cas d'un instrument possédant une trappe à ions, un remplissage un peu plus rapide de celle-ci (dû à une arrivée massive de précurseurs) peut déclencher une acquisition précoce. Quant à l'échantillonnage de la gamme m/z, c'est encore plus compliqué : le signal en sortie du spectromètre n'est pas du tout un signal brut, mais un signal qui a déjà été partiellement traité, nettoyé, et compressé. Ainsi, en plus du problème de l'irrégularité de l'échantillonnage en m/z, se pose celui du contrôle de l'amplitude des distorsions induites par le ré-échantillonnage d'un signal prétraité. En pratique, les outils classiques du traitement du signal sont à disposition, et permettent de résoudre ce problème de ré-échantillonnage, afin que le signal puisse être encodé sous la forme d'un chromatospectre : notamment dans le cas de jeux de données simples (peu de protéines, faible gamme dynamique, etc.) cela n'a pratiquement pas d'influence. En revanche, il est probable que plus le jeu de données est compliqué, plus un ré-échantillonnage sous-optimal fasse perdre à la fois de la sensibilité et de la résolution au démultiplexage. Il est donc indispensable que nous y restions attentifs durant l'avancement du projet.

Maintenant passons à la seconde hypothèse : pour bien comprendre celle-ci, revenons à l'interprétation des matrices résultant de SAGA. Comme illustré sur la Fig. 4.6, la première matrice est censée contenir les chromatogrammes des peptides demultiplexés. Plus précisément, pour chaque ion précurseur dont le spectre a été démultiplexé, il doit y avoir une colonne de la première matrice qui représente son chromatogramme. De par les règles du calcul matriciel, cette colonne ne doit contenir que le chromatogramme de cet ion, et aucun autre (dans le cas contraire, au moins deux ions n'auraient pas été démultiplexés). Cependant, par construction, cette première matrice est un CSS, c'est-à-dire un sous ensemble de colonnes extraites du chromatospectre. Or chaque colonne du chromatospectre est susceptible de contenir plusieurs pics d'élution, pour la raison évidente que des fragments de différents ions précurseurs éluant à différents instants peuvent avoir exactement la même masse (comme par exemple illustré sur la Fig. 4.5). Un élément essentiel de SAGA est donc de réussir à extraire suffisamment de chromatogrammes simples de la jungle des chromatogrammes multiples contenus dans le chromatospectre pour définir le CSS. Cependant, ce n'est pas toujours possible : considérons un ion précurseur donné, et supposons que pour tous ses fragments, qui correspondent à une valeur m/z précise chacun, il y a un autre ion précurseur ayant un fragment avec exactement la même valeur m/z. En conséquence directe, il sera impossible de récupérer une colonne du chromatospectre qui corresponde au chromatogramme de cet ion précurseur, et seulement à lui. Et, par conséquent, il sera impossible de le démultiplexer. Cela nous amène à formuler la seconde hypothèse de la manière suivante : pour chaque ion précurseur, il existe au moins un fragment dont le ratio m/z ne correspond à aucun autre fragment d'aucun autre ion précurseur. Tout comme la première hypothèse, la seconde ne pose pas vraiment de difficulté sur des jeux de données contrôlés, dont la complexité est intrinsèquement limitée : dans de tels jeux, il y a peu d'ions précurseurs, donc un nombre total de raies spectrales (MS2) plus faible ; et d'autant moins de chances de ne pas pouvoir trouver de tels chromatogrammes "simples". À l'inverse, sur de vrais jeux de données provenant d'échantillons biologiques complexes, cette hypothèse sera moins facilement respec-

tée, ce qui pourra potentiellement limiter la profondeur d'analyse. Cependant, nous savons dès maintenant que sur de tels jeux de données, il sera possible de décomposer des chromatogrammes "complexes", en une somme de chromatogrammes "simples", en utilisant les outils classiques du traitement de signal [24]. Encore une fois donc, ces outils nous fournissent par avance une solution à un problème qui ne se présentera que quand la complexité des jeux de données étudiés augmentera, afin de se rapprocher des jeux de données réalistes.

5.4 Un cadre pour un travail doctoral

Olga Permiakova débute sa thèse au laboratoire sous ma direction au moment-même de la finalisation de ce rapport, et son sujet de thèse s'inscrit dans la droite ligne du projet Reveal-MS. Cependant, comme expliqué au Chap. 2, Sec. 5, je tiens à ce qu'elle ait ses objectifs propres, indépendamment des contraintes du projet, afin de lui éviter le piège d'un travail d'ingénierie visant principalement à accélérer la production du logiciel Reveal-MS. Par ailleurs, il faut tenir compte du fait que Reveal-MS a déjà commencé, et qu'il peut être difficile pour un jeune doctorant de s'approprier un sujet déjà partiellement défriché.

Afin d'éviter ces deux écueils, il me semble intéressant de proposer à Olga d'étudier l'adéquation entre les deux hypothèses décrites ci-dessus, et les données de protéomique produites en pratique par un protocole DIA. Bien que n'ouvrant pas tout de suite vers des problématiques originales (comme je l'ai dit, les outils classiques du traitement du signal devraient fournir des solutions au moins partiellement satisfaisantes), cela peut constituer une entrée en matière intéressante : tout d'abord, parce que cela permettra à Olga de se familiariser à la fois avec les données et les aspects plus théoriques du projet. Ensuite parce que cela permettra un positionnement précoce sur une question fortement interdisciplinaire. Enfin, parce qu'il s'agira d'une excellente préparation à des questions plus ouvertes, notamment concernant la super-résolution des signaux de spectrométrie de masse.

Le principe de la super-résolution est de reconstruire les détails d'un signal à une résolution inférieure à celle de l'acquisition (par exemple, être capable d'inverser un flou gaussien, par analogie avec ce qui a été discuté au Chap. 1, Sec. 4). Cela peut sembler irréaliste, puisque cela nécessiterait de créer *ex nihilo* de l'information. Cependant, il existe des situations dans lesquelles cela est possible : par exemple tout humain est capable de reconnaître l'identité d'une personne sur une photo qui a subi un flou gaussien, ou encore de deviner la position des yeux sur celle-ci même si le sujet porte des lunettes de soleil. En pratique, nous connaissons tellement bien la structure du signal étudié que nous disposons *a priori* d'un contexte (au sens de la pyramide DIC de la Fig. 1.1) sur lequel nous appuyer. Plus prosaïquement, la super-résolution est initialement apparue avec le *Robust Uncertainty Principle* [181], qui stipule que la plupart des signaux réels (naturels ou artificiels), même s'ils ne sont pas suffisamment réguliers pour permettre une reconstruction de l'information manquante, sont tels que leur représentation dans l'espace de Fourier l'est. Ainsi, la reconstruction d'un signal tronqué peut-être réalisée par la recherche du signal le plus régulier possible dans l'espace de Fourier, dont certaines parties concordent avec celles observées. Dire que le signal (ou sa représentation de Fourier) est régulier est équivalent à dire que sa dérivée est souvent nulle ; et donc qu'il est possible

d'en fournir une représentation parcimonieuse (avec peu de coefficients non-nuls, tels que décrit en Chap. 1, Sec. 5.2). Il n'est donc pas étonnant que les méthodes de réduction de dimensionnalité et d'approximation de faible rang (dont la factorisation de matrice) soient intrinsèquement liées à la super-résolution [182–184].

Finalement, les liens entre SAGA, la super-résolution et l'amélioration des signaux de spectrométrie de masse, semblent constituer un cadre intéressant vers lequel orienter les travaux d'Olga. Cependant, en fonction de ses appétences, d'autres sujets tout aussi intéressants et prenant racine sur le projet Reveal-MS peuvent être explorés : par exemple, comment valoriser l'expertise que nous sommes en train d'acquérir sur les noyaux caractérisant les chromatogrammes pour proposer de nouvelles méthodes d'alignement des temps de rétentions des peptides (qui par nature, sont très difficilement reproductibles) ? Il serait tout aussi intéressant d'exploiter les raffinements possible de SAGA, afin de traiter directement plusieurs réplicats analytiques (et donc réaliser une factorisation jointe), ou afin d'augmenter la profondeur d'analyse (en mettant en place des stratégies basées sur des factorisations itératives, voire multi-échelles). Tous ces sujets me semblent également intéressants, et Olga fera le choix de l'un d'entre eux au fil des premiers mois de thèse.

6 Conclusions : recherche de financements

À l'heure actuelle, je suis en recherche de financements pour supporter la réalisation du travail restant sur le projet Reveal-MS, à la frontière entre traitement du signal, spectrométrie de masse et génie logiciel. Notamment, j'ai échoué à obtenir un financement "Jeunes Chercheurs" de l'ANR. Compte-tenu du taux de succès particulièrement bas de ce type d'appels à projets, je ne prends pas cet échec comme une remise en cause de la pertinence scientifique de ma proposition ; notamment parce qu'aux deux étapes de sélection de l'ANR, les rapporteurs ont soulignés l'intérêt du projet pour la protéomique, son originalité vis-à-vis de l'état de l'art, la pertinence des compétences interdisciplinaires du groupe et du laboratoire pour le mener, et son positionnement stratégique vis-à-vis des évolutions probables de la discipline.

Malgré tout cela, je n'ai pas réussi à convaincre le jury sur d'autres éléments, que je perçois comme secondaires, mais qui, de leur point de vue, sont au contraire déterminants : objectifs de recherche perçus comme très (trop ?) théoriques, manque de précision sur la définition de certains jeux de données d'évaluation (à produire au laboratoire ou issus de la littérature), manque d'un projet d'application directe dans le domaine biomédical, et résultats préliminaires et/ou intermédiaires pas assez incrémentaux. De mon point de vue, ces arguments sont justes, mais je pense que certains n'impactent que peu le projet, alors que d'autres au contraire, vont à l'encontre de ce qui permettrait sa réussite. Cela m'amène à penser que je n'ai pas vraiment le même bagage scientifique que le jury, qui semble principalement issu de la culture "biologie, chimie, protéomique", où la recherche est "de découverte" plutôt que "méthodologique" ; où l'on approche la découverte par une accumulation itérative de preuves complémentaires permettant des résultats incrémentaux ; où l'application médicale est une justification importante du travail ; et où le choix du jeu de données ainsi que de sa préparation sont d'importance capitale. Tous ces éléments ne sont au contraire pas constitutifs de la culture de la *data science*.

Néanmoins, pour rester honnête, j'aurais pu prévoir la culture du jury, puisque j'ai soumis ce projet dans le défi intitulé "Vie, santé et bien-être". Cependant, quelles auraient été les chances d'un tel projet d'être accepté dans un autre défi ? "Société de l'information et de la communication" ? À la place du rapporteur, je pourrais probablement moi-même écrire : "*Le travail théorique correspondant à SAGA est déjà réalisé. Le reste n'est qu'application, et ne participe pas à l'amélioration de l'état de l'art dans un quelconque domaine de la* data science*. Il convient donc de faire financer ce projet par les domaines d'application qui vont en bénéficier*". Défi "Des autres savoirs" ? Ce défi me semble surtout permettre le financement de recherches fondamentales en astrophysique, physique des particules, mathématique, et physico-chimie, de sorte que je crains d'y faire un réel hors sujet... Mais c'est ce qu'il me reste à tenter cette année.

Finalement, cela peut donner l'impression que l'ANR n'a pas vocation à financer prioritairement des travaux du type de Reveal-MS, et qu'il est prévu que cette tâche soit remplie par d'autres instances, comme la Mission pour l'Interdisciplinarité du CNRS. Cependant, celle-ci a des capacités moindres : ainsi, il y a quelques années, nous avons pu bénéficier d'un financement pour le projet Prospectom, mais plus faible d'un ordre de grandeur. Tout cela me laisse un peu perplexe : dois-je modifier mon projet en profondeur, quitte à aller dans une direction qui me semble moins prometteuse en termes de résultats méthodologiques ? Certes, il est toujours plus facile de critiquer les institutions que de remettre en cause ses choix scientifiques, et j'espère éviter ce travers facile. Malgré tout, j'ai l'impression que chez nous, le travail interdisciplinaire semble plus difficile que dans la plupart des autres pays européens, anglo-saxons, ou au Japon. Je ne prétends pas connaître tous ces pays, mais à ce que j'ai pu juger par mes quelques collaborations internationales, j'ai l'impression que dans plusieurs d'entre eux, les cloisonnements entre recherche et ingénierie, entre recherche appliquée et recherche fondamentale, mais aussi entre les différents domaines disciplinaires, ne sont pas aussi importants que chez nous. Ce constat est à mettre en parallèle avec un autre : sous l'influence des standards internationaux, notre modèle de fonctionnement historique est en train d'évoluer très rapidement. D'une organisation très "collective" structurée par des équipes de tailles importantes dans lesquels de nombreux chercheurs titulaires travaillaient ensemble, nous nous orientons vers une organisation plus individuelle, centrée sur des *principal investigators* dont l'objectif est de ramener des financements en leur nom propre, afin de monter une équipe de taille réduite autour d'eux. De manière naïve, j'ai tendance à penser qu'un fonctionnement individuel est moins apte à permettre l'émergence de travaux fortement interdisciplinaires. Dès lors, il ne faudrait pas prendre le risque de garder de chacun des modèles, le moins favorable à l'interdisciplinarité, et ainsi aboutir à un environnement particulièrement cloisonné disciplinairement, dans lequel vivraient de petites équipes.

Malgré tout, l'environnement dans lequel je travaille actuellement compense en partie cela : il s'agit d'une équipe de grande taille, dotée d'un gros support en ingénierie (car jumelée à une plateforme de service), et dans laquelle de nombreuses disciplines collaborent. Je prends cela comme une réelle chance de pouvoir continuer le projet Reveal-MS malgré les difficultés de financement.

Conclusion générale

Ce document n'a pas la forme d'un rapport d'HDR moderne, dit "sur articles" (c'est-à-dire constitué d'une compilation des principales publications du candidat, chapeautée d'une introduction générale). Un tel format a l'avantage de la concision, tout en permettant d'embrasser de manière globale la problématique de recherche principalement abordée. Cependant, je pense qu'il n'est pas adapté à mon travail pour deux raisons : tout d'abord, mes travaux ne s'articulent pas autour d'une problématique, mais plutôt autour d'un champ d'applications. Ensuite, une compilation de résultats ne permet pas d'aborder ce qui dans mon cas, fait la difficulté de la direction de recherche. J'ai donc fait le choix d'une forme différente, laissant moins de place à ce qui fait le contenu des articles de recherche, pour me concentrer sur d'autres éléments que je trouve fondamentaux : l'animation d'équipe, l'équilibre subtil imposé à tout travail interdisciplinaire entre recherche et supervision d'activité d'ingénierie, et enfin, la dimension pédagogique indispensable à la fertilisation mutuelle de plusieurs disciplines. Ce choix m'a amené à rédiger un document assez long, et je remercie les rapporteurs ayant eu la lourde tâche de lire l'intégralité ma prose. Dans l'ensemble, j'espère avoir atteint les objectifs suivants :

- Donner, grâce au Chap. 1 un aperçu aussi large et simple que possible de la science des données à un protéomicien, qu'il soit chimiste ou biologiste de formation, qu'il soit ingénieur de recherche ou chercheur ;
- Donner, grâce au début du Chap. 2, l'envie à de nombreux *data scientists* de s'intéresser aux problèmes de la protéomique, et plus généralement aux "*omics*" un peu plus confidentielles que la génomique et la transcriptomique : ces domaines contiennent une multitude de problèmes intéressants, et compte-tenu des besoins, trop peu de *data scientists* s'y intéressent pour l'instant ;
- Montrer grâce à la fin du Chap. 2, le travail d'animation d'équipe et d'encadrement d'activité de recherche que j'ai mis en œuvre ces dernières années, conformément aux attentes de l'HDR ;
- Montrer que des problèmes de recherche intéressants peuvent émerger d'une activité de supervision d'ingénierie (Chap. 3) et que réciproquement, même une recherche méthodologique innovante par elle-même doit se rapprocher de l'ingénierie (Chap. 4) afin de permettre une véritable interdisciplinarité, et donc une application à court terme pour de la recherche de découverte.

Par ailleurs, j'ai essayé de fournir, tout au long de ce document, suffisamment de pistes de recherche et de questions ouvertes pour garantir que la *to-do list* du groupe KDPD est déjà intégralement remplie pour les 5 années à venir ; réduisant d'autant le risque pour le contribuable de nous voir désœuvrés. Voici les principaux éléments de cette liste :

Outils de Visualisation : dans la continuité de ce qui a été exploré durant le stage de Shivani (cf. Chap. 3, Sec. 2.1), j'aimerais compléter les outils d'extraction automatique de connaissances par des outils permettant au protéomicien de visualiser, comprendre et interagir avec ses données.

Alignement de graphes : il s'agit, de manière complémentaire au point précédent, de permettre la mise en correspondance automatique de graphes peptides-protéines, afin de faciliter le diagnostic et la correction d'erreurs d'identification.

Agrégation de p-valeurs : le besoin du protéomicien de contrôler les fausses découvertes au niveau des protéines (entités ayant le plus de sens d'un point de vue biologique) n'est à l'heure actuelle rempli par aucune méthode statistiquement valide (et robuste). Plutôt que d'essayer de définir de nouveaux FDR, comme cela est classiquement fait en protéomique (cf. Chap. 2, Sec. 2.4), de manière un peu "artisanale", je pense qu'il serait plus pertinent de réfléchir à des moyens de définir les p-valeurs au niveau protéique, afin d'y appliquer ensuite les outils de contrôle classiques.

Prise en compte des données temporelles : comme cela est décrit au Chap. 3, Sec. 3.1, il est indispensable d'étendre les fonctionnalités de ProStaR afin de pouvoir prendre en charge de manière routinière des données de protéomiques où la quantification relative s'effectue entre différents instants d'un processus biologique.

TAP-TAG : de même que le point précédent, il est nécessaire d'étendre le schéma de comparaisons relatives à des "comparaisons de comparaisons", comme demandé par les expériences de *Tandem Affinity Purification* (cf. Chap. 3, Sec. 3.1).

Bioanalyse (cf. Chap. 3, Sec. 3.1) : enrichir ProStaR de fonctionnalités supplémentaires permettant de remettre les protéines sélectionnées comme différentiellement abondantes dans un contexte biologique pertinent.

Faire le lien avec mes précédents travaux sur les fonctions de croyance : plusieurs sujets semblent prometteurs (cf. Chap. 3, Sec. 3.2) ; combinaison de scores issus de plusieurs moteurs d'identification de peptides ; réinterprétation de la p-valeur en termes de plausibilité ; tests statistiques sur des données imprécises. Au-delà de ces quelques exemples, il y a probablement d'autres pistes à explorer.

Financer et terminer Reveal-MS : comme expliqué à la page précédente.

Alignement de temps de rétention : profiter du travail réalisé sur les chromatogrammes dans le cadre de Reveal-MS pour améliorer les méthodes d'alignement de chromatogrammes.

Super-résolution : exploiter le lien entre super-résolution et approximations matricielles de faible rang, afin de profiter des avancées du projet Reveal-MS pour améliorer le traitement bas niveau des signaux spectrométriques.

À la réflexion, il y a même trop de travail pour le petit groupe que nous sommes. J'invite donc toute personne qui le souhaite à venir s'emparer d'un de ces sujets, afin de l'aborder seule ou en collaboration avec nous !

Références

[1] URL : http://www.academie-sciences.fr/fr/Colloques-conferences-et-debats-par-et-pour-la-communaute-scientifique/la-datamasse-directions-et-enjeux-pour-les-donnees-massives.html (cf. p. viii).

[2] Stigler, S. M. *The history of statistics : The measurement of uncertainty before 1900*. Harvard University Press, 1986 (cf. p. 5).

[3] URL : https://fr.wikipedia.org/wiki/Noam_Chomsky (cf. p. 5).

[4] Rigo, M. *notes du cours de Théorie des automates et langages formels*. Université de Liège, 2009. URL : http://www.discmath.ulg.ac.be/cours/main_autom.pdf (cf. p. 5).

[5] URL : https://fr.wikipedia.org/wiki/Alan_Turing (cf. p. 6).

[6] URL : https://fr.wikipedia.org/wiki/Claude_Shannon (cf. p. 6).

[7] Zins, C. "Conceptual approaches for defining data, information, and knowledge". In : *Journal of the American society for information science and technology* 58.4 (2007), p. 479–493. DOI : 10.1002/asi.20508 (cf. p. 6).

[8] Saporta, G. *Probabilités, analyse des données et statistique*. Editions Technip, 2006 (cf. p. 8).

[9] URL : http://www.cs.cornell.edu/courses/cs4780/2015fa/web/lecturenotes/lecturenote12.html (cf. p. 11).

[10] Perkins, D. N. Pappin, D. J. C. Creasy, D. M. Cottrell, J. S. "Probability-based protein identification by searching sequence databases using mass spectrometry data". In : *electrophoresis* 20.18 (1999), p. 3551–3567. DOI : 10.1002/(SICI)1522-2683(19991201)20:18<3551::AID-ELPS3551>3.0.CO;2-2 (cf. p. 11, 62).

[11] Efron, B. *Local false discovery rates*. Techincal Report, Division of Biostatistics, Stanford University, 2005. URL : https://statistics.stanford.edu/sites/default/files/BIO%20234.pdf (cf. p. 12).

[12] URL : https://fr.wikipedia.org/wiki/Test_(statistique) (cf. p. 14).

[13] Cohen, J. "The Earth is Round (p<. 05)". In : *American Psychologist* (1994), p. 997–1003. URL : http://ist-socrates.berkeley.edu/~maccoun/PP279_Cohen1.pdf (cf. p. 14, 80).

[14] Gigerenzer, G. "Mindless statistics". In : *The Journal of Socio-Economics* 33.5 (2004), p. 587–606. DOI : 10.1016/j.socec.2004.09.033 (cf. p. 14, 80).

[15] Trafimow, D. Marks, M. "Editorial". In : *Basic and Applied Social Psychology* 37.1 (2015), p. 1–2. DOI : 10.1080/01973533.2015.1012991 (cf. p. 14, 80).

[16] Goodman, S. N. "Aligning statistical and scientific reasoning". In : *Science* 352.6290 (2016), p. 1180–1181. DOI : 10.1126/science.aaf5406 (cf. p. 14, 80).

[17] Efron, B. *Large-scale inference : empirical Bayes methods for estimation, testing, and prediction*. T. 1. Cambridge University Press, 2010. URL : http://statweb.stanford.edu/~ckirby/brad/LSI/monograph_CUP.pdf (cf. p. 14, 79).

[18] Tibshirani, R. "Regression shrinkage and selection via the lasso". In : *Journal of the Royal Statistical Society. Series B (Methodological)* (1996), p. 267–288. DOI : 10.1111/j.1467-9868.2011.00771.x (cf. p. 14, 33).

[19] Rumelhart, D. E. Hinton, G. E. Williams, R. J. *Learning internal representations by error propagation*. Rapp. tech. DTIC Document, 1985. URL : http://deeplearning.cs.cmu.edu/pdfs/Chap8_PDP86.pdf (cf. p. 14).

[20] Dorigo, M. Maniezzo, V. Colorni, A. "Ant system : optimization by a colony of cooperating agents". In : *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)* 26.1 (1996), p. 29–41. DOI : 10.1109/3477.484436 (cf. p. 15).

[21] Zadeh, L. "Fuzzy Sets". In : *Information and Control* 8 (1965), p. 338–353. DOI : 10.1016/S0019-9958(65)90241-X (cf. p. 15).

[22] Anscombe, F. J. Aumann, R. J. "A definition of subjective probability". In : *The annals of mathematical statistics* 34.1 (1963), p. 199–205. DOI : 10.1214/aoms/1177704255 (cf. p. 15).

[23] Cornuéjols, A. Miclet, L. *Apprentissage artificiel : concepts et algorithmes*. Editions Eyrolles, 2011 (cf. p. 15).

[24] Mallat, S. *A wavelet tour of signal processing*. Academic press, 1999. DOI : 10.1016/B978-0-12-374370-1.00001-X (cf. p. 15, 19, 20, 112).

[25] Dreo, J. Petrowski, A. Siarry, P. Taillard, E. *Metaheuristics for hard optimization : methods and case studies*. 2006. DOI : 10.1007/3-540-30966-7 (cf. p. 15).

[26] Dubois, D. Prade, H. *Possibility theory : an approach to computerized processing of uncertainty*. Springer Science & Business Media, 2012. DOI : 10.1007/978-1-4684-5287-7 (cf. p. 15).

[27] Jensen, F. V. *An introduction to Bayesian networks*. T. 210. UCL press London, 1996 (cf. p. 16).

[28] Pearl, J. *Causality*. 2009 (cf. p. 16).

[29] URL : https://fr.wikipedia.org/wiki/Paradoxe_de_Hempel (cf. p. 16).

[30] Shafer, G. R. Shenoy, P. P. "Local computation in hypertrees". In : (1991). URL : https://kuscholarworks.ku.edu/bitstream/handle/1808/143/WP201.pdf?sequence=1&isAllowed=y (cf. p. 16).

[31] Aji, S. M. McEliece, R. J. "The generalized distributive law". In : *IEEE transactions on Information Theory* 46.2 (2000), p. 325–343. DOI : 10.1109/18.825794 (cf. p. 16).

[32] Kschischang, F. R. Frey, B. J. Loeliger, H.-A. "Factor graphs and the sum-product algorithm". In : *IEEE Transactions on information theory* 47.2 (2001), p. 498–519. DOI : 10.1109/18.910572 (cf. p. 16).

[33] URL : http://math.stackexchange.com/questions/29204/inference-bayes-network (cf. p. 17).

[34] Rabiner, L. R. "A tutorial on hidden Markov models and selected applications in speech recognition". In : *Proceedings of the IEEE* 77.2 (1989), p. 257–286. DOI : 10.1109/5.18626 (cf. p. 17).

[35] Dempster, A. P. "Upper and lower probabilities induced by a multivalued mapping". In : *The annals of mathematical statistics* (1967), p. 325–339. DOI : 10.1007/978-3-540-44792-4_3 (cf. p. 17).

[36] Dempster, A. P. "A generalization of Bayesian inference". In : *Journal of the Royal Statistical Society. Series B (Methodological)* (1968), p. 205–247. DOI : 10.1007/978-3-540-44792-4_4 (cf. p. 17).

[37] Shafer, G. *A mathematical theory of evidence*. T. 1. Princeton university press Princeton, 1976 (cf. p. 17).

[38] URL : http://fr.wikipedia.org/wiki/Ondelette_chapeau_mexicain (cf. p. 22).

[39] URL : http://en.wikipedia.org/wiki/Haar_wavelet (cf. p. 22).

[40] Bennett, C. H. Gács, P. Li, M. Vitányi, P. M. Zurek, W. H. "Information distance". In : *IEEE Transactions on information theory* 44.4 (1998), p. 1407–1423. DOI : 10.1109/18.681318 (cf. p. 22).

[41] Donoho, D. L. "Compressed sensing". In : *IEEE Transactions on information theory* 52.4 (2006), p. 1289–1306. DOI : 10.1109/TIT.2006.871582 (cf. p. 22, 103).

[42] Candes, E. J. "The restricted isometry property and its implications for compressed sensing". In : *Comptes Rendus Mathematique* 346.9 (2008), p. 589–592. DOI : doi:10.1016/j.crma.2008.03.014 (cf. p. 22).

[43] Mairal, J. Bach, F. Ponce, J. Sapiro, G. "Online learning for matrix factorization and sparse coding". In : *Journal of Machine Learning Research* 11.Jan (2010), p. 19–60. URL : http://www.jmlr.org/papers/volume11/mairal10a/mairal10a.pdf (cf. p. 22).

[44] Wright, J. Ma, Y. Mairal, J. Sapiro, G. Huang, T. S. Yan, S. "Sparse representation for computer vision and pattern recognition". In : *Proceedings of the IEEE* 98.6 (2010), p. 1031–1044. DOI : 10.1109/JPROC.2010.2044470 (cf. p. 22, 98, 103).

[45] URL : http://en.wikipedia.org/wiki/JPEG_2000 (cf. p. 23).

[46] Schölkopf, B. Smola, A. J. *Learning with kernels : Support vector machines, regularization, optimization, and beyond*. MIT press, 2002 (cf. p. 25, 26, 29).

[47] URL : https://www.youtube.com/watch?v=3liCbRZPrZA (cf. p. 26).

[48] Miclet, L. Cornuéjols, A. "What is the place of Machine Learning between Pattern Recognition and Optimization ?" In : *TML 2008 Conference :(Teaching Machine Learning)*. 2008. URL : http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.361.6285&rep=rep1&type=pdf (cf. p. 27).

[49] Vapnik, V. N. *Statistical learning theory*. T. 1. Wiley New York, 1998 (cf. p. 27, 31).

[50] URL : https://fr.wikipedia.org/wiki/Jean-Paul_Benz%C3%A9cri (cf. p. 27).

[51] Beyer, K. Goldstein, J. Ramakrishnan, R. Shaft, U. "When is "nearest neighbor" meaningful ?" In : *International conference on database theory*. Springer. 1999, p. 217–235. DOI : 10.1007/3-540-49257-7_15 (cf. p. 29).

[52] Francois, D. Wertz, V. Verleysen, M. "The concentration of fractional distances". In : *IEEE Transactions on Knowledge and Data Engineering* 19.7 (2007), p. 873–886. DOI : 10.1109/TKDE.2007.1037 (cf. p. 29).

[53] Saul, L. K. Roweis, S. T. "Think globally, fit locally : unsupervised learning of low dimensional manifolds". In : *The Journal of Machine Learning Research* 4 (2003), p. 119–155. URL : http://www.jmlr.org/papers/volume4/saul03a/saul03a.pdf (cf. p. 30).

[54] Lafferty, J. Lebanon, G. "Diffusion kernels on statistical manifolds". In : *Journal of Machine Learning Research* 6.Jan (2005), p. 129–163. URL : http://www.jmlr.org/papers/volume6/lafferty05a/lafferty05a.pdf (cf. p. 30, 92).

[55] Lee, J. A. Verleysen, M. *Nonlinear dimensionality reduction*. 2007. DOI : 10.1007/978-0-387-39351-3 (cf. p. 30).

[56] Coifman, R. R. Lafon, S. Lee, A. B. Maggioni, M. Nadler, B. Warner, F. Zucker, S. W. "Geometric diffusions as a tool for harmonic analysis and structure definition of data : Diffusion maps". In : *Proceedings of the National Academy of Sciences of the United States of America* 102.21 (2005), p. 7426–7431. DOI : 10.1073/pnas.0500334102 (cf. p. 31).

[57] Hofstadter, D. "Gödel, Escher, Bach, Les brins d'une Guirlande éternelle". In : *InterÉditions, Paris* (1985) (cf. p. 34).

[58] URL : http://en.wikipedia.org/wiki/Protein (cf. p. 36).

[59] URL : http://fr.wikipedia.org/wiki/Acide_amin%C3%A9 (cf. p. 36).

[60] URL : http://pubs.niaaa.nih.gov/publications/aa86/aa86.htm (cf. p. 39).

[61] URL : http://www.genomicglossaries.com/content/omes.asp (cf. p. 41).

[62] URL : https://en.wikipedia.org/wiki/List_of_omics_topics_in_biology (cf. p. 41).

[63] Ginsberg, J. Mohebbi, M. H. Patel, R. S. Brammer, L. Smolinski, M. S. Brilliant, L. "Detecting influenza epidemics using search engine query data". In : *Nature* 457.7232 (2008), p. 1012–1014. DOI : 10.1038/nature07634 (cf. p. 42).

[64] URL : http://www.google.org/flutrends/ (cf. p. 42).

[65] URL : http://www.creative-proteomics.com/Services/Peptide-mass-fingerprinting-PMF.htm (cf. p. 45).

[66] URL : http://fr.wikipedia.org/wiki/S%C3%A9quen%C3%A7age_par_spectrom%C3%A9trie_de_masse (cf. p. 46).

[67] URL : http://en.wikipedia.org/wiki/Protein_mass_spectrometry (cf. p. 46).

[68] Nesvizhskii, A. I. "A survey of computational methods and error rate estimation procedures for peptide and protein identification in shotgun proteomics". In : *Journal of proteomics* 73.11 (2010), p. 2092–2123. DOI : 10.1016/j.jprot.2010.08.009 (cf. p. 47, 49).

[69] Johnson, K. J. Wright, B. W. Jarman, K. H. Synovec, R. E. "High-speed peak matching algorithm for retention time alignment of gas chromatographic data for chemometric analysis". In : *Journal of Chromatography A* 996.1 (2003), p. 141–155. DOI : 10.1016/S0021-9673(03)00616-2 (cf. p. 48).

[70] Bantscheff, M. Schirle, M. Sweetman, G. Rick, J. Kuster, B. "Quantitative mass spectrometry in proteomics : a critical review". In : *Analytical and bioanalytical chemistry* 389.4 (2007), p. 1017–1031. DOI : 10.1007/s00216-007-1486-6 (cf. p. 48).

[71] Bantscheff, M. Lemeer, S. Savitski, M. M. Kuster, B. "Quantitative mass spectrometry in proteomics : critical review update from 2007 to the present". In : *Analytical and bioanalytical chemistry* 404.4 (2012), p. 939–965. DOI : 10.1007/s00216-012-6203-4 (cf. p. 48).

[72] Tran, J. C. Zamdborg, L. Ahlf, D. R. Lee, J. E. Catherman, A. D. Durbin, K. R. Tipton, J. D. Vellaichamy, A. Kellie, J. F. Li, M. "Mapping intact protein isoforms in discovery mode using top-down proteomics". In : *Nature* 480.7376 (2011), p. 254–258. DOI : 10.1038/nature10575 (cf. p. 49).

[73] Chapman, J. D. Goodlett, D. R. Masselon, C. D. "Multiplexed and data-independent tandem mass spectrometry for global proteome profiling". In : *Mass spectrometry reviews* 33.6 (2014), p. 452–470. DOI : 10.1002/mas.21400 (cf. p. 51).

[74] Cox, J. Neuhauser, N. Michalski, A. Scheltema, R. A. Olsen, J. V. Mann, M. "Andromeda : a peptide search engine integrated into the MaxQuant environment". In : *Journal of proteome research* 10.4 (2011), p. 1794–1805. DOI : 10.1021/pr101065j (cf. p. 55, 62).

[75] Cox, J. Mann, M. "MaxQuant enables high peptide identification rates, individualized ppb-range mass accuracies and proteome-wide protein quantification". In : *Nature biotechnology* 26.12 (2008), p. 1367–1372. DOI : 10.1038/nbt.1511 (cf. p. 55, 62, 82).

[76] Tyanova, S. Temu, T. Sinitcyn, P. Carlson, A. Hein, M. Y. Geiger, T. Mann, M. Cox, J. "The Perseus computational platform for comprehensive analysis of (prote) omics data". In : *Nature Methods* (2016). DOI : 10.1038/nmeth.3901 (cf. p. 55, 82).

[77] Dupierris, V. Barthe, D. Bruley, C. "ePIMS : un LIMS pour la gestion des données de spectrométrie de masse". In : *Spectra Analyse* 38.269 (2009), p. 36 (cf. p. 62).

[78] Dupierris, V. Masselon, C. Kieffer-Jaquinod, S. Bruley, C. "A toolbox for validation of mass spectrometry peptides identification and generation of database : IRMa". In : *Bioinformatics* 25.15 (2009), p. 1980–1981. DOI : 10.1093/bioinformatics/btp301 (cf. p. 62).

[79] Hesse, A.-M. Dupierris, V. Adam, C. Court, M. Barthe, D. Emadali, A. Masselon, C. Ferro, M. Bruley, C. "hEIDI : An intuitive application tool to organize and treat large-scale proteomics data". In : *Journal of Proteome Research* (2016). DOI : 10.1021/acs.jproteome.5b00853 (cf. p. 62).

[80] URL : http://www.profiproteomics.fr/ (cf. p. 62).

[81] Liu, H. Sadygov, R. G. Yates, J. R. "A model for random sampling and estimation of relative protein abundance in shotgun proteomics". In : *Analytical chemistry* 76.14 (2004), p. 4193–4201. DOI : 10.1021/ac0498563 (cf. p. 62).

[82] URL : http://shiny.rstudio.com/ (cf. p. 66).

[83] Giardine, B. Riemer, C. Hardison, R. C. Burhans, R. Elnitski, L. Shah, P. Zhang, Y. Blankenberg, D. Albert, I. Taylor, J. "Galaxy : a platform for interactive large-scale genome analysis". In : *Genome research* 15.10 (2005), p. 1451–1455. DOI : 10.1101/gr.4086505 (cf. p. 67).

[84] Gentleman, R. C. Carey, V. J. Bates, D. M. Bolstad, B. Dettling, M. Dudoit, S. Ellis, B. Gautier, L. Ge, Y. Gentry, J. "Bioconductor : open software development for computational biology and bioinformatics". In : *Genome biology* 5.10 (2004), p. 1. DOI : 10.1186/gb-2004-5-10-r80 (cf. p. 67).

[85] Zaslavskiy, M. Bach, F. Vert, J.-P. "Global alignment of protein–protein interaction networks by graph matching methods". In : *Bioinformatics* 25.12 (2009), p. i259–1267. DOI : 10.1093/bioinformatics/btp196 (cf. p. 68).

[86] Gao, X. Xiao, B. Tao, D. Li, X. "A survey of graph edit distance". In : *Pattern Analysis and applications* 13.1 (2010), p. 113–129. DOI : 10.1007/s10044-008-0141-y (cf. p. 68).

[87] URL : http://prospectom.liglab.fr/ (cf. p. 68).

[88] Alper, B. Bach, B. Henry Riche, N. Isenberg, T. Fekete, J.-D. "Weighted graph comparison techniques for brain connectivity analysis". In : *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM. 2013, p. 483–492. DOI : 10.1145/2470654.2470724 (cf. p. 69).

[89] URL : http://biovis.net/2016/index.html (cf. p. 69).

[90] Karpievitch, Y. Stanley, J. Taverner, T. Huang, J. Adkins, J. N. Ansong, C. Heffron, F. Metz, T. O. Qian, W.-J. Yoon, H. "A statistical framework for protein quantitation in bottom-up MS-based proteomics". In : *Bioinformatics* 25.16 (2009), p. 2028–2034. DOI : 10.1093/bioinformatics/btp362 (cf. p. 70).

[91] Michalski, A. Cox, J. Mann, M. "More than 100,000 detectable peptide species elute in single shotgun proteomics runs but the majority is inaccessible to data-dependent LC- MS/MS". In : *Journal of proteome research* 10.4 (2011), p. 1785–1793. DOI : 10.1021/pr101060v (cf. p. 70).

[92] Karpievitch, Y. V. Dabney, A. R. Smith, R. D. "Normalization and missing value imputation for label-free LC-MS analysis". In : *BMC bioinformatics* 13.Suppl 16 (2012), S5. DOI : 10.1186/1471-2105-13-S16-S5 (cf. p. 70, 72).

[93] Deeb, S. J. D'Souza, R. C. Cox, J. Schmidt-Supprian, M. Mann, M. "Super-SILAC allows classification of diffuse large B-cell lymphoma subtypes by their protein expression profiles". In : *Molecular & Cellular Proteomics* 11.5 (2012), p. 77–89. DOI : 10.1074/mcp.M111.015362 (cf. p. 70).

[94] Piehowski, P. D. Petyuk, V. A. Orton, D. J. Xie, F. Moore, R. J. Ramirez-Restrepo, M. Engel, A. Lieberman, A. P. Albin, R. L. Camp, D. G. "Sources of technical variability in quantitative LC–MS proteomics : human brain tissue sample analysis". In : *Journal of proteome research* 12.5 (2013), p. 2128–2137. DOI : 10.1021/pr301146m (cf. p. 70).

[95] Rubin, D. B. "Inference and missing data". In : *Biometrika* 63.3 (1976), p. 581–592. DOI : 10.1093/biomet/63.3.581 (cf. p. 70).

[96] Rubin, D. B. Schenker, N. "Interval Estimation from Multiply-Imputed Data : A Case Study Using Census Agriculture Industry Codes". In : *Journal of Official Statistics* 3.4 (1987), p. 375–387 (cf. p. 70).

[97] Schafer, J. L. *Analysis of incomplete multivariate data*. CRC press, 1997 (cf. p. 70).

[98] Royston, P. "Multiple imputation of missing values". In : *Stata journal* 4.3 (2004), p. 227–41. URL : http://ageconsearch.umn.edu/bitstream/116244/2/sjart_st0067.pdf (cf. p. 70).

[99] Little, R. J. Rubin, D. B. *Statistical analysis with missing data*. John Wiley & Sons, 2014 (cf. p. 70).

[100] Webb-Robertson, B.-J. M. Wiberg, H. K. Matzke, M. M. Brown, J. N. Wang, J. McDermott, J. E. Smith, R. D. Rodland, K. D. Metz, T. O. Pounds, J. G. "Review, evaluation, and discussion of the challenges of missing value imputation for mass spectrometry-based label-free global proteomics". In : *Journal of proteome research* 14.5 (2015), p. 1993–2001. DOI : 10.1021/pr501138h (cf. p. 70).

[101] Luo, R. Colangelo, C. M. Sessa, W. C. Zhao, H. "Bayesian analysis of iTRAQ data with nonrandom missingness : identification of differentially expressed proteins". In : *Statistics in biosciences* 1.2 (2009), p. 228–245. DOI : 10.1007/s12561-009-9013-2 (cf. p. 70).

[102] Taylor, S. L. Leiserowitz, G. S. Kim, K. "Accounting for undetected compounds in statistical analyses of mass spectrometry 'omic studies". In : *Statistical applications in genetics and molecular biology* 12.6 (2013), p. 703–722. DOI : 10.1515/sagmb-2013-0021 (cf. p. 70).

[103] Ryu, S. Y. Qian, W.-J. Camp, D. G. Smith, R. D. Tompkins, R. G. Davis, R. W. Xiao, W. "Detecting differential protein expression in large-scale population proteomics". In : *Bioinformatics* (2014). DOI : 10.1093/bioinformatics/btu341 (cf. p. 70).

[104] Chen, L. S. Prentice, R. L. Wang, P. "A penalized EM algorithm incorporating missing data mechanism for Gaussian parameter estimation". In : *Biometrics* 70.2 (2014), p. 312–322. DOI : 10.1111/biom.12149 (cf. p. 70).

[105] O'Brien, J. Gunawardena, H. Chen, X. Ibrahim, J. Qaqish, B. "The Midpoint Mixed Model with a Missingness Mechanism (M5) : A Likelihood-Based Framework for Quantification of Mass Spectrometry Proteomics Data". In : *(Preprint)* (2015). URL : https://arxiv.org/pdf/1507.06907 (cf. p. 70).

[106] Eidhammer, I. Barsnes, H. Eide, G. E. Martens, L. *Computational and statistical methods for protein quantification by mass spectrometry*. John Wiley & Sons, 2012. Chap. 10.7.1. DOI : 10.1002/9781118494042 (cf. p. 72).

[107] Bø, T. H. Dysvik, B. Jonassen, I. "LSimpute : accurate estimation of missing values in microarray data with least squares methods". In : *Nucleic acids research* 32.3 (2004), e34–e34. DOI : 10.1093/nar/gnh026 (cf. p. 73).

[108] Robertson, T. Wright, F. T. Dykstra, R. L. *Order restricted statistical inference*. John Wiley & Sons, 1988. DOI : 10.1007/BF02924337 (cf. p. 74).

[109] Bukhman, Y. V. Dharsee, M. Ewing, R. Chu, P. Topaloglou, T. Le Bihan, T. Goh, T. Duewel, H. Stewart, I. I. Wisniewski, J. R. "Design and analysis of quantitative differential proteomics investigations using LC-MS technology". In : *Journal of Bioinformatics and Computational Biology* 6.01 (2008), p. 107–123. DOI : 10.1142/S0219720008003321 (cf. p. 75).

[110] Podwojski, K. Eisenacher, M. Kohl, M. Turewicz, M. Meyer, H. E. Rahnenführer, J. Stephan, C. "Peek a peak : a glance at statistics for quantitative label-free proteomics". In : *Expert review of proteomics* 7.2 (2010), p. 249–261. DOI : 10.1586/epr.09.107 (cf. p. 75).

[111] Blein-Nicolas, M. Xu, H. Vienne, D. Giraud, C. Huet, S. Zivy, M. "Including shared peptides for estimating protein abundances : A significant improvement for quantitative proteomics". In : *Proteomics* 12.18 (2012), p. 2797–2801. DOI : 10.1002/pmic.201100660 (cf. p. 75, 76).

[112] Tusher, V. G. Tibshirani, R. Chu, G. "Significance analysis of microarrays applied to the ionizing radiation response". In : *Proceedings of the National Academy of Sciences* 98.9 (2001), p. 5116–5121. DOI : 10.1073/pnas.091062498 (cf. p. 75, 79, 81, 82).

[113] Neyman, J. Pearson, E. S. "On the problem of the most efficient tests of statistical hypotheses". In : *Breakthroughs in Statistics*. Springer, 1992, p. 73–108. DOI : 10.1007/978-1-4612-0919-5_6 (cf. p. 76).

[114] Wilks, S. S. "The large-sample distribution of the likelihood ratio for testing composite hypotheses". In : *The Annals of Mathematical Statistics* 9.1 (1938), p. 60–62 (cf. p. 76).

[115] Benjamini, Y. Hochberg, Y. "Controlling the false discovery rate : a practical and powerful approach to multiple testing". In : *Journal of the Royal Statistical Society. Series B (Methodological)* (1995), p. 289–300. URL : http://www.stat.purdue.edu/~doerge/BIOINFORM.D/FALL06/Benjamini%20and%20Y%20FDR.pdf (cf. p. 79, 80).

[116] Benjamini, Y. Hochberg, Y. "On the adaptive control of the false discovery rate in multiple testing with independent statistics". In : *Journal of Educational and Behavioral Statistics* 25.1 (2000), p. 60–83. URL : http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.477.4915&rep=rep1&type=pdf (cf. p. 79).

[117] Benjamini, Y. Yekutieli, D. "The control of the false discovery rate in multiple testing under dependency". In : *Annals of statistics* (2001), p. 1165–1188 (cf. p. 79).

[118] Benjamini, Y. Krieger, A. M. Yekutieli, D. "Adaptive linear step-up procedures that control the false discovery rate". In : *Biometrika* 93.3 (2006), p. 491–507. DOI : 10.1093/biomet/93.3.491 (cf. p. 79).

[119] Efron, B. Tibshirani, R. Storey, J. D. Tusher, V. "Empirical Bayes analysis of a microarray experiment". In : *Journal of the American statistical association* 96.456 (2001), p. 1151–1160. DOI : 10.1198/016214501753382129 (cf. p. 79).

[120] Storey, J. D. "A direct approach to false discovery rates". In : *Journal of the Royal Statistical Society : Series B (Statistical Methodology)* 64.3 (2002), p. 479–498. DOI : 10.1111/1467-9868.00346 (cf. p. 79).

[121] Storey, J. D. Tibshirani, R. "Statistical significance for genomewide studies". In : *Proceedings of the National Academy of Sciences* 100.16 (2003), p. 9440–9445. DOI : 10.1073/pnas.1530509100 (cf. p. 79).

[122] Storey, J. D. Taylor, J. E. Siegmund, D. "Strong control, conservative point estimation and simultaneous conservative consistency of false discovery rates : a unified approach". In : *Journal of the Royal Statistical Society : Series B (Statistical Methodology)* 66.1 (2004), p. 187–205. DOI : 10.1111/j.1467-9868.2004.00439.x (cf. p. 79).

[123] Serang, O. Käll, L. "The solution to statistical challenges in proteomics is more statistics, not less". In : *Journal of proteome research* (2015). DOI : 10.1021/acs.jproteome.5b00568 (cf. p. 81).

[124] Savitski, M. M. Wilhelm, M. Hahne, H. Kuster, B. Bantscheff, M. "A scalable approach for protein false discovery rate estimation in large proteomic data sets". In : *Molecular & Cellular Proteomics* (2015), mcp–M114. DOI : 10.1074/mcp.M114.046995 (cf. p. 81).

[125] Wilhelm, M. Schlegl, J. Hahne, H. Gholami, A. M. Lieberenz, M. Savitski, M. M. Ziegler, E. Butzmann, L. Gessulat, S. Marx, H. Mathieson, T. Lemeer, S. Schnatbaum, K. Reimer, U. Wenschuh, H. Mollenhauer, M. Slotta-Huspenina, J. Boese, J.-H. Bantscheff, M. Gerstmair, A. Faerber, F. Kuster, B. "Mass-spectrometry-based draft of the human proteome". In : *Nature* 509.7502 (2014), p. 582–587. DOI : 10.1038/nature13319 (cf. p. 81).

[126] Andlauer, T. F. Scholz-Kornehl, S. Tian, R. Kirchner, M. Babikir, H. A. Depner, H. Loll, B. Quentin, C. Gupta, V. K. Holt, M. G. "Drep-2 is a novel synaptic protein important for learning and memory". In : *Elife* 3 (2014), e03895. DOI : 10.7554/eLife.03895 (cf. p. 82).

[127] Hubner, N. C. Bird, A. W. Cox, J. Splettstoesser, B. Bandilla, P. Poser, I. Hyman, A. Mann, M. "Quantitative proteomics combined with BAC TransgeneOmics reveals in vivo protein interactions". In : *The Journal of cell biology* 189.4 (2010), p. 739–754. DOI : 10.1083/jcb.200911091 (cf. p. 82).

[128] Hornburg, D. Drepper, C. Butter, F. Meissner, F. Sendtner, M. Mann, M. "Deep proteomic evaluation of primary and cell line motoneuron disease models delineates major differences in neuronal characteristics". In : *Molecular & Cellular Proteomics* 13.12 (2014), p. 3410–3420. DOI : 10.1074/mcp.M113.037291 (cf. p. 82).

[129] Lundby, A. Andersen, M. N. Steffensen, A. B. Horn, H. Kelstrup, C. D. Francavilla, C. Jensen, L. J. Schmitt, N. Thomsen, M. B. Olsen, J. V. "In vivo phosphoproteomics analysis reveals the cardiac targets of β-adrenergic receptor signaling". In : *Sci. Signal.* 6.278 (2013), rs11–rs11. DOI : 10.1126/scisignal.2003506 (cf. p. 82).

[130] Geiger, T. Wehner, A. Schaab, C. Cox, J. Mann, M. "Comparative proteomic analysis of eleven common cell lines reveals ubiquitous but varying expression of most proteins". In : *Molecular & Cellular Proteomics* 11.3 (2012), p. M111–014050. DOI : 10.1074/mcp.M111.014050 (cf. p. 82).

[131] Smits, A. H. Jansen, P. W. Poser, I. Hyman, A. A. Vermeulen, M. "Stoichiometry of chromatin-associated protein complexes revealed by label-free quantitative mass spectrometry-based proteomics". In : *Nucleic acids research* (2012), gks941. DOI : 10.1093/nar/gks941 (cf. p. 82).

[132] Smaczniak, C. Li, N. Boeren, S. America, T. Dongen, W. Goerdayal, S. S. Vries, S. Angenent, G. C. Kaufmann, K. "Proteomics-based identification of low-abundance signaling and regulatory protein complexes in native plant tissues". In : *Nature protocols* 7.12 (2012), p. 2144–2158. DOI : 10.1038/nprot.2012.129 (cf. p. 82).

[133] Clough, T. Thaminy, S. Ragg, S. Aebersold, R. Vitek, O. "Statistical protein quantification and significance analysis in label-free LC-MS experiments with complex designs". In : *BMC bioinformatics* 13.Suppl 16 (2012), S6. DOI : 10.1186/1471-2105-13-S16-S6 (cf. p. 84).

[134] URL : https://fr.wikipedia.org/wiki/Analyse_de_la_variance (cf. p. 84).

[135] Hüllermeier, E. "Does machine learning need fuzzy logic ?" In : *Fuzzy Sets and Systems* 281 (2015), p. 292–299. DOI : 10.1016/j.fss.2015.09.001 (cf. p. 87).

[136] Searle, B. C. Turner, M. Nesvizhskii, A. I. "Improving sensitivity by probabilistically combining results from multiple MS/MS search methodologies". In : *The Journal of Proteome Research* 7.1 (2008), p. 245–253. DOI : 10.1021/pr070540w (cf. p. 87).

[137] Vaudel, M. Burkhart, J. M. Zahedi, R. P. Oveland, E. Berven, F. S. Sickmann, A. Martens, L. Barsnes, H. "PeptideShaker enables reanalysis of MS-derived proteomics data sets". In : *Nature biotechnology* 33.1 (2015), p. 22–24. DOI : 10.1038/nbt.3109 (cf. p. 87).

[138] Destercke, S. Strauss, O. "Kolmogorov-smirnov test for interval data". In : *International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems*. Springer International Publishing. 2014, p. 416–425. DOI : 10.1007/978-3-319-08852-5_43 (cf. p. 88).

[139] Denœux, T. "Likelihood-based belief function : justification and some extensions to low-quality data". In : *International Journal of Approximate Reasoning* 55.7 (2014), p. 1535–1547. DOI : 10.1016/j.ijar.2013.06.007 (cf. p. 88).

[140] Kanjanatarakul, O. Kaewsompong, N. Sriboonchitta, S. Denœux, T. "Estimation and prediction using belief functions : Application to stochastic frontier analysis". In : *Econometrics of Risk*. Springer International Publishing, 2015, p. 171–184. DOI : 10.1007/978-3-319-13449-9_12 (cf. p. 88).

[141] Tsuda, K. "Subspace classifier in the Hilbert space". In : *Pattern Recognition Letters* 20.5 (1999), p. 513–519. DOI : 10.1016/S0167-8655(99)00023-9 (cf. p. 92).

[142] Ding, C. He, X. "K-means clustering via principal component analysis". In : *Proceedings of the twenty-first international conference on Machine learning*. ACM. 2004, p. 29. DOI : 10.1145/1015330.1015408 (cf. p. 92).

[143] Cutler, A. Breiman, L. "Archetypal analysis". In : *Technometrics* 36.4 (1994), p. 338–347. DOI : 10.1080/00401706.1994.10485840 (cf. p. 93).

[144] Mørup, M. Hansen, L. K. "Archetypal analysis for machine learning and data mining". In : *Neurocomputing* 80 (2012), p. 54–63. DOI : 10.1016/j.neucom.2011.06.033 (cf. p. 93).

[145] Awate, S. Yu, Y.-Y. Whitaker, R. "Kernel principal geodesic analysis". In : *European Conference on Machine Learning (ECML) and Practice of Knowledge Discovery in Databases (PKDD)*. Springer LNAI. 2014, p. 82–98. DOI : 10.1007/978-3-662-44848-9_6 (cf. p. 93).

[146] Awate, S. P. Koushik, N. N. "Robust Dictionary Learning on the Hilbert Sphere in Kernel Feature Space". In : *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer International Publishing. 2016, p. 731–748. DOI : 10.1007/978-3-319-46128-1_46 (cf. p. 93).

[147] Gillet, L. C. Navarro, P. Tate, S. Röst, H. Selevsek, N. Reiter, L. Bonner, R. Aebersold, R. "Targeted data extraction of the MS/MS spectra generated by data-independent acquisition : a new concept for consistent

and accurate proteome analysis". In : *Molecular & Cellular Proteomics* 11.6 (2012), O111–016717. DOI : 10.1074/mcp.O111.016717 (cf. p. 94).

[148] Jutten, C. Herault, J. "Blind separation of sources, part I : An adaptive algorithm based on neuromimetic architecture". In : *Signal processing* 24.1 (1991), p. 1–10. DOI : 10.1016/0165-1684(91)90079-X (cf. p. 94, 97).

[149] Venable, J. D. Dong, M.-Q. Wohlschlegel, J. Dillin, A. Yates, J. R. "Automated approach for quantitative analysis of complex peptide mixtures from tandem mass spectra". In : *Nature methods* 1.1 (2004), p. 39–45. DOI : 10.1038/nmeth705 (cf. p. 94).

[150] Gallien, S. Bourmaud, A. Kim, S. Y. Domon, B. "Technical considerations for large-scale parallel reaction monitoring analysis". In : *Journal of proteomics* 100 (2014), p. 147–159. DOI : 10.1016/j.jprot.2013.10.029 (cf. p. 94).

[151] URL : https://shop.biognosys.ch/spectronaut (cf. p. 94).

[152] Röst, H. L. Rosenberger, G. Navarro, P. Gillet, L. Miladinović, S. M. Schubert, O. T. Wolski, W. Collins, B. C. Malmström, J. Malmström, L. "OpenSWATH enables automated, targeted analysis of data-independent acquisition MS data". In : *Nature biotechnology* 32.3 (2014), p. 219–223. DOI : 10.1038/nbt.2841 (cf. p. 95).

[153] MacLean, B. Tomazela, D. M. Shulman, N. Chambers, M. Finney, G. L. Frewen, B. Kern, R. Tabb, D. L. Liebler, D. C. MacCoss, M. J. "Skyline : an open source document editor for creating and analyzing targeted proteomics experiments". In : *Bioinformatics* 26.7 (2010), p. 966–968. DOI : 10.1093/bioinformatics/btq054 (cf. p. 95).

[154] Panchaud, A. Scherl, A. Shaffer, S. A. Haller, P. D. Kulasekara, H. D. Miller, S. I. Goodlett, D. R. "Precursor acquisition independent from ion count : how to dive deeper into the proteomics ocean". In : *Analytical chemistry* 81.15 (2009), p. 6481–6488. DOI : 10.1021/ac900888s (cf. p. 95).

[155] Egertson, J. D. Kuehn, A. Merrihew, G. E. Bateman, N. W. MacLean, B. X. Ting, Y. S. Canterbury, J. D. Marsh, D. M. Kellmann, M. Zabrouskov, V. "Multiplexed MS/MS for improved data-independent acquisition". In : *Nature methods* 10.8 (2013), p. 744–746. DOI : 10.1038/nmeth.2528 (cf. p. 95).

[156] Weisbrod, C. R. Eng, J. K. Hoopmann, M. R. Baker, T. Bruce, J. E. "Accurate peptide fragment mass analysis : multiplexed peptide identification and quantification". In : *Journal of proteome research* 11.3 (2012), p. 1621–1632. DOI : 10.1021/pr2008175 (cf. p. 96).

[157] Zhang, N. Li, X.-j. Ye, M. Pan, S. Schwikowski, B. Aebersold, R. "ProbIDtree : An automated software program capable of identifying multiple peptides from a single collision-induced dissociation spectrum collected by a tandem mass spectrometer". In : *Proteomics* 5.16 (2005), p. 4096–4106. DOI : 10.1002/pmic.200401260 (cf. p. 96).

[158] Silva, J. C. Denny, R. Dorschel, C. A. Gorenstein, M. Kass, I. J. Li, G.-Z. McKenna, T. Nold, M. J. Richardson, K. Young, P. "Quantitative proteomic analysis by accurate mass retention time pairs". In : *Analytical chemistry* 77.7 (2005), p. 2187–2200. DOI : 10.1021/ac048455k (cf. p. 96).

[159] Wang, J. Bourne, P. E. Bandeira, N. "MixGF : spectral probabilities for mixture spectra from more than one peptide". In : *Molecular & Cellular Proteomics* 13.12 (2014), p. 3688–3697. DOI : 10.1074/mcp.O113.037218 (cf. p. 96).

[160] URL : http://www.physikron.com/ (cf. p. 96).

[161] Kryuchkov, F. Verano-Braga, T. Hansen, T. A. Sprenger, R. R. Kjeldsen, F. "Deconvolution of mixture spectra and increased throughput of peptide identification by utilization of intensified complementary ions formed in tandem mass spectrometry". In : *Journal of proteome research* 12.7 (2013), p. 3362–3371. DOI : 10.1021/pr400210m (cf. p. 96).

[162] Elias, J. E. Gibbons, F. D. King, O. D. Roth, F. P. Gygi, S. P. "Intensity-based protein identification by machine learning from a library of tandem mass spectra". In : *Nature biotechnology* 22.2 (2004), p. 214–219. DOI : 10.1038/nbt930 (cf. p. 96).

[163] Käll, L. Storey, J. D. MacCoss, M. J. Noble, W. S. "Assigning significance to peptides identified by tandem mass spectrometry using decoy databases". In : *Journal of proteome research* 7.01 (2007), p. 29–34. DOI : 10.1021/pr700600n (cf. p. 96).

[164] Geiger, T. Cox, J. Mann, M. "Proteomics on an Orbitrap benchtop mass spectrometer using all-ion fragmentation". In : *Molecular & Cellular Proteomics* 9.10 (2010), p. 2252–2261. DOI : 10.1074/mcp.M110.001537 (cf. p. 97).

[165] Carvalho, P. C. Han, X. Xu, T. Cociorva, D. Gloria Carvalho, M. Barbosa, V. C. Yates, J. R. "XDIA : improving on the label-free data-independent analysis". In : *Bioinformatics* 26.6 (2010), p. 847–848. DOI : 10.1093/bioinformatics/btq031 (cf. p. 97).

[166] Bern, M. Finney, G. Hoopmann, M. R. Merrihew, G. Toth, M. J. MacCoss, M. J. "Deconvolution of mixture spectra from ion-trap data-independent-acquisition tandem mass spectrometry". In : *Analytical chemistry* 82.3 (2009), p. 833–841. DOI : 10.1021/ac901801b (cf. p. 97).

[167] Tsou, C.-C. Avtonomov, D. Larsen, B. Tucholska, M. Choi, H. Gingras, A.-C. Nesvizhskii, A. I. "DIA-Umpire : comprehensive computational framework for data-independent acquisition proteomics". In : *Nature methods* 12.3 (2015), p. 258–264. DOI : 10.1038/nmeth.3255 (cf. p. 97).

[168] Lee, D. D. Seung, H. S. “Learning the parts of objects by non-negative matrix factorization”. In : *Nature* 401.6755 (1999), p. 788–791. DOI : 10.1038/44565 (cf. p. 98, 102).

[169] Paatero, P. Tapper, U. “Positive matrix factorization : A non-negative factor model with optimal utilization of error estimates of data values”. In : *Environmetrics* 5.2 (1994), p. 111–126. DOI : 10.1002/env.3170050203 (cf. p. 102).

[170] Bioucas-Dias, J. M. Plaza, A. Dobigeon, N. Parente, M. Du, Q. Gader, P. Chanussot, J. “Hyperspectral unmixing overview : Geometrical, statistical, and sparse regression-based approaches”. In : *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* 5.2 (2012), p. 354–379. DOI : 10.1109/JSTARS.2012.2194696 (cf. p. 103).

[171] Recht, B. Re, C. Tropp, J. Bittorf, V. “Factoring nonnegative matrices with linear programs”. In : *Advances in Neural Information Processing Systems*. 2012, p. 1214–1222. URL : http://papers.nips.cc/paper/4518-factoring-nonnegative-matrices-with-linear-programs.pdf (cf. p. 103).

[172] Arora, S. Ge, R. Kannan, R. Moitra, A. “Computing a nonnegative matrix factorization–provably”. In : *Proceedings of the forty-fourth annual ACM symposium on Theory of computing*. ACM. 2012, p. 145–162. DOI : 10.1145/2213977.2213994 (cf. p. 103).

[173] Kumar, A. Sindhwani, V. Kambadur, P. “Fast Conical Hull Algorithms for Near-separable Non-negative Matrix Factorization”. In : *ICML (1)*. 2013, p. 231–239. URL : http://www.jmlr.org/proceedings/papers/v28/kumar13b.pdf (cf. p. 103).

[174] Donoho, D. Stodden, V. “When does non-negative matrix factorization give a correct decomposition into parts ?” In : *Advances in neural information processing systems*. 2003, None. URL : http://machinelearning.wustl.edu/mlpapers/paper_files/NIPS2003_LT10.pdf (cf. p. 103).

[175] Thurau, C. Kersting, K. Wahabzada, M. Bauckhage, C. “Descriptive matrix factorization for sustainability adopting the principle of opposites”. In : *Data Mining and Knowledge Discovery* 24.2 (2012), p. 325–354. DOI : 10.1007/s10618-011-0216-z (cf. p. 104).

[176] Wahabzada, M. Mahlein, A.-K. Bauckhage, C. Steiner, U. Oerke, E.-C. Kersting, K. “Metro maps of plant disease dynamics-automated mining of differences using hyperspectral images”. In : *PloS one* 10.1 (2015), e0116902. DOI : 10.1371/journal.pone.0116902 (cf. p. 105).

[177] URL : http://i-want-to-study-engineering.org/q/tetrahedron_volume/ (cf. p. 105).

[178] Kyrillidis, A. T. Becker, S. Cevher, V. Koch, C. “Sparse projections onto the simplex”. In : *ICML (2)*. 2013, p. 235–243. URL : http://www.jmlr.org/proceedings/papers/v28/kyrillidis13.pdf (cf. p. 107).

[179] Candes, E. J. Tao, T. “Decoding by linear programming”. In : *IEEE transactions on information theory* 51.12 (2005), p. 4203–4215. DOI : 10.1109/TIT.2005.858979 (cf. p. 107).

[180] Gill, P. E. Golub, G. H. Murray, W. Saunders, M. A. “Methods for modifying matrix factorizations”. In : *Mathematics of Computation* 28.126 (1974), p. 505–535. DOI : 10.1090/S0025-5718-1974-0343558-6 (cf. p. 108).

[181] Candès, E. J. Romberg, J. Tao, T. “Robust uncertainty principles : Exact signal reconstruction from highly incomplete frequency information”. In : *IEEE Transactions on information theory* 52.2 (2006), p. 489–509. DOI : 10.1109/TIT.2005.862083 (cf. p. 112).

[182] Cadzow, J. A. “Signal enhancement-a composite property mapping algorithm”. In : *IEEE Transactions on Acoustics, Speech, and Signal Processing* 36.1 (1988), p. 49–62. DOI : 10.1109/29.1488 (cf. p. 113).

[183] Chiron, L. Agthoven, M. A. Kieffer, B. Rolando, C. Delsuc, M.-A. “Efficient denoising algorithms for large experimental datasets and their applications in Fourier transform ion cyclotron resonance mass spectrometry”. In : *Proceedings of the National Academy of Sciences* 111.4 (2014), p. 1385–1390. DOI : 10.1073/pnas.1306700111 (cf. p. 113).

[184] Condat, L. Hirabayashi, A. “Super-resolution of positive spikes by Toeplitz low-rank approximation”. In : *Signal Processing Conference (EUSIPCO), 2015 23rd European*. IEEE. 2015, p. 459–463. DOI : 10.1109/EUSIPCO.2015.7362425 (cf. p. 113).

NB : Les publications co-signées par l’auteur se trouvent listée en Annexe 2. Par ailleurs, elles sont toutes téléchargeables à l’adresse suivante :
https://sites.google.com/site/thomasburgerswebpage/publications

Annexes

1 Curriculum Vitæ

2015-... : Chargé de recherche première classe au CNRS.

2013-... : Réferant de la thématique Knowledge Discovery from Proteomic Data, au sein du laboratoire EDyP (Etude de la Dynamique des Protéomes).

2011-2015 : Chargé de recherche deuxième classe au CNRS, affecté à BIG (FR3425 - Institut de Biosciences et Biotechnologies de Grenoble). Rattaché au laboratoire EDyP (Etude de la Dynamique des Protéomes) de l'unité de recherche BGE (U1038 - Biologie à Grande Echelle).

2008-2011 : Maître de Conférences à l'Université de Bretagne Sud (Morbihan) et au Lab-STICC.

2007-2008 : Qualifications aux fonctions de Maître de Conférences (sections 27 et 61). ATER à mi-temps à l'Institut Polytechnique de Grenoble (Ensimag).

2004-2007 : Doctorat Signal Image Parole & Telecom à l'Institut Polytechnique de Grenoble : *Reconnaissance automatique des gestes de la Langue Française Parlée Complétée* ; financement CIFRE (Orange), sous la direction d'Alice Caplier (GispaLab) et de Pascal Perret (Orange Labs).

2001-2004 : Diplôme d'Ingénieur en Télécommunication (spécialités Applications Réparties & Réseaux) à l'Institut Polytechnique de Grenoble (Ensimag). Master Recherche Mathématiques/Informatique (spécialité Recherche Opérationnelle & Combinatoire) à l'Institut Polytechnique de Grenoble (Ensimag). Projet de Fin d'Etude : *Caractérisation Labiale des phonèmes de la Langue Française Parlée Complétée* ; sous la direction de Denis Beautemps (Gipsa-Lab).

1999-2001 : Classes Préparatoires, Lycée Masséna (Nice).

2 Communications et publications

2.1 Articles de journaux internationaux

[J'1] Lazar, C. Gatto, L. Ferro, M. Bruley, C. **Burger, T.** "Accounting for the Multiple Natures of Missing Values in Label-Free Quantitative Proteomics Data Sets to Compare Imputation Strategies". In : *Journal of proteome research* 15.4 (2016), p. 1116–1125 (cf. p. 64, 66, 70).

[J'2] Gianetto, Q. G. Couté, Y. Bruley, C. **Burger, T.** "Uses and misuses of the fudge factor in quantitative discovery proteomics". In : *Proteomics* 16.14 (2016), p. 1955–1960 (cf. p. 64, 82).

[J'3] Giai Gianetto, Q. Combes, F. Ramus, C. Bruley, C. Couté, Y. **Burger, T.** "Calibration plot for proteomics : A graphical tool to visually check the assumptions underlying FDR control in quantitative experiments". In : *Proteomics* 16.1 (2016), p. 29–32 (cf. p. 64, 78, 82).

[J'4] Wieczorek, S. Combes, F. Lazar, C. Giai Gianetto, Q. Gatto, L. Dorffer, A. Hesse, A.-M. Coute, Y. Ferro, M. Bruley, C. **Burger, T.** "DAPAR & ProStaR : software to perform statistical analyses in quantitative discovery proteomics". In : *Bioinformatics* (2016) (cf. p. 65).

[J'5] Giai Gianetto, Q. Lazar, C. Bruley, C. Coute, Y. **Burger, T.** "Multiple imputation strategy for mass spectrometry-based proteomic data". In : (in prepration) (cf. p. 73, 74).

[J'6] **Burger, T.** Combes, F. Jacob, L. "More powerful differential analysis of relative quantitative proteomics data by leveraging shared peptides". In : (in prepration) (cf. p. 76).

[J'7] Aran, O. **Burger, T.** Caplier, A. Akarun, L. "A belief-based sequential fusion approach for fusing manual signs and non-manual signals". In : *Pattern Recognition* 42.5 (2009), p. 812–822 (cf. p. 86).

[J'8] Tomizioli, M. Lazar, C. Brugière, S. **Burger, T.** Salvi, D. Gatto, L. Moyet, L. Breckels, L. M. Hesse, A.-M. Lilley, K. S. "Deciphering thylakoid sub-compartments using a mass spectrometry-based approach". In : *Molecular & Cellular Proteomics* 13.8 (2014), p. 2147–2167 (cf. p. 89).

[J'9] Gatto, L. Breckels, L. M. Wieczorek, S. **Burger, T.** Lilley, K. S. "Mass-spectrometry-based spatial proteomics data analysis using pRoloc and pRolocdata". In : *Bioinformatics* 30.9 (2014), p. 1322–1324 (cf. p. 93).

[J'10] Courty, N. Gong, X. Vandel, J. **Burger, T.** "SAGA : Sparse And Geometry-Aware non-negative matrix factorization through non-linear local embedding". In : *Machine Learning* 97.1-2 (2014), p. 205–226 (cf. p. 104).

[J'11] **Burger, T.** "Geometric views on conflicting mass functions : From distances to angles". In : *International Journal of Approximate Reasoning* 70 (2016), p. 36–50.

[J'12] Kessentini, Y. **Burger, T.** Paquet, T. "A Dempster–Shafer Theory based combination of handwriting recognition systems with multiple rejection strategies". In : *Pattern Recognition* 48.2 (2015), p. 534–544.

[J'13] Pichon, F. Destercke, S. **Burger, T.** "A consistency-specificity trade-off to select source behavior in information fusion". In : *Cybernetics, IEEE Transactions on* 45.4 (2015), p. 598–609.

[J'14] Gatto, L. Breckels, L. M. **Burger, T.** Nightingale, D. J. Groen, A. J. Campbell, C. Mulvey, C. M. Christoforou, A. Ferro, M. Lilley, K. S. "A foundation for reliable spatial proteomics data analysis". In : *Molecular & Cellular Proteomics* (2014).

[J'15] Chapel, L. **Burger, T.** Courty, N. Lefèvre, S. "PerTurbo Manifold Learning Algorithm for Weakly Labeled Hyperspectral Image Classification". In : *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* (2014).

[J'16] **Burger, T.** Destercke, S. "How to randomly generate mass functions". In : *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems* 21.05 (2013), p. 645–673.

[J'17] Destercke, S. **Burger, T.** "Toward an Axiomatic Definition of Conflict Between Belief Functions". In : *Cybernetics, IEEE Transactions on* 43.2 (2013), p. 585–596.

[J'18] Caplier, A. Stillittano, S. Aran, O. Akarun, L. Bailly, G. Beautemps, D. Aboutabit, N. **Burger, T.** "Image and video for hearing impaired people". In : *Journal on Image and Video Processing* 2007.5 (2007), p. 1–14.

[J'19] **Burger, T.** Caplier, A. Perret, P. "Cued speech gesture recognition : a first prototype based on early reduction". In : *EURASIP Journal on Image and Video Processing* 2007.1 (2008), p. 1–19.

2.2 Sélections de l'éditeur et chapitres de livres

[S'1] **Burger, T.** Aran, O. Urankar, A. Caplier, A. Akarun, L. "A Dempster-Shafer Theory Based Combination of Classifiers for Hand Gesture Recognition". In : *Computer vision and computer graphics : theory and applications : international conference VISIGRAPP 2007, Barcelona, Spain, March 8-11, 2007, revised selected papers*. T. 21. Springer-Verlag New York Inc. 2008, p. 137–150 (cf. p. 86).

[S'2] Aran, O. **Burger, T.** Akarun, L. Caplier, A. "Gestural Interfaces for Hearing-Impaired Communication". In : *Multimodal User Interfaces* (2008), p. 219–250 (cf. p. 86).

[S'3] Aran, O. **Burger, T.** Caplier, A. Akarun, L. "Sequential Belief-Based Fusion of Manual and Non-manual Information for Recognizing Isolated Signs". In : *Gesture-Based Human-Computer Interaction and Simulation : 7th International Gesture Workshop, GW 2007, Lisbon, Portugal, May 23-25, 2007, Revised Selected Papers*. T. 5085. Springer Verlag. 2009, p. 134 (cf. p. 86).

[S'4] Allègre, W. **Burger, T.** Antoine, J.-Y. Berruet, P. Departe, J.-P. "A non-intrusive context-aware system for ambient assisted living in smart home". English. In : *Health and Technology* 3.2 (2013), p. 129–138. ISSN : 2190-7188.

2.3 Articles dans des actes internationaux

[P'1] Allègre, W. **Burger, T.** Berruet, P. "Model-driven flow for assistive home automation system design". In : *18th IFAC World Congress*. 2011 (cf. p. 86).

[P'2] Courty, N. **Burger, T.** Laurent, J. "Perturbo : A new classification algorithm based on the spectrum perturbations of the laplace-beltrami operator". In : *Machine Learning and Knowledge Discovery in Databases*. Springer, 2011, p. 359–374 (cf. p. 86, 92, 105).

[P'3] **Burger, T.** Caplier, A. "A generalization of the pignistic transform for partial bet". In : *Symbolic and Quantitative Approaches to Reasoning with Uncertainty*. Springer, 2009, p. 252–263 (cf. p. 86).

[P'4] Kessentini, Y. **Burger, T.** Paquet, T. "Evidential Combination of Multiple HMM Classifiers for Multiscript Handwritting Recognition". In : *Computational Intelligence for Knowledge-Based Systems Design*. Sous la dir. d'Eyke Hüllermeier, Rudolf Kruse et Frank Hoffmann. T. 6178. Lecture Notes in Computer Science. Heidelberg : Springer, 2010, p. 445–454. ISBN : 978-3-642-14048-8 (cf. p. 86).

[P'5] **Burger, T.** Kessentini, Y. Paquet, T. "Dealing with precise and imprecise decisions with a Dempster-Shafer Theory based algorithm in the context of handwriting recognition". In : *proceeedings of ICFHR 2010* (2010) (cf. p. 86).

[P'6] Kessentini, Y. **Burger, T.** Paquet, T. "Constructing dynamic frames of discernment in cases of large number of classes". In : *Symbolic and Quantitative Approaches to Reasoning with Uncertainty*. Springer, 2011, p. 275–286 (cf. p. 86).

[P'7] **Burger, T.** Kessentini, Y. Paquet, T. "Dempster-Shafer based rejection strategy for handwritten word recognition". In : *Document Analysis and Recognition (ICDAR), 2011 International Conference on*. IEEE. 2011, p. 528–532 (cf. p. 86).

[P'8] Courty, N. **Burger, T.** Marteau, P.-F. "Geodesic analysis on the gaussian RKHS hypersphere". In : *Machine Learning and Knowledge Discovery in Databases*. Springer, 2012, p. 299–313 (cf. p. 92).

[P'9] Courty, N. **Burger, T.** "A kernel view on manifold sub-sampling based on Karcher variance optimization". In : *Geometric Science of Information*. Springer, 2013, p. 751–758 (cf. p. 92, 93).

[P'10] **Burger, T.** "Geometric Interpretations of Conflict : A Viewpoint". In : *Belief Functions : Theory and Applications*. Springer, 2014, p. 412–421.

[P'11] Pichon, F. Destercke, S. **Burger, T.** "Selecting source behavior in information fusion on the basis of consistency and specificity". In : *Symbolic and Quantitative Approaches to Reasoning with Uncertainty*. Springer, 2013, p. 473–484.

[P'12] Allègre, W. **Burger, T.** Berruet, P. Antoine, J.-Y. "A non-intrusive monitoring system for ambient assisted living service delivery". In : *Impact Analysis of Solutions for Chronic Disease Prevention and Management*. Springer, 2012, p. 148–156.

[P'13] **Burger, T.** Destercke, S. "Random generation of mass functions : A short howto". In : *Belief Functions : Theory and Applications*. Springer, 2012, p. 145–152.

[P'14] Destercke, S. **Burger, T.** "Revisiting the notion of conflicting belief functions". In : *Belief Functions : Theory and Applications*. Springer, 2012, p. 153–160.

[P'15] **Burger, T.** Urankar, A. Aran, O. Akarun, L. Caplier, A. "Cued speech hand shape recognition". In : *2nd International Conference on Computer Vision Theory and Applications (VISAPP'07), Spain*. 2007.

[P'16] **Burger, T.** Aran, O. Caplier, A. "Modeling hesitation and conflict : A belief-based approach for multi-class problems". In : *Machine Learning and Applications, 2006. ICMLA'06. 5th International Conference on*. IEEE. 2006, p. 95–100.

[P'17] **Burger, T.** Benoit, A. Caplier, A. "Extracting static hand gestures in dynamic context". In : *Image Processing, 2006 IEEE International Conference on*. IEEE. 2006, p. 2081–2084.

[P'18] **Burger, T.** Caplier, A. Mancini, S. "Cued speech hand gestures recognition tool". In : *Proceedings of the 13th European Signal Processing Conference (EUSIPCO'05)*.

[P'19] Beautemps, D. **Burger, T.** Girin, L. "Characterizing and classifying cued speech vowels from labial parameters." In : *INTERSPEECH*. 2004.

2.4 Autres publications

[O'1] Gianetto, Q. G. Couté, Y. Bruley, C. **Burger, T.** "Estimer la proportion de valeurs manquantes complètement aléatoirement dans des jeux de données protéomiques". In : *48ièmes Journées de Statistique de la SFdS* (2016) (cf. p. 73).

[O'2] Despagne, W. **Burger, T.** "Introduction de l'ingénierie ontologique dans la méthodologie de développement des progiciels de gestion des collectivités territoriales". In : *Actes de ECG 2011* (2011) (cf. p. 86).

[O'3] Boulabiar, M.-I. **Burger, T.** Poirier, F. Coppin, G. "A low-cost natural user interaction based on a camera hand-gestures recognizer". In : *Human-Computer Interaction. Interaction Techniques and Environments*. Springer, 2011, p. 214–221 (cf. p. 86).

[O'4] Allègre, W. **Burger, T.** Berruet, P. "Aide à la conception d'un système domotique pour l'assistance aux personnes à mobilité réduite". In : *acte de la conférence Majestic 2010*. 2010 (cf. p. 86).

[O'5] Allègre, W. **Burger, T.** Berruet, P. Departe, J.-P. "Conception d'un système domotique pour l'assistance aux personnes dépendantes". In : *Sciences et Technologies pour le Handicap* (2010) (cf. p. 86).

[O'6] Allègre, W. Seguin, C. **Burger, T.** De Lamotte, F. Berruet, P. Philippe, J.-L. Diguet, J.-P. "Ambient Assisted Living with Linux". In : *eWili workshop*. 2011 (cf. p. 86).

[O'7] **Burger, T.** Dhorne, T. "Classification supervisée avec second étage optionnel pour variables de covariance conditionnelle hétérogène". In : *SFC 2009* (2009) (cf. p. 86).

[O'8] Kessentini, Y. Paquet, T. **Burger, T.** "Comparaison des méthodes probabilistes et évidentielles de fusion de classifieurs pour la reconnaissance de mots manuscrits". In : *CIFED'2010* (2010) (cf. p. 86).

[O'9] **Burger, T.** "Defining new approximations of belief functions by means of Dempster's combination". In : *Workshop on Theory of Belief Functions, Brest, France*. 2010, p. 11–16 (cf. p. 86).

[O'10] **Burger, T.** Cuzzolin, F. "The barycenters of the k-additive dominating belief functions and the pignistic k-additive belief functions". In : *Workshop on Theory of Belief Functions, Brest, France*. 2010, p. 1–6 (cf. p. 86).

[O'11] Chapel, L. **Burger, T.** Courty, N. Lefevre, S. "Classwise hyperspectral image classification with PerTurbo method". In : *Geoscience and Remote Sensing Symposium (IGARSS), 2012 IEEE International*. 2012, p. 6883 –6886.

[O'12] Allègre, W. **Burger, T.** Berruet, P. Antoine, J.-Y. "Modèle de supervision d'interactions non-intrusif basé sur les ontologies." In : *EGC*. 2012, p. 285–290.

[O'13] Beautemps, D. Girin, L. Aboutabit, N. Bailly, G. Besacier, L. Breton, G. **Burger, T.** Caplier, A. Cathiard, M.-A. Chêne, D. "Telma : Telephony for the hearing-impaired people. from models to user tests". In : *Proceedings of ASSISTH* (2007), p. 201–208.

2.5 Communications diverses

[M'1] **Burger, T.** Dhorne, T. "A Graphical Tool for the Detection of Modes in Continuous Data". In : *Use'R 2009*. 2009 (cf. p. 86).

[M'2] **Burger, T.** "Bridging modern machine learning to belief function theory". In : *ArXiv :1504.03874*. 2015 (cf. p. 87).

[M'3] Wieczorek, S. Combes, F. Lazar, C. Giai Gianetto, Q. Gatto, L. Dorffer, A. Hesse, A.-M. Coute, Y. Ferro, M. Bruley, C. **Burger, T.** "DAPAR & ProStaR : software to perform statistical analyses in quantitative discovery proteomics". In : *HUPO, Taipei, Taiwan*. 2016.

[M'4] Wieczorek, S. Combes, F. Lazar, C. Giai Gianetto, Q. Gatto, L. Dorffer, A. Hesse, A.-M. Coute, Y. Ferro, M. Bruley, C. **Burger, T.** "DAPAR & ProStaR : software to perform statistical analyses in quantitative discovery proteomics". In : *SFEAP, Chambery, France*. 2016.

[M'5] Wieczorek, S. Combes, F. Dorffer, A, Hesse, A.-M. Giai Gianetto, Q. Lazar, C. Ramus, C. Tardiff, M. Brugière, S. Coutè, Y. Bruley, C. **Burger, T.** "DAPAR (Differential Analysis of Proteins Abundance with R) : a new package with fancy graphical user interfaces". In : *SMAP, Ajaccio, France*. 2015.

[M'6] Lazar, C. Ferro, M. Bruley, C. **Burger, T.** "QRILC : a quantile regression approach to left-censored missing data imputation in quantitative high-throughput proteomics". In : *missDATA, Rennes, France*. 2015.

[M'7] Courty, N. Gong, X. Vandel, J. **Burger, T.** "SAGA : Sparse And Geometry-Aware non-negative matrix factorization through non-linear local embedding". In : (2015).

[M'8] Wieczorek, S. Combes, F. Hesse, A.-M. Lazar, C. Ramus, C. Brugière, S. Coutè, Y. Bruley, C. **Burger, T.** "Une application web et un package R pour l'analyse de données protéomiques quantitatives". In : *Atelier Prospectom, Grenoble, France*. 2015.

[M'9] Rolland, N. Tomizioli, M. Lazar, C. Salvi, D. Brugière, S. Moyet, L. **Burger, T.** Finazzi, G. Berny, D. Ferro, M. "Subcellular and subplastidial proteomics". In : *1st INPPO World Congress on Plant Proteomics : Methodology to Biology, Hamburg, Germany*. 2014.

[M'10] Gatto, L. Breckels, L. M. Wieczorek, S. **Burger, T.** Lilley, K. S. "A state-of-the-art machine learning pipeline for the analysis of organelle proteomics data". In : *HUPO, Madrid, Spain*. 2014.

[M'11] Wieczorek, S. Combes, F. Hesse, A.-M. Lazar, C. Ramus, C. Coutè, Y. Bruley, C. **Burger, T.** "A package R and a web application for the analysis of quantitative proteomics data". In : *SMAP, Lyon, France*. 2014.

[M'12] Lazar, C. Matringe, M. Brugière, S. Dziekan, J. Salvi, D. Rolland, N. Ferro, M. **Burger, T.** "To which extent spectral counting can be used to predict protein localization ?" In : *AdO'13 workshop (Apprentissage et données Omiques), CAP, PFIA, Lilles, France*. 2013.

[M'13] Bouveret, S. **Burger, T.** "Identification de protéines et prise en compte des peptides partagés". In : *Atelier Prospectom*. 2012.

[M'14] **Burger, T.** "Les défauts du calcul du FDR (False Discovery Rate)". In : *Atelier Prospectom*. 2012.

[M'15] **Burger, T.** Wieczorek, S. Masselon, C. Salvi, D. Rolland, N. Ferro, M. "Prediction of subplastidial localization of chloroplast proteins from spectral count data-Comparison of machine learning algorithms". In : *RECOMB sat. conf. on proteomics 2012*. 2012.

[M'16] Aran, O. **Burger, T.** Caplier, A. Akarun, L. "Sequential Belief-Based Fusion of Manual and Non-Manual Signs". In : *The 7th International Workshop on Gesture in Human-Computer Interaction and Simulation 2007, GW 2007, Lisbon, Portugal*. 2007.

3 Projets et financements

3.1 Protéomique (depuis 2012)

Bourse de thèse **IRTELIS** (financement CEA du contrat d'Olga Permiakova, 2016-2019) : Scalable computational methods for big proteomics data : application to demultiplexing of mass spectrometry signals. **Prospectom** est un projet nait de l'appel à manifestations d'intérêt MASTODONS de la Mission Interdisciplinaire du CNRS, dont l'objectif est de fournir des méthodes de fouilles de données adaptées aux grands volumes de spectres produits dans les études protéomiques. Financements : 20k€ (2012), 20k€ (2013), 25k€ (2014), 15k€ (2015). Ce projet a donné lieu à l'atelier du même nom en novembre 2012 et 2014 (http ://prospectom.liglab.fr/). Il a permis le financement des stages de Khawla, Sang, Hatem et Shivani.

Participation à divers projets impliquant tout le laboratoire, et dont je ne suis pas porteur : **Prime-XS**, JRA1 ; **ProFI** ; Labex **GRAL**. Participations mineures à de nombreux projets de protéomiques, financés par l'ANR, et impliquant entre autre, une phase d'extraction de connaissance (**ChloroPro**, finançant Cosmin Lazar, ou **RNAGermSilence**, finançant Quentin Giai Gianetto).

3.2 Apprentissage et science des données (depuis 2012)

Uncertainty in Machine Learning cooperation network (**UML-Net**) supervisé par Sébastien Destercke. Coordination élargie de l'équipe-action **Khronos** du labex **Persyval**, supervisé par Zaid Harchaoui et Massih-Reza Amini. Action **ATLAS** auprès du GdR **MaDICS**, coordonnée par Marianne Clausel et Massih-Reza Amini.

3.3 Sciences de l'ingénieur (2008-2012, suivi 2012-2015)

IntelHome (Bourse région de 2009-2012), thèse de Willy Allègre. **ASIM** est un projet ANR e-santé courant sur 18 mois (2012-2013) en collaborations avec les PME VITY et KAPTALIA. Il propose une suite et un transfert des travaux de Willy Allègre. Un **contrat OSEO** (115 k€ pour les années universitaires 2010-2012) entre la société éditrice de solutions logicielles MGDIS et l'UBS a permis entre autre de financer les travaux post-doctoraux de Wilfried Despagne, de juillet 2010 à fin 2011, sur la définition d'ontologies pour les métadonnées statistiques. **DyDom** (2012-2015) est un financement CG56 / VITY pour la thèse de Valère Alibert.

3.4 Reconnaissance de gestes (2004-2012)

TELMA (Télephonie à l'usage des malentendants, Projet RNTS de 2006 à 2009) a prolonger le cadre de ma thèse. **SIMILAR** (2004 - 2007) est un European Net-

work Of Excellence proposant l'étude d'interfaces hommes-machines multimodales. Il finança durant ma collaboration de thèse avec l'Université Boğaziçi d'Istanbul. **Rev-TV** (2010-2013) est un projet FUI sur le développement d'émissions de télévision interactives. Il est le cadre du stage et du CDD de Mohamed Ikbel Boulabiar, ainsi que du post-doctorat de Mathieu Simonet.

4 Encadrement d'activités de recherche

Les noms soulignés correspondent à des personnels permanents, ou à des contrats en cours. Les autres correspondent à des contrats terminées.

1. Olga Permiakova (Doctorante à l'UGA) : Scalable computational methods for big proteomics data : application to demultiplexing of mass spectrometry signals (3 ans à partir d'octobre 2016, co-encadré avec Thomas Fortin).
2. Iban Fouad Djama (Stagiaire de M2) : Visualisation et traitement de signaux de spectrométrie de masse pour le séquençage de protéines (3 mois en 2016, co-encadré avec Thomas Fortin).
3. Hugo Fortin (Stagiaire de L3) : Développement d'une librairie JAVA pour la parallélisation des opérateurs de base en algèbre linéaire (1.5 mois en 2015, co-encadré avec Thomas Fortin).
4. Alexia Dorffer (Stagiaire de L3) : Développement et diffusion du logiciel de statistique pour la protéomique ProStaR (6 mois en 2015, co-encadré avec Samuel Wieczorek).
5. Shivani Shah (Stagiaire de M2, au LIG) : Visual analytics for peptide-based proteins inference (6 mois en 2015, co-encadré avec Renaud Blanch et Jean-Philippe Menetrey).
6. Quentin Giai Gianetto (post-doctorant, INSERM) : Test d'hypothèse et correction pour tests multiples en proteomique quantitative (2 ans, à partir d'Octobre 2014 - co-encadré avec Yohann Couté).
7. Thomas Fortin (Ingénieur-Chercheur en CDI au CEA) : groupe KDPD (depuis Aout 2014).
8. Hatem Loukil (Doctorant à l'université de Sfax) : Modèles graphiques pour l'identification automatique de peptides à partir de spectrogrammes de fragmentation (3 ans, à partir d'Octobre 2013 - co-encadré avec Yousri Kessentini).
9. Jimmy Vandel (post-doctorant, CEA) : Extraction du signal de quantification des peptides à partir de données de spectrométrie label free pour la sélection de protéines différentiellement exprimées (18 mois à partir d'Octobre 2013 - co-encadré avec Yohann Coute).
10. Florence Combes (Ingénieur-Chercheur en CDI au CEA) : groupe KDPD (depuis Janvier 2014).
11. Cosmin Lazar (post-doctorant, CEA) : Apprentissage et prediction de la localisation subplastidiale de protéines à partir de données de protéomiques quantitative différentielle (18 mois en 2013 - co-encadré avec Myriam Ferro).
12. Samuel Wieczorek (Ingénieur-Chercheur en CDI au CEA) : groupe KDPD (depuis Janvier 2013).

13. Khawla Seddiki (stagiaire de M2) : Application de méthodes de sélection de variables pénalisée à des quantifications de protéines pour la découverte de bio-marqueurs (6 mois en 2013).
14. Van Sang Dao (stagiaire de M2) : Définition de co-similarités à partir de marches aléatoires sur les graphes bipartis sur des données "omics" (6 mois en 2013 - co-encadré avec Gilles Bisson).
15. Valère Alibert (Doctorant à l'UBS) : DyDom : Gestion intelligente et dynamique de la consommation énergétique d'un habitat domotisé (3 ans, à partir de la rentrée 2012 - co-encadré avec Pascal Berruet et Mounir Lallali).
16. Mathieu Simonet (Post-Doctorant à Télécom Bretagne) : Reconnaissance de gestes pour émissions de télévision interactives (10 mois à partir de mars 2011).
17. Johann-Dan Laurent (Stagiaire en Master 2 ISD à l'UBS) : Apprentissage actif pour l'annotation de fichiers multi-modaux : application à la Langue Parlée Complétée ou à la Langue des Signes Française (6 mois en 2011).
18. Mohamed-Ikbel Boulabiar (CDD ingénieur à Telecom Bretagne) : Interfaces Homme-Machine mixtes tactiles et gestuelles (15 mois, à partir de septembre 2010 - co-encadré avec Gilles Coppin).
19. Wilfried Despagne (Post-Doctorant à l'UBS et intervenant pour MGDIS) : Mise en correspondance d'ontologies pour les métadonnées statistiques (18 mois à partir de juillet 2010).
20. Willy Allègre (Doctorant à l'UBS) : Approche composant pour la gestion intelligente d'un habitat domotisé (d'octobre 2009 à octobre 2012 - co-encadré avec Pascal Berruet). Prix IFRATH (Institut Fédératif de Recherche sur les Aides Techniques pour personnes Handicapées) de la meilleure thèse.
21. Mohamed-Ikbel Boulabiar (Stagiaire de M2) : Reconnaissance de gestes pour interaction TV 3D (4 mois en 2010 - co-encadré avec Gilles Coppin).
22. Adyl Kenouche (Prestataire de service en informatique) : Intégration logicielle du prototype du projet TELMA (4 mois en 2008).
23. Pierre Lemaire (Stagiaire de M2) : Etude de la pertinence topologique des descripteurs d'images utilisés dans les algorithmes de détection de visages par apprentissage (5 mois en 2008 - co-encadré avec Alice Caplier).
24. Pierre Lemaire (Stagiaire de M2) : Développement d'une application de reconnaissance gestuelle pour malentendants (5 mois en 2007).
25. Alexandra Urankar (Stagiaire de M2) : Développement et optimisation d'une application de traitements d'images temps réel pour une application de téléphonie à l'usage des malentendants (6 mois en 2006).

Cela correspond à un total de **3 ingénieurs-chercheurs permanents, 5 post-doctorants, 4 doctorants, 2 Ingénieurs contractuels et 11 stagiaires**. **Durant ma thèse**, j'ai co-supervisé avec ma directrice, les stages de fin d'étude d'Alexandra Urankar (2006), et de Pierre Lemaire (2007), ainsi que le mémoire de recherche de ce dernier en 2008 ; en même temps qu'un ingénieur en CDD, Adyl Kenouche. Toutes ces personnes ont travaillé sur des sujets connexes à ma thèse. **À l'université de Bretagne Sud**, j'ai co-encadré avec Gilles Coppin, de Telecom Bretagne, à Brest, le stage puis le CDD de Mohamed-Ikbel Boulabiar (2010), et le

post-doctorat de Mathieu Simonet (2011), qui portaient sur de la reconnaissance de gestes pour interfaces homme-machine (IHM) gestuelles. Il y a eu aussi le stage de Johann-Dan Laurent (2011), résultant d'une collaboration avec Jean-Yves Antoine, de l'université de Tours : Sur un sujet initiale très proche des IHM gestuelles, le travail a évolué vers l'apprentissage automatique, en collaboration avec Nicolas Courty. Notons aussi Wilfried Despagne, post-doctorant (2010-2011), qui travaillait sur la mise en place d'ontologie pour une PME (MGDIS). Enfin, j'ai co-encadré avec Pascal Berruet les thèses de Willy Allègre (2009-2012) puis de Valère Alibert (2012-2015) en domotique. Un peu moins d'un an et demi après être arrivé **à EDyP** (à la fin de l'hiver 2013), la direction m'a proposé de prendre en charge l'animation d'une des huit thématiques scientifiques du laboratoire. La thématique en question, KDPD, a pris vie grâce à la participation de trois ingénieurs-chercheurs permanents (en CDI au CEA) : Samuel Wieczorek, Florence Combes et Thomas Fortin. Ensemble, nous avons suivi les travaux de de Van Sang Dao (M2) et de Khawla Seddiki (M2) en 2013, suivi en 2015 de Shivani Sha (M2) d'Alexia Dorffer (L3) et d'Hugo Fortin (L3), puis en 2016, d'Iban Fouad Djama (M2). De plus, 3 post-doctorants participent ou ont participé aux travaux décrits dans ce document : Cosmin Lazar en 2013-2014, Jimmy Vandel en 2014-2015 et Quentin Giai Gianetto en 2015-2016. Enfin, les doctorants : j'ai participé à la supervision d'un étudiant de l'université de Sfax, Hatem Loukil (2013-2016) ; enfin, je suis le directeur de thèse d'Olga Permiakova (2016-2019).

5 Activités diverses

5.1 Rapporteur et relecteur

Protéomique : Proteomics (2015), Journal of Proteome Research (2015), Plos One (2015), Amino Acids (2013), ECCB computational proteomics workshop (2012), Molecular and Cellular Proteomics (2012).

Science des données : Int. J. of Pattern Recognition and Artificial Intelligence (2015), Soft Computing (2015), Int. Conf. on Belief Functions (2014), Annals of Mathematics and Artificial Intelligence (2014), Information Sciences (2014-2016), Int. J. of Uncertainty, Fuzziness and Knowledge-based Systems (2013), J. of Knowledge-Based Systems (2012, 2013), IEEE trans. on Fuzzy Systems (2012, 2014), Pattern Analysis and Applications (2011, 2012), Information Fusion (2010), Int. J. of Approximate Reasoning (2009, 2011-2016), IEEE trans. on Circuits and Systems for Video Technologies (2008), Int. J. of Image and Video Processing (2007).

Autres : Sensors (2016), Expert ANR (2014, 2016), IEEE J. of Selected Topics in Applied Earth Observations and Remote Sensing (2014), J. of Language Resources and Evaluation (2007), Intern. Symposium on circuits and system (2007).

5.2 Animation scientifique

- Prospectom 2012 (29 et 30 Nov.) et 2014 (19,20 et 31 Nov.) - http ://prospectom.liglab.fr/ : Comité de pilotage, comité d'organisation et comité scientifique
- IPMU 2014 - http ://www.ipmu2014.univ-montp2.fr/index.html : Organisation de la session speciale Uncertainty Management in Machine Learning : http ://www.ipmu2014.univ-montp2.fr/pages/sessions/MachineLearning.pdf

5.3 Séminaires

1. Sparse and geometry aware matrix factorization : application to mass spectrometry based proteomics (GIPSA-Lab, Grenoble, 2016)
2. Open problems in computational proteomics in the UML language (Workshop on Uncertainty in Machine Learning, 2015)
3. Traitement statistique des données de protéomique quantitative label-free (journée scientifique SFSM-SFEAP, Paris, 2014)
4. From clustering to efficient and sparse nonnegative matrix factorization (Tim-C, Grenoble, 2014)
5. Processing, analyzing and exploring high-throughput proteomics data (Grenoble Inderdisciplinary Days, 2013)
6. De FT R&D vers la recherche académique (Journée des doctorants de Orange Lab, Issy-les-Moulineaux, 2013).
7. Introduction to machine learning, Organelle Proteomics Data Analysis Workshop (Cambridge Center for Proteomics, UK, 2012).
8. PerTurbo : A new classification algorithm based on the spectrum perturbations of the Laplace-Beltrami operator (LJK ; LIG, Grenoble, 2012).
9. Apprentissage statistique en contexte de forte incertitude - Application à l'inférence de protéines (CEA/iRTSV/BGE, Grenoble, 2011).
10. Théorie de Dempster-Shafer et apprentissage semi-supervisé ou actif (LIG, Grenoble ; Orange-Labs, Meylan, 2010).
11. De l'intérêt de la théorie de Dempster-Shafer pour l'apprentissage automatique probabiliste (LIG, Grenoble, 2010).
12. La reconnaissance de langages gestuels et la gestion de leur imprécision naturelle (LITIS, Rouen, 2009).
13. Reconnaissance de gestes pour malentendants (Séminaires de candidature de MCF, 2008).
14. Reconnaissance automatique de systèmes gestuels pour malentendants (Tim-C, Grenoble, 2007).
15. French Cued Speech recognition (Boğaziçi, Istanbul, 2006).

5.4 Enseignements

2016-2017 : Institut Polytechnique de Grenoble (Ensimag, Master Complémentaire Big Data), (niveau bac+6) : Apprentissage statistique (cours - 24h - 36heqTD)
Formation CNRS : Utilisation du logiciel Proline pour le traitement de données de protéomique quantitative (Cours - 3h)

2015-2016 : Institut Polytechnique de Grenoble (Ensimag, Master Complémentaire Big Data), (niveau bac+6) : Apprentissage statistique (cours - 12h - 18heqTD)
Université Pierre-Mendès-France (IAE, L3 MGE, avec Céline Schiavon) : Mathématiques appliqués (CTD - 21h - 31.5heqTD), Statistiques (CTD - 24h - 36heqTD)

2014-2015 : Institut Polytechnique de Grenoble (Ensimag), 2A (niveau M1) : Projets de spécialité, Modélisation Aléatoire et Statistique (suivi de projets - 12h)

2013-2014 : Ecole Thématique du CNRS (spectrométrie de masse FT-MS, #1404036) : Méthodes statistiques pour la quantification relative label-free (Cours - 2h)

2012-2013 : Université Pierre-Mendès-France : L3 MGE : Statistiques (CTD - 20h - 20heqTD - L3, avec Céline Schiavon)

2011-2012 : Université de Bretagne Sud : IUT STID (niveau L2) : Logiciel Spécialisé : Introduction à SAS (CTD - 24h - 24heqTD, avec E Castello et F Michel)

2010-2011 : Université de Bretagne Sud : IUT STID (niveaux L1 et L2), Master Ingénierie Statistique & Décisionnelle (niveau M2) :

- Projet Informatique de Statistique (TD - 25.5h - 25.5heqTD - L1).
- Stat. Expl. & Descr. (TP - 72h - 72heqTD - L1, avec Thierry Dhorne)
- Logiciel Spécialisé : Introduction à SAS (CTD - 60h - 60heqTD - L2).
- Extraction de Connaissances à partir des Données (Cours/TD - 22h/22h - 33heqTD/22heqTD - M2, avec Philippe Lenca)
- Suivi de Stages L2 et L3.

2009-2010 : Université de Bretagne Sud : IUT STID (niveaux L1 et L2), Master Ingénierie Statistique & Décisionnelle (niveaux M1 et M2)

- Projet Informatique de Statistique (TD - 25.5h - 25.5heqTD - L1).
- Stat. Expl. & Descr. (TP - 72h - 72heqTD - L1, avec Thierry Dhorne)
- Logiciel Spécialisé : Introduction à SAS (CTD - 60h - 60heqTD - L2)
- Extraction de Connaissances à partir des Données (Cours/TD - 12h/12h - 18heqTD/12heqTD - M2, avec PF Marteau, V Monbet, P Lenca et A Antoni)
- Méthodes de Réduction Dimensionnelle et Etude de Variables Cachées (Cours/TD - 6h/8h - 9heqTD/8heqTD - M1, avec Valerie Monbet et Thierry Dhorne)
- Suivi de Stages L2 et L3

2008-2009 : Université de Bretagne Sud : IUT STID (niveaux L1 et L2), Ensibs (niveau L3), Master Ingénierie Statistique & Décisionnelle (niveau M1 et M2)

- Projet Informatique de Statistique (TD - 25.5h - 25.5heqTD - L1).
- Stat. Expl. & Descr. (TP - 72h - 72heqTD - L1, avec Thierry Dhorne)
- Logiciel Spécialisé : Introduction à SAS (CTD - 60h - 60heqTD - L2)
- Programmation Web (CTD - 48h - 48heqTD - L2).
- Méthodes de Réduction Dimensionnelle et Etude de Variables Cachées (Cours/TD - 6h/8h - 9heqTD/8heqTD - M1, avec Valerie Monbet et Thierry Dhorne)
- Programmation Web (TD - 20h - 20heqTD - L3, avec Salma Ben Sassi)
- Suivi de Stages L2 et L3

2007-2008 : Institut Polytechnique de Grenoble (Ensimag), niveaux L3 et M1

- Processus Aléatoires (Cours Magistraux/TD - 18h/18h - 27heqTD/18heqTD - M1), Dpt. Télécom, 2A (avec Hervé Guiol et Yann Dijoux)
- Random Processes (TD - 18h - 18heqTD - M1), Dpt. Math. Fi., 2A (avec Hervé Guiol)
- Probabilités Appliquées (TD - 18h - 18heqTD - L3), Tronc commun, 1A (avec Hervé Guiol)
- Projet de Spécialité en Informatique : Modélisation Aléatoire et Statistiques (Projet - 18h - 18heqTD - M1), Dpt. Math. Fi., 2A, (avec Hervé Guiol)

2006-2007, 2007-2008 : Institut Polytechnique de Grenoble (Ensimag), Processus Aléatoires (TD - 18h - 18heqTD - 2A), Dpt. Télécom (avec Hervé Guiol)

Résumé

Ce mémoire présente mon travail d'encadrement d'activités de recherche pour les années 2012-2016, ainsi que des perspectives pour les cinq années à venir. Au travers de la présentation de deux de mes projets de recherche, j'analyse les différentes questions suscitées par l'animation d'un groupe de recherche dont l'objectif est le développement d'outils et de méthodes permettant l'extraction de connaissances automatisées à partir de données de quantification relative en protéomique *label-free* obtenues par spectrométrie de masse haut-débit. Ces questions concernent notamment *(i)* l'encadrement de jeunes chercheurs et la valorisation de leurs activités ; *(ii)* la recherche de financements ; et surtout *(iii)* la gestion des difficultés spécifiques au contexte interdisciplinaire (mode de diffusion/valorisation, équilibre entre recherche et ingénierie, pilotage et priorisation des sujets de recherche, etc.). Le premier projet présenté, ProStaR, est un outil logiciel permettant de faciliter l'analyse statistique de données protéomiques. Au-delà de l'important travail d'ingénierie que sa réalisation a nécessité, je montre qu'il peut être le support de nombreux petits projets relativement indépendants mais novateurs en science des données. Le second projet, Reveal-MS, propose de résoudre le démultiplexage de spectres de peptides par des méthodes innovantes de factorisation non-négative de matrices de grandes tailles. À l'inverse du précédent projet et dans une logique complémentaire, celui-ci est moins motivé par les besoins quotidiens de la protéomique que par la possibilité à long terme de permettre une rupture dans l'état de l'art.

Title

Knowledge discovery from high-throughput discovery proteomic data

Abstract

This dissertation presents my research activity supervision during years 2012 through 2016, as well as prospects for the next five years. Through the presentation of two of my research projects, I analyze the various issues raised by leading a research group focused on the development of tools and methods for automated knowledge extraction from relative quantification of high-throughput mass spectrometry data for label-free quantitative proteomics. Mainly, these issues relates to *(i)* advising young researchers and publishing their result ; *(ii)* grant chasing ; and most importantly *(iii)* managing the interdisciplinary context (publication strategies, balance between research and engineering, management and prioritization of research subjects, etc.). The first project presented, referred to as ProStaR, is a software tool to ease the statistical analysis of proteomic data. Beyond the important engineering workload it required, I show it is now a support to many small and independent yet innovative projects in data science. The second project, referred to as Reveal-MS, proposes to address the demultiplexing of peptide spectra by innovative non-negative factorization methods for large matrices. In contrast with the previous project, it is less motivated by the daily needs of proteomic labs than by the long-term hope of a state-of-the-art breakthrough.