Étude des déterminants de la puissance statistique en spectrométrie de masse

Thomas Jouve

▶ **To cite this version:**

Thomas Jouve. Étude des déterminants de la puissance statistique en spectrométrie de masse. Sciences agricoles. Université Claude Bernard - Lyon I, 2009. Français. NNT : 2009LYO10251 . tel-00635493

HAL Id: tel-00635493

https://theses.hal.science/tel-00635493

Submitted on 25 Oct 2011

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

N^O d'ordre : 251-2009 Année 2009

THÈSE DE L'UNIVERSITÉ DE LYON

délivrée par

L'UNIVERSITÉ CLAUDE BERNARD - LYON 1

Ecole doctorale Évolution Écosystèmes Microbiologie Modélisation (ED341)

DIPLOME DE DOCTORAT
(arrêté du 7 août 2006)

Titre : Étude des déterminants de la puissance statistique en spectrométrie de masse

Dirigée par Pascal ROY
Co-dirigée par Patrick DUCOROY

Jury :
Philippe BESSE
Hubert CHARLES
Patrick DUCOROY
Jean-Louis GOLMARD
Bart MERTENS
Nicolas MOLINARI
Pascal ROY

Rapporteurs :
Jean-Louis GOLMARD
Bart MERTENS
Nicolas MOLINARI

par **Thomas JOUVE**

soutenue publiquement le 3 décembre 2009

Résumé

La spectrométrie de masse fait partie des technologies *haut débit* et offre à ce titre un regard inédit, à une échelle nouvelle, sur les protéines contenues dans divers échantillons biologiques. Les études biomédicales utilisant cette technologie sont de plus en plus nombreuses et visent à détecter de nouveaux *biomarqueurs* de différents processus biologiques, notamment de processus pathologiques à l'origine de cancers. Cette utilisation comme outil de criblage pose des questions quant à la capacité même des expériences de spectrométrie de masse dans cette détection. La puissance statistique traduit cette capacité et rappelle que les études doivent être calibrées pour offrir des garanties suffisantes de succès. Toutefois, cette exploration de la puissance statistique en spectrométrie de masse n'a pas encore été réalisée. L'objet de cette thèse est précisément l'étude des déterminants de la puissance pour la détection de biomarqueurs en spectrométrie de masse.

Une revue de la littérature a été réalisée, reprenant l'ensemble des étapes nécessaires du traitement du signal, afin de bien comprendre les techniques utilisées. Les méthodes statistiques disponibles pour l'analyse du signal ainsi traité sont revues et mises en perspective. Les situations de tests multiples, qui émergent notamment de ces données de spectrométrie de masse, suggèrent une redéfinition de la puissance, détaillée par la suite.

La puissance statistique dépend du plan d'expérience. La taille d'échantillon, la répartition entre groupes étudiés et l'effet différentiel ont été investigués, par l'intermédiaire de simulations d'expériences de spectrométrie de masse. On retrouve ainsi les résultats classiques de la puissance, faisant notamment ressortir le besoin crucial d'augmenter la tailles des études pour détecter des biomarqueurs, particulièrement lorsque ceux-ci présentent un faible effet différentiel.

Au delà de ces déterminants classiques de la puissance, des déterminants propres à la spectrométrie de masse apparaissent. Une chute importante de puissance est mise en évidence, due à l'erreur de mesure des technologies de spectrométrie de masse. Une synergie péjorative existe de plus entre erreur de mesure et procédure de contrôle du risque de première espèce de type FDR. D'autre part, les méthodes de détection des pics, par leurs imperfections (faux pics et pics manqués), induisent un contrôle suboptimal de ce risque de première espèce, conduisant à une autre chute de puissance.

Ce travail de thèse met ainsi en évidence trois niveaux d'intervention possibles pour améliorer la puissance des études : la meilleure calibration des plans d'expérience, la minimisation de l'erreur de mesure et l'amélioration des algorithmes de prétraitement. La technologie même de spectrométrie de masse ne pourra conduire de façon fiable à la détection de nouveaux biomarqueurs qu'au prix d'un travail à ces trois niveaux.

Mots-clés : Puissance statistique, protéomique, spectrométrie de masse, haut-débit, calibration, tests multiples, perte séquentielle de puissance

Abstract

Mass-spectrometry (MS) belongs to the *high-throughput* technologies and therefore offers an original perspective on proteins contained in various biological samples, at a new scale. Biomedical studies using this technology are increasingly frequent. They aim at detecting new *biomarkers* of different biological processes, especially pathological processes leading to cancer. This use as a screening tool asks questions regarding the very detection effectiveness of MS experiments. Statistical power is the direct translation of this effectiveness and reminds us that calibrated studies are required to offer sufficient guarantees of success. However, this exploration of statistical power in mass-spectrometry has not been performed yet. The theme of this work is precisely the study of power determinants for the detection of biomarkers in MS studies.

A literature review was performed, summarizing all necessary pretreatment steps of the signal analysis, in order to understand the utilized techniques. Available statistical methods for the analysis of this pretreated signal are also reviewed and put into perspective. Multiple testing settings arising from MS data suggest a power redefinition. This power redefinition is detailed.

Statistical power depends on the study design. Sample sizes, group repartition and the differential effect were investigated through MS experiment simulations. Classical results of statistical power are acknowledged, with an emphasis on the crucial need to increase sample sizes for biomarker detection, especially when these markers show low differential effects.

Beyond these classical power determinants, mass-spectrometry specific determinants appear. An important power drop is experienced when taking into account the high measurement variability encountered in mass-spectrometry. A detrimental synergy exists between measurement variability and type 1 error control procedures (e.g. FDR). Furtheremore, the imperfections of peak detection methods (false and missed peaks) induce a sub-optimal control of this type 1 error, leading to another power drop.

This work shows three possible intervention levels if we want to improve power in MS studies : a better study design, measurement variability minimisation and pretreatment algorithms improvements. Only a work at these three levels can guarantee reliable biomarker detections in these studies.

Keywords : Statistical power, proteomics, mass-spectrometry, high-throughput, calibration, multiple testing, sequential power loss

Table des matières

Avant-propos et remerciements

Faire de la recherche, c'est questionner l'évidence. Cette réflexion, peut-être tout sauf inédite, s'est développée en moi pendant ce travail de thèse et illustre la démarche qui a été la mienne pendant ces années de travail. Ce questionnement offre de nouvelles perspectives, permet parfois d'innover et toujours de réfléchir. Vient alors le temps de la construction, de nouvelles évidences que d'autres questionneront à leur tour.

Faire de la recherche, c'est aussi accepter et choisir la simplicité. Ma philosophie des sciences fait la part belle au principe de parcimonie, au célèbre rasoir d'Ockham dont la juste coupe guide dans les avancées scientifiques. Cette simplicité est un objectif constant du travail réalisé, dans l'idée de le rendre simple et abordable, même lorsqu'il s'agit d'objets d'études complexes. C'est encore cette simplicité qui est le plus souvent la source des plus grandes inspirations. C'est enfin cette simplicité qui permet de clarifier le discours, lorsque la science est partagée - ce qu'elle doit toujours être.

Faire de la recherche, c'est ainsi, dans le monde actuel, envisager sa dimension plurielle. On ne cherche pas seul, on n'avance pas isolément, on ne progresse pas en restant cantonné à son champ de recherche. Questionner l'évidence suppose avant tout l'existence d'un savoir commun. La richesse de la science actuelle tient dans son caractère partagé, dans la beauté de ses constructions intellectuelles et matérielles, dont la grandeur m'étonne toujours.

Participer à ces constructions demande aussi de toujours se poser la question de sa reponsabilité. Même si le travail réalisé ici ne pose *a priori* pas de problème éthique, le développement de la science impose toujours de s'interroger sur le développement parallèle de la sphère de notre capacité technologique et de celle de notre capacité à *répondre* de nos actes, selon une belle idée de Hans Jonas dans son ouvrage *Pour une éthique du futur*.

Enfin, un travail de thèse repose toujours sur plus de personnes qu'un doctorant ou qu'une équipe de recherche. Mon travail a ainsi bénéficié pour des raisons très diverses de l'expérience, de la sagesse, de l'intelligence, de la compréhension, de la patience, de nombreuses personnes, que je remercie ici.

A Gaëlle, tout d'abord et avant tout, pour avoir partagé et supporté mes choix, pour avoir été là et être celle qu'elle est (je ne peux qu'emprunter des mots) : mon or, mon sud, mon est et mon ouest.

A ma famille, ensuite, dont le soutien indéfectible et la fierté ont su m'encourager quand parfois la route était longue. A mes parents, mes soeurs, mes grands-mères.

A mes amis, qui se reconnaîtront sans aucun doute, dont la présence à mes côtés a toujours été une source de plaisir et une solide base pour avancer plus loin que je ne l'aurais pensé. A Sophie, Bruno, Magali, Sylvain, pour des heures ludiques magnifiques. A Greg, Rudy, Damien, JM, Gilles, nos chemins se suivent toujours sans se ressembler ! To my ducky friend Mark, whose enthusiasm for research and love of life showed me both can simultaneously be part of life.

A l'équipe du laboratoire de biostatistiques des HCL, dont la bonne humeur et la gourmandise m'ont toujours réjoui. Plus particulièrement, merci à Hadrien pour nos dialogues scientifiques et son pertinent optimisme bien à lui. Merci à Delphine, pour sa disponibilité, son écoute et ses conseils. Merci à Emmanuelle et Aurélien, mes voisins occasionnels dont l'anglais restera pour moi une belle image. Merci à Mariethé, dont la gentillesse, l'organisation et la disponibilité m'ont permis de mener à bien ce projet. Enfin, merci à Pascal, pour sa vision du point noir sur

une toile blanche, qui a guidé mes premiers pas dans la recherche.

A Caroline, Delphine et Patrick, gardiens du précieux spectromètre à leurs heures et partenaires pour faire avancer ce projet dans la réalité.

Aux membres de mon jury de thèse et à mes rapporteurs, qui ont tous accepté avec un enthousiasme formidable de contribuer à ce travail. Dr Mertens, special thanks for your visit to France and your work.

Enfin, je dédie ce travail à mes grands-pères, qui ont chacun contribué à me changer et à faire de moi ce que je suis aujourd'hui. Je marche aussi dans les traces qu'ils ont laissées.

Chapitre 1

Introduction

La biologie, à l'aube du XXIème siècle, vit une révolution technologique avec le développement de nouvelles technologies dites *haut débit* permettant de changer l'échelle des études. Le séquençage du génome humain, dont une première version a été rendue publique en 2001, est un exemple de ce changement d'échelle, permettant l'étude globale de l'ensemble des gènes humains. De nouvelles investigations sont possibles, de nouveaux champs d'étude ouverts, de nouvelles questions posées.

La biologie offre ainsi de nouvelles possibilités prometteuses et donne lieu à de nombreux travaux de recherche. Si ces nouvelles technologies paraissent alléchantes, force est de constater qu'elles doivent encore faire leurs preuves. Le séquençage du génome humain n'a pas encore livré les secrets attendus et les autres technologies haut débit n'ont pas encore acquis leurs lettres de noblesse.

Ces nouvelles technologies permettent une exploration inédite des échantillons biologiques, notamment en recherche biomédicale. Parmi ces technologies, le travail réalisé au cours de cette thèse se concentre sur la spectrométrie de masse. Celle-ci fait figure de doyenne. Elle existe en effet depuis les années 1960, mais son application aux nouvelles échelles de la biologie est récente. L'espoir des explorations inédites qu'elle permet est l'identification de nouvelles signatures de maladies directement à partir d'un échantillon, par exemple de sang. L'évaluation de ses capacités exploratoires devient primordiale, du fait notamment de son application de plus en plus fréquente à la recherche biomédicale. Il s'agit principalement de bien évaluer la capacité de détection de nouveaux *biomarqueurs*. Cette capacité de détection se traduit par le concept de *puissance statistique*. Cette puissance est influencée par de nombreux paramètres de l'expérience et de son exploitation. L'objet de cette thèse est ainsi d'étudier ces différents déterminants de la puissance statistique dans les études de spectrométrie de masse.

Le chapitre 2 précise le cadre conceptuel et méthodologique de ce travail. Le concept de biomarqueur y est détaillé, ainsi que les idées majeures des technologies haut-débit, particulièrement les applications de la spectrométrie de masse en médecine. Afin de bien comprendre les questions soulevées par le traitement de données de spectrométrie, la technologie est en outre présentée plus en détails, ainsi que les stratégies utilisées pour répondre à ces questions.

Le chapitre 3 présente une étude de puissance basée sur la simulation d'expériences de spectrométrie de masse. Le développement d'expériences simulées s'intègre naturellement dans ce travail comme une démarche d'ingénierie inversée, basée sur les connaissances du chapitre 2,

permettant la construction de spectres de masse virtuels réalistes. Ces simulations permettent d'une part la quantification de la perte de puissance attendue au cours de l'analyse des données de spectrométrie de masse, et d'autre part l'étude des variations de la puissance avec ses déterminants classiques, accessibles à l'expérimentateur et propres aux marqueurs recherchés.

L'analyse de données de spectrométrie induit de plus des déterminants propres de la puissance. Précisément, des étapes du prétraitement interfèrent d'une part avec les outils statistiques utilisés et agissent d'autre part directement sur la puissance. Le chapitre 4 explore à son tour l'effet de ces déterminants, afin d'en comprendre le mécanisme d'action et d'en évaluer l'importance.

L'étude de ces déterminants de la puissance n'est pas une fin en soi : elle a pour but de guider la démarche expérimentale et analytique dans les expériences à venir. Elle permet de valider les schémas expérimentaux développés jusqu'à présent et de calibrer les schémas des études à venir. Ces schémas doivent ainsi permettre de fixer des garanties quant à la rentabilité de l'expérience en terme de découvertes. Une discussion des résultats présentés ainsi que les perspectives de recherche sont ainsi présentées dans le chapitre 5.

Chapitre 2

Applications de la spectrométrie de masse en recherche biomédicale

La médecine actuelle fait face à de nouveaux défis diagnostiques, pronostiques et thérapeutiques. La recherche est plus que jamais nécessaire dans ce contexte. Les outils de la biologie moderne, particulièrement les technologies haut-débit, offrent de nouvelles perspectives pour aborder ces défis. De nouvelles études voient le jour, nécessitant une bonne compréhension des outils et de l'analyse des données. Ces éléments sont développés dans cette section.

2.1 Contexte biomédical

2.1.1 Concept de biomarqueur

Le concept de biomarqueur est à la mode en recherche biomédicale. Une simple recherche sur la base de données publique PubMed[1] avec le terme *biomarker* retourne plusieurs milliers d'articles. Si ce terme est aussi fréquemment utilisé dans la littérature actuelle, c'est notamment parce qu'il s'applique à de nombreux domaines, mais aussi parce qu'il représente l'objet de recherche de nombreuses équipes.

Un biomarqueur est défini comme une caractéristique phénotypique représentative d'un état d'intérêt. De façon plus appliquée, on s'intéressera ici aux biomarqueurs moléculaires utilisés en médecine. Il s'agit alors d'une molécule dont le profil de concentration est corrélé à une caractéristique clinique d'intérêt (par exemple, une maladie, ou la réponse à un traitement). Ainsi, la troponine T, une protéine humaine, est un exemple de biomarqueur, dans ce cas précis de l'ischémie myocardique, c'est à dire la carence d'apport en oxygène au myocarde. Lorsque la concentration plasmatique de troponine augmente, elle signe cette ischémie. Il est donc possible, sur la base de la connaissance de la concentation plasmatique de troponine T, d'évaluer l'oxygénation myocardique.

De nombreux biomarqueurs sont connus pour différentes pathologies humaines. On citera par exemple la créatinine pour l'insuffisance rénale, le NT-proBNP pour l'insuffisance cardiaque, le Prostate Specific Antigen (PSA) pour le cancer de la prostate, le CA15.3 pour le cancer du sein, la protéine C-réactive (CRP) pour les états inflammatoires de façon générique... Ces biomar-

[1] http ://www.pubmed.org

queurs sont quotidiennement utiles au médecin et par conséquent au patient, en permettant des diagnostics éventuellement plus précis, plus rapides ou plus précoces. Cela peut permettre une meilleure prise en charge thérapeutique et contribuer ainsi à une augmentation de la qualité des soins, par là-même de la qualité ou de l'espérance de vie.

Les caractéristiques d'un bon biomarqueur sont nombreuses. Il doit être sensible, c'est à dire permettre la détection de la majorité des cas. La CRP est un exemple de marqueur très sensible, avec une élévation presque systématique dans les premières heures de l'inflammation. En outre, un biomarqueur doit être spécifique, c'est à dire ne pas présenter un profil anormal alors que l'individu ne présente pas la caractéristique clinique d'intérêt (par exemple, lorsque l'individu n'est pas malade). La troponine T est un marqueur très spécifique, dont l'élévation de la concentration traduit presque systématiquement l'ischémie myocardique. Ces deux paramètres de sensibilité et de spécificité sont précisés dans la section 2.3. Au delà de ces considérations théoriques, un biomarqueur doit de préférence être facilement accessible (par exemple par un simple recueil d'urines, ou un simple prélèvement sanguin). Le recueil doit pouvoir se faire sans danger pour le patient, avec souvent un coût assez bas pour permettre, par exemple, un dépistage de masse, ou une utilisation courante en milieu hospitalier.

Même s'il existe de nombreux biomarqueurs pour différentes pathologies ou états cliniques, leur recherche constitue toujours un enjeu important de la biologie et de la médecine actuelles. En effet, si certaines pathologies sont bien pourvues en marqueurs fiables et précoces (en cardiologie notamment), d'autres ne disposent pas d'un tel arsenal. La créatinine, marqueur de l'insuffisance rénale, n'est ainsi qu'un reflet tardif du mauvais fonctionnement rénal. En outre, il s'agit d'une molécule dont l'élévation de la concentration ne traduit pas exclusivement l'insuffisance rénale. La néphropathie diabétique par exemple, conduit *in fine* à une insuffisance rénale et ne peut être diagnostiquée que tardivement car on ne dispose pas d'un marqueur précoce de son installation. Une prise en charge plus adaptée serait possible avec l'utilisation de nouveaux biomarqueurs plus précoces.

La détection d'un nouveau biomarqueur est basée sur deux étapes majeures :

- l'acquisition d'un signal d'intérêt, typiquement les concentrations protéiques d'intérêt,
- la recherche au sein des données ainsi obtenues des molécules correspondant à des biomarqueurs, typiquement présentant un différentiel d'expression entre deux conditions médicales.

Ces deux étapes correspondent à différents temps de travail nécessaires. L'acquisition du signal d'intérêt est soumise à un temps expérimental et à un temps de traitement du signal. La recherche de biomarqueurs demande un temps d'analyse statistique et parfois un temps d'identification par des méthodes bioinformatiques. On conçoit donc que le travail de détection de biomarqueurs fasse intervenir de nombreuses disciplines, de l'acquisition des spectres avec la meilleure qualité possible, à l'analyse statistique et bioinformatique avec les meilleurs outils disponibles.

2.1.2 Les méthodes haut-débit

La recherche de nouveaux biomarqueurs apparaît comme essentielle pour de nombreuses pathologies, dont le diagnostic précoce est un élément clé d'une bonne prise en charge. Des stratégies d'identification de ces biomarqueurs sont nécessaires. Toutefois, ces stratégies sont parfois complexes et souvent peu généralisables. Une compréhension des mécanismes intimes

de la pathologie est souvent nécessaire, obligeant à une analyse poussée du fonctionnement biologique. Cette compréhension requiert un investissement temporel et financier important.

Aussi, à l'image de la chimie combinatoire utilisée en pharmacie pour l'identification de nouvelles molécules thérapeutiques, de nouvelles méthodes sont apparues qui permettent d'explorer rapidement et à moindre coût un vaste espace de recherche. Le développement récent de la biologie haut-débit permet une telle exploration. Les techniques associées sont apparues à la fin des années 1990. Les différents champs d'application sont souvent collectivement dénommés *omiques*, d'après le suffixe qui leur est appliqué. La génomique s'intéresse à l'étude des gènes au sein du génome, la transcriptomique aux transcrits (i.e. aux ARN messagers et par conséquent à l'expression des gènes associés), la protéomique aux protéines. Ces trois principales applications de la biologie haut-débit ont été rejointes par d'autres applications telles que la métabolomique, qui s'intéresse à l'ensemble des métabolites d'un échantillon biologique.

Toutes ces applications permettent l'étude simultanée d'un grand nombre d'entités biologiques, que ce soit des gènes, des ARN messagers, des protéines... Ainsi, le niveau d'expression de plusieurs milliers de gènes peut être obtenu sur une unique *biopuce*, principale technologie utilisée en transcriptomique. De même, la spectrométrie de masse permet l'étude du contenu protéique total d'un échantillon complexe de sang, par exemple. Dans les deux cas, on obtient une photographie à un instant donné de l'état, à l'échelle considérée, d'un système biologique. La quantité de données acquise en une seule expérience confère à ces technologies leur statut réellement haut-débit.

Pour être bien utilisées, les méthodes haut-débit ne dispensent pas de comprendre le système biologique étudié. En revanche, elles permettent d'étudier l'état de nombreux objets biologiques en une unique expérience, ce qui accélère sensiblement les investigations. De plus, la grande taille de l'espace biologique investigué peut permettre de formuler de nouvelles hypothèses quant aux relations unissant les entités biologiques entre elles et avec les pathologies d'intérêt. Ces méthodes présentent donc deux attraits majeurs de rapidité et d'échelle d'exploration, par rapport aux expériences classiques, non haut-débit.

La transcriptomique repose essentiellement sur l'utilisation d'une unique technologie, à savoir les puces à ADN. A l'inverse existent de nombreuses technologies pertinentes en protéomique. Historiquement se sont développés les expériences d'électrophorèse bidimensionnelle, permettant l'études de fluides biologiques totaux. Toutefois, ces techniques peuvent difficilement être rendues haut-débit. Les techniques de spectrométrie de masse, apparues plus tôt, dans les années 60, se sont ainsi peu à peu imposées comme l'outil idéal pour la protéomique à haut débit. Au sein même des technologies de spectrométrie de masse existent de nombreuses techniques et applications possibles. Le séquençage protéique par des techniques de spectrométrie en tandem (MS/MS) en est un exemple, ainsi que l'identification protéique. L'étude détaillé d'un protéome peut faire appel à la technique dite de Liquid Chromatography / Mass Spectrometry (LC/MS). Les études comparatives à la recherche de biomarqueurs reposent quant à elles majoritairement sur les techniques *temps de vol*, dont le fonctionnement est présenté en section 2.2.1.

La transcriptomique et la protéomique offrent deux perspectives différentes sur les données, l'une au niveau des ARN messagers et l'autre au niveau protéique. Toutefois, l'ARN messager est sujet à des modifications post-transcriptionnelles, ce qui peut rendre difficile la corrélation entre

l'activité d'un gène et le nombre de copies de messagers mesuré. Les protéines représentent en outre les effecteurs biologiques et peuvent en ce sens être considérées comme un meilleur marqueur de l'activité biologique. Elles sont en revanche sujettes à des modifications post-traductionnelles (coupure entre pro-peptide et protéine finale par exemple) et à des activations / inactivations (phosphorylations par exemple), ce qui rend parfois complexe le lien entre la concentration protéique et l'activité biologique.

Enfin, la transcriptomique a fait l'objet de nombreux travaux de recherche, dont une partie est directement applicable à la protéomique. Les données issues des deux champs présentent en effet des formats très comparables. Toutefois, des questions propres à la protéomique se posent de façon de plus en plus fréquente, que ce soit pour des réponses quantitatives à des questions ouvertes en transcriptomique (par exemple, la normalisation) ou pour des questions conceptuellement nouvelles. Ce travail se concentre donc sur la protéomique.

2.1.3 Applications cliniques

Les cancers représentent aujourd'hui la première cause de mortalité dans les pays industrialisés, pour deux raisons principales : la meilleure prise en charge et donc la moins grande mortalité des pathologies infectieuses et cardiaques, et l'augmentation de la fréquence (plus précisément, de l'incidence) de certains cancers, pour des raisons parfois mal comprises, souvent en lien avec des facteurs environnementaux. A ce titre, la cancérologie est un domaine dans lequel la recherche est particulièrement active. Les enjeux sont importants car il s'agit de mieux prendre en charge des maladies au pronostic souvent sombre, avec un impact majeur sur la vie des sujets atteints.

Une meilleure prise en charge des cancers passe à la fois par une meilleure prévention et un meilleur traitement. Dans les deux cas, les biomarqueurs jouent un rôle essentiel. En effet, ils peuvent permettre un dépistage facilité ou un diagnostic précoce, mais ils permettent aussi dans d'autres cas de prédire la réponse au traitement, par exemple. Il s'agit donc de contribuer tant au diagnostic qu'à l'évaluation de la réponse au traitement ou du pronostic d'une maladie chez un patient.

La complexité des mécanismes moléculaires pathologiques à l'oeuvre dans les cellules cancéreuses rend leur compréhension souvent difficile. La biologie non haut-débit permet d'investiguer ces mécanismes, ce qui demande beaucoup de temps. La recherche de nouveux biomarqueurs de cancer est donc une tâche longue et difficile. Les méthodes haut-débit peuvent utilement jouer un rôle dans cette recherche en l'accélérant d'une part et en proposant d'autre part d'autres perspectives sur les systèmes biologiques (on pensera notamment aux recherches sur les réseaux de gènes et plus généralement à la biologie dite *systémique*). De nombreuses études tant en transcriptomique qu'en protéomique ont donc été réalisées, dans le but d'identifier des biomarqueurs. Les hémopathies ont fait l'objet d'un célèbre jeu de données par Golub et al. [1] en transcriptomique. De nombreux cancers ont fait l'objet d'études en protéomique, comme par exemple le cancer du sein par l'analyse d'échantillons sanguins [2, 3, 4, 5] ou par l'analyse de fluide d'aspiration canalaire [6], le cancer du pancréas [7], du colon [8, 9], de la prostate [10, 11, 12, 13], différentes tumeurs cérébrales par l'analyse du liquide céphalo-rachidien [14, 15]... On ne dispose pas de marqueur précoce de la pathologie pour certains de ces processus malins, rendant d'autant

plus intéressante leur recherche. Le cancer du pancréas, par exemple, est souvent diagnostiqué de façon tardive car il n'existe pas de marqueur sensible et spécifique de son apparition. Une revue pertinente de ces études oncologiques a été proposée en 2008 par Whelan [16].

Toutefois, les applications cliniques potentielles de la spectrométrie de masse pour la détection de biomarqueurs ne se limitent pas aux processus malins. La néphropathie diabétique a fait l'objet de différentes études [17, 18, 19], avec la particularité de faire intervenir un fluide secrété par l'organe atteint, à savoir l'urine. En gastro-entérologie, on peut citer la réponse aux anti-TNF alpha dans la maladie de Crohn [20]. Comme pour la recherche oncologique, il n'existe pas encore de biomarqueur issu d'expérience de spectrométrie de masse présentant une application clinique. Des défis restent à relever quant à l'identification de protéines pertinentes et non des protéines classiques de l'inflammation [21, 22, 23].

2.2 Comprendre la technologie

La spectrométrie de masse fournit des données complexes à analyser, à l'issue d'un processus d'analyse tout aussi complexe, reposant sur différentes techniques interagissant entre elles. Une revue de la littérature a donc été réalisée au cours de ce travail de thèse, dont la finalité est la bonne compréhension de ces techniques de traitement des données de spectrométrie de masse.

2.2.1 Principes de fonctionnement

La spectrométrie de masse est une technique d'analyse permettant la détermination de la composition moléculaire d'un échantillon. Toutes les technologies reposent sur trois grands éléments :

- un ioniseur, permettant de transformer en ions et en phase gazeuse, un échantillon biologique,
- un analyseur de masse, permettant de séparer les ions ainsi produits selon leur rapport masse (m) sur charge (z), classiquement noté m/z,
- un détecteur, permettant de réaliser un comptage des ions présents et de fournir un signal électrique enregistrable.

Les technologies dites *temps de vol* utilisent un analyseur reposant sur le temps de vol d'une particule chargée dans un champ électrique connu. Elles sont principalement représentées par les technologies *Matrix Assisted Laser Desorption / Ionization* (MALDI) et *Surface Enhanced Laser Desorption / Ionization* (SELDI). Dans les deux cas, l'échantillon a analysé est co-crystalisé avec une matrice moléculaire. Ce mélange est ionisé par un laser. La matrice (le plus souvent composée d'acide sinapinique) sert à absorber l'énergie du tir laser, afin d'éviter l'altération de l'échantillon d'intérêt, tout en fournissant des protons aux analytes. La technologie SELDI est une variation de la technologie MALDI. Elle comprend une étape de filtrage protéique. Ce filtrage est réalisé par une étape de fixation sur une surface aux propriétés chimiques particulières, puis un lavage des protéines non fixées. On peut ainsi réaliser l'analyse d'un sous-échantillon protéique.

L'histograme du nombre d'ions percutant le détecteur en fonction du temps de vol (i.e. du temps entre la mise en tension du champ dans le tube à vide et l'impact sur le détecteur) constitue le spectre de masse. La relation entre temps de vol et rapport m/z est déduite des lois

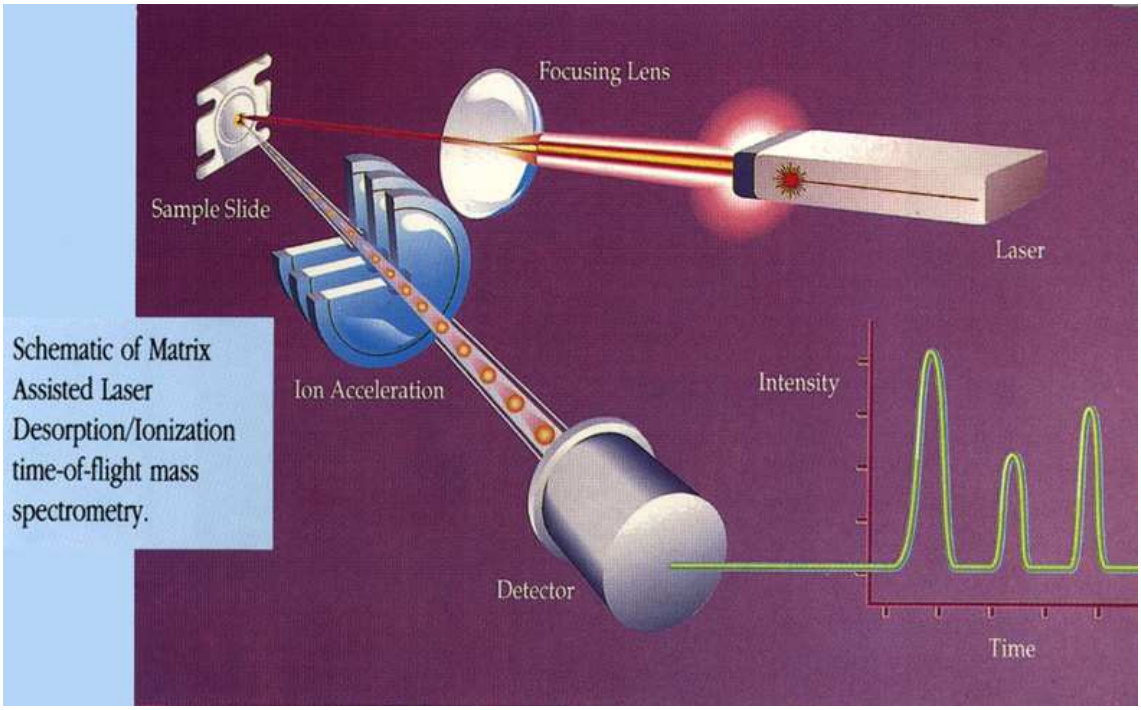

FIG. 2.1 – Principe de fonctionnement d'un appareil MALDI. L'impulsion laser atteint l'échantillon (*Sample Slide*), qui est ionisé et accéléré dans un tube à vide (*Ion Acceleration*). Le signal ionique est enregistré sur le détecteur (*Detector*) et converti en spectre de masse. Image propriété de l'entreprise Finnigan MAT.

classiques de l'électricité et de conservation de l'énergie (équivalence entre énergie de potentiel et énergie cinétique).

2.2.2 Traitement du signal de spectrométrie de masse

Les spectres produits par un spectromètre doivent subir des étapes de prétraitement avant qu'il ne soit possible d'en exploiter le contenu. Deux raisons principales rendent ce prétraitement nécessaire :

- les spectres intègrent différentes sources de variabilité instrumentale, ou bruit, qu'il faut corriger au mieux avant analyse,
- le spectre lui-même, en tant que mesure échantillonnée d'un signal analogique, n'est pas le signal d'intérêt, dans la majorité des approches adoptées.

Les données de spectrométrie de masse sont donc *bruitées* et fonctionnelles. Le prétraitement a pour principe général de transformer un ensemble de spectres en un ensemble de valeurs d'intensités lues, tout en opérant une correction des bruits présents dans le signal. L'extraction de ces intensités repose sur la lecture de l'intensité associée à chaque pic. Cette discrétisation du signal de spectrométrie repose sur des hypothèses quant aux pics, reliefs représentatifs des peptides dans les spectres. Ces hypothèses sont variables d'une équipe à l'autre et sont décrites plus en détail dans la revue de la littérature proposée dans ces pages.

Il faut noter l'existence d'approches fonctionnelles, développées essentiellement par l'équipe de Morris [24] et Antoniadis [25], ou plus récemment par Alexandrov et al. [26]. Elles restent peu ou pas utilisées en 2009. Leur principe est de comparer les individus sur la base des spectres

entiers, c'est-à-dire en utilisant le signal fonctionnel total. Cette approche est séduisante dans la mesure où elle évite la formulation d'hypothèses quant aux pics. Elle présente l'inconvénient de mettre en évidence des différences entre spectres portant sur les paramètres de modélisation de ces derniers. Il est alors difficile de faire le lien entre un paramètre de modélisation et une protéine, ce qui ne permet pas l'identification de nouveaux biomarqueurs. En revanche, cette approche fonctionnelle offre des possibilités de classification des spectres, sans fournir directement la liste des protéines permettant cette classification. Elles sortent en ce sens du cadre des méthodes d'identification des biomarqueurs, en l'état actuel.

En outre, des études de mise en évidence de *profils*, définis comme des ensembles de peptides d'intérêt, ont vu le jour. Celles-ci visent à classer les spectres selon plusieurs et non un unique peptide, sans chercher à identifier chacun de ces peptides. Bien que ces études ne visent pas explicitement la détection de nouveaux biomarqueurs, elles reposent quand même sur la détection d'éléments spectraux différentiels et justifient à ce titre tout autant le contrôle de la puissance statistique. L'idée est alors d'offrir des garanties quant à la capacité à détecter les différents éléments du profil.

Local features based methods in mass spectrometry proteomics: a review

Thomas Jouve*†, Delphine Maucort-Boulch†, Patrick Ducoroy‡and Pascal Roy†

July 16, 2009

Abstract

Background Mass spectrometry (MS) is an increasingly used technique in proteomics. Time-Of-Flight techniques enable the study of biological fluids, e.g. human blood. Analysis of these samples can lead to the discovery of new biomarkers which can ease the diagnosis and prognosis of several diseases, e.g. cancers. **Methods** We review methods for MS signal analysis. This analysis extends from raw spectra pre-processing, to selection of biomarkers in the signal. We focus on methods based on local features. **Results** Three main analysis steps are identified with specific sub-steps: i)pre-processing; ii)identification of features of interest; iii)selection of biomarkers among these features. At each level, different methods are considered and put into perspective. **Conclusions** A genereal workflow is derived from the set of available tools, with identification of key concepts. The dominant concept of local feature is defined, enabling a better understanding of MS data properties.

Keywords: Proteomics, mass-spectrometry, MALDI, SELDI, pre-processing, biomarkers

Introduction

Proteomics is a rapidly developing field among other *omics* disciplines, focusing on large biological datasets. High throughput methods are required to collect data in these fields. This kind of study requires the use of appropriate high throughput methods. Mass spectrometry (MS) offers an interesting insight on biological samples containing large numbers of proteins such as plasma. Its analysis can lead to the discovery of new biomarkers, thereby offering new diagnostic or prognostic tools. This works for a wide range of diseases supposed to affect the concentrations of circulating proteins. While transcriptomics focuses on the RNA level, proteomics is concerned with the next biological level in the universal dogma of genetic, namely proteins. Down- and up-regulation (amplification) affect RNA translation differentially and make RNA relative concentrations not as good a proxy to biological activity as proteins. Proteins are the actual effectors of biological functions and the measure of their expression level is in direct connection with their activity. Proteins are less influences by down- and up-regulation than RNA and might offer a better proxy to biological activity. Besides, proteins are exported out of the cell and can be detected in various biological fluids like blood, easily sampled. Nevertheless, it should also be pointed out that protein expression levels and activities are not exactly correlated.

*corresponding author: thomas.jouve@chu-lyon.fr

†Universite de Lyon, F-69000, Lyon ; Universite Lyon 1 ; CNRS, UMR5558, Laboratoire Biostatistique Sante, F-69622, Villeurbanne, France; Hospices Civils de Lyon, Service de Biostatistique, Lyon, F-69003, France

‡Faculties of Medicine and Pharmacy - INSERM U517, IFR 100 - 21000 Dijon - France

MS relies on proteins identification by their time of flight (TOF), related to their mass (m) to charge (z) ratio, usually written m/z. MS output is a mass spectrum, associating TOF with signal intensities. These intensities are related to concentrations of proteins in the processed sample. Despite numerous studies stating that MS methods are a powerful approach to detect diseases in medicine and biology, they still require improvement and validation [1][2][3][4]. Surface Enhanced Laser Desorption/Ionization and Matrix Assisted Laser Desorption/Ionization are two promising MS technology very commonly used in medical studies. We focus on the use of MS (SELDI or MALDI) for clinical applications (diagnosis, prognosis). Important issues concerning the reproducibility of the method as well as data analysis remain open questions that need further validation [5][6]. Our focus here is not reproducibility or the intrinsic quality of the method, but rather data analysis steps. Hence, we will only develop data processing steps for MALDI-TOF and SELDI-TOF mass spectra. Both MALDI and SELDI yield the same type of results, although MALDI-TOF offers a broader range of analysis (from 0 to about 50 kDA, whereas SELDI-TOF rather restricts to 0-30 kDa). Mass spectra from both methods require the same processing steps and will be considered hereafter together under the generic name *spectrum*. Whether these methods are quantitative is still questioned by some authors [3]. This aspect still deserves investigations before judging the potential of MS for quantitative analysis [7][8]. Mass spectrometry nevertheless appears as an exciting tool with great potential [9]. Two broad approaches are developed for MS data analysis. The identification of local features (e.g. a peak or a spectrum region) corresponding to proteins is necessary for the *local approach*. On the opposite, the *functional approach* deals with spectra as the basic unit for analysis [10][11]. This approach is not discussed here. The focus of this review is MS data analysis based on identification of local features. Issues concerning this analytical process can be divided in three main steps: pre-processing, identification of local features and selection of features of interest. A summary of the whole analytical process is presented in figure as the workflow schematic.

Working object changes throughout signal processing. Starting with raw spectrum, a series of signal improvement steps is performed, described in section 1. After such *cleaning* of the signal, identification of features of interest is performed, as developed in section 2. This in turn enables the discovery of biomarkers and possibly samples classification, as explained in section 3. Each step represents a specific issue.

1 Pre-processing steps

All steps described in this section aim at removing all forms of noise and artifacts introduced in the data by specific properties of the method. All these steps are related to the following equation:

$$I_i(t) = k \cdot S_i(t) + B_i(t) + N_i(t) \tag{1}$$

where i is a sample, I is the measured signal (i.e. the raw spectrum), S_i is the true signal, k is a scaling coefficient applied to the true signal, B_i and N_i are noise terms, respectively a baseline and random noise. t refers to time of flight values (proportional to m/z). The aim of pre-processing step is to isolate the true signal S_i. The order in which these steps should be performed still remains an open question and different positions were adopted by different authors. Furthermore, some methods are very likely to perform better together with some methods and worse with other. The order of pretreatment steps and interactions between methods were investigated by Arneberg et al.[12] through an original modeling approach. The interested reader is refered to this paper for more details.

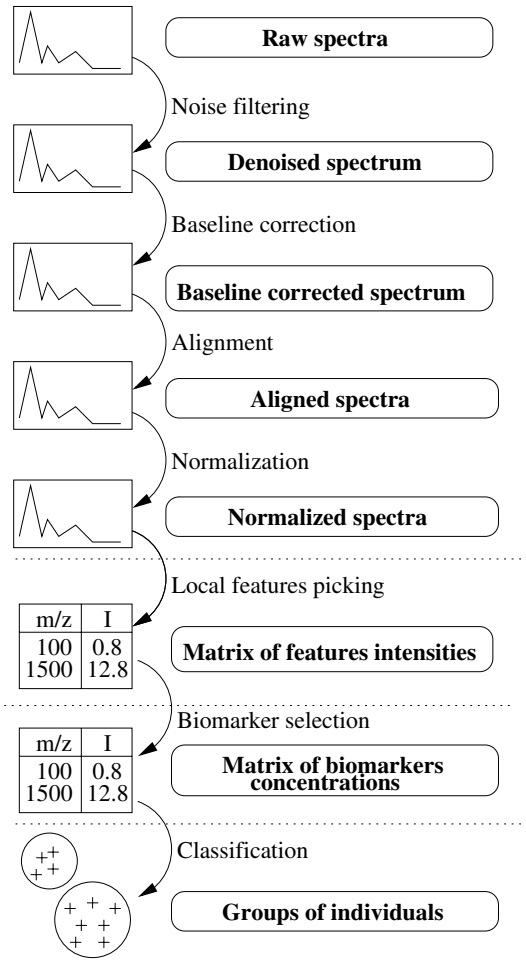

Figure 1: Workflow for MS data analysis : from raw spectra to identification of biomarkers and groups of individuals

1.1 Noise filtering

As is the case with every measure, some *random noise* adds up to the desired signal on the intensity scale. This random noise is commonly associated to electrical noise: it is therefore instrument related. MS data analysis methods benefit from development of noise filtering tools in other domains such as electronics, speech recognition... One first possible approach to handle the problem uses classical smoothing filters, defined as signal processing functions specifically designed to remove unwanted signal components. A Savitzky-Golay filter is used for instance in softwares from Bruker Daltonics [1]. A Kaiser filter is used in [13] as part of a whole processing strategy. This last article compare Savitzky-Golay filter, Kaiser filter and wavelet based approaches. Finally, de Noo et al. [14] mention the use of a Whittaker smoother.

Coombes et al. [15][16] pointed out wavelets-based methods as a really effective denoising method. The Undecimated discrete wavelet transform (UDWT) is used to separate signal from noise in spectrum, by shrinking small wavelets coefficients supposed to correspond to noise. Two possibilities are offered for this shrinking: i) *hard thresholding* replaces small coefficients below a given threshold with zeroes e.g. [2], ii) *soft thresholding* shrinks coefficients, possibly to zero. Hard thresholding might distort the signal more than soft thresholding, but it remains a simpler approach with good peformances. Similarly, Du et al. [17] propose an algorithm that simultaneously filters out noise and baseline by using *quasi-wavelet*.

This algorithm uses quasi-wavelets that have the property to make wavelet coefficients representative of a typical MS baseline equal zero: no attempt is made to model this artifact. Baseline-representative coefficients are automatically zeroed.

Time-series techniques can also be applied to this topic. Malyarenko et al. [18] develop a *resolution enhancement* through a deconvolution filter. Their algorithm are based in the time domain, that is to say using TOF and not m/z. They build a filter on *autocorrelation properties* for signal and noise.

It uses a target shape for signal (so called *signal wavelet*, without any link to wavelet transform) and defines a weight for noise in the signal. As a time-series technique, the filter is built on M earlier positions for each TOF, using autocorrelations for signal and noise. Mathematically, the filtered signal y_t is:

$$y_t = \sum_{k=1}^{M+1} a_k s_{(t-k)} \tag{2}$$

where a_k are filter coefficients defined with the help of target shape and noise's weight, s_t being the actual incoming signal.

1.2 Baseline correction

A baseline trend distorts the true signal: it is commonly admitted that this artifact arises from chemical noise. It has to be removed before intensities can be reliably read. Two main alternative approaches for this baseline correction were discussed: i) baseline is either the part of the signal remaining after features of interest have been removed (type 1 methods), or ii) some kind of smooth curve *underlying* the spectrum (type 2 methods).

An example of type 1 methods was developed by Coombes et al. [19]. This method heavily depends on detection of *peaks* as local features of interest. The algorithm first filters out peaks (i.e. potential local features of interest) and then identifies the baseline in the remaining signal by computing local minima in successive signal windows. Mantini et al. [13] describe a similar

[1] http://www.bdal.com/

approach: they identify windows in the signal containing sharp features and discard them to get an estimate of the baseline in the remaining parts of the signal.

In type 2 methods, baseline can be thought as a slow varying signal while true signal is mainly made of sharp features. Therefore, any slow varying function can be fit to the signal, like low-order polynomial. The software sold with SELDI spectrometers estimates the baseline as the *convex hull* of the signal, using a topological notion [20]. Another simple yet efficient procedure uses monotone local minimum [16], in close connection to the convex hull algorithm. Antoniadis [10] proposed a method using penalized quantile regression splines. The idea is to model the baseline with splines, adjusting them through quantile regression. A similar idea is mentioned, though not developed, in an article from Tan et al. [21].

Malyarenko et al. [18] offered a time-series perspective on the problem. In their article, baseline arises from a constant offset and a slowly decaying charge, plus some shift after detector overload events.With this approach, baseline does not have to be stricly decreasing.

Another original approach is developed by Dijkstra et al. [22]. These authors set up a mixture model to deconvolve each signal component of a spectrum, as in equation (1). An appropriate component in this mixture model corresponds to the baseline.

1.3 Alignment of spectra

Due to physical principles underlying MS, a shift on the m/z axis can appear. Although frequent instrument is required, some alignment procedure remains necessary.

In addition to the error due to instrument calibration, spectrum features can be shifted on the m/z axis during the physical process occurring in the mass spectrometer. In other words, a given peptide will be associated with different m/z labels for different spectra. For a reliable identification of local features, one need to carefully associate each feature of interest with a specific m/z. Automated recalibration procedures are required. Most if not all of these recalibration procedures are based on the identification of some obvious reliable local features (e.g. peaks). Alignment of these particular features is supposed to perform appropriate alignment for the whole spectra.

Jeffries [23] describes a method to perform this alignment. While the method requires some *anchors* to set up a new calibration function, it does not require identification of all local features of interest. One can use expected peaks from a protein known to be found in the biological sample. It builds a recalibration function that smoothly stretch the original signal so that anchors' positions corresponds to their true position.

Wong et al. [24] develop a strategy that aligns spectra also using selected local features (usually peaks from the average spectrum) as anchors. This strategy is implemented in SpecAlign[2]. By inserting or deleting some points on the m/z axis, a spectrum is locally shifted. Features from one spectrum are aligned to features from a reference spectrum. In the same flavor, Sauve and Speed [25] mention a method based on dynamic programming, resembling DNA sequences alignment methods. The advantage of such algorithm is that it does not use any reference spectrum. However, this approach requires identification of all local features used in later parts of the analysis.

Finally, an original method is proposed by Pratapa et al. [26] for alignment of repetitions of mass spectra, i.e. multiple spectra for the same sample. In this particular case, all spectrum are thought of as variations of one and same *latent spectrum*. A Hidden Markov Model (HMM) allows to infer the unknown latent spectrum. This model is adjusted on measured spectra and standart HMM algorithms are used to estimate the latent spectrum. With this strategy, local features can later be associated with a unique m/z position.

[2]http://physchem.ox.ac.uk/ jwong/specalign/

1.4 Normalization

A normalization step is necessary to ensure comparable spectra on the intensity scale. The idea is to consider that the total amount of protein is roughly the same between samples. In other terms, a small proportion of proteins are differentially expressed among all proteins in the sample. The same hypothesis is used for normalization of RNA microarray. Total Ion Current (TIC) is a useful proxy to the total amount of protein. It is related to the total number of ion collisions with the detector and output by the mass spectrometer. Each intensity in the spectrum is divided by TIC. More robust approaches can be sought: a good comparison study of these methods is available[27].

2 Finding features of interest

For MS, a spectrum obviously does not contain labels pointing to features and allowing identification of peptides or proteins. Localization (finding a m/z position) and labeling (associating a peak with a protein name) of features is therefore an essential task in MS data. We will only focus here on localization. Labeling is important for the study of precise biological mechanisms but does not bring an added value to profiling. Therefore, it falls out of the scope of this article.

After all the pre-processing work performed on spectra, the localization step represents a major shift from spectra to local features within spectra. While spectra were the basic unit of interest until this step, we now focus on features within spectra. This focus makes use of the intuitive paradigm that *local features* are the unit of interest (the best example of this being a peak). This allows a matrix representation of data at the end of this step, which is the same as for transcriptomic studies. This matrix contains an intensity for each local feature and for each spectrum. Filling this matrix with adequate values requires to define *what* a feature is, *where* it is located and *how much* signal it represents.

2.1 Main concepts for *peaks*: *what* is a peak ?

2.1.1 Local maxima

Mathematical definitions are required to automate the process of peak finding. The most commonly used definition, initially developed by Yasui et al. [28], relies on *local maxima*, that is to say a point with higher intensity than other points in a neighborhood. This is usually performed by searching for maxima in a window of defined width, repeating the process so that windows cover the entire spectrum. A list of potential peaks is returned. Too small widths will result in keeping small noisy fluctuations as peaks, while too big values will potentially filter out true peaks. Most authors use a window width of about 20 to 40 points.Even with noise reduction techniques, defined previously (section 1.1), a too large number of local maxima is found in a spectrum, not all corresponding to peptides. Other constraints have to be used to filter local maxima.

2.1.2 Filtering a list of local maxima

The large number of identified peaks when handling spectra requires strategies to select peaks that (ideally) really correspond to a peptide. Several hundreds of peaks do not offer a tractable analytical approach for later identification of biomarkers (as defined in section 3.1). Two steps can be distinguished: i) single spectrum-based filtering first, ii) the use of information of multiple spectra from a group of samples.

For a given spectrum, two main hypotheses can be made to select peaks. First, a peak must clearly stand out of noise. The most common approach to select peaks standing out of noise is to compute a *Signal to Noise Ratio* (SNR) at each position. By setting a threshold t on SNR, it is then possible to consider as a true peak only a peak that satisfies the condition $SNR > t$. Second, another arbitrary threshold can be considered, only based on the intensity of the peak. If the peak intensity is not higher than this threshold, then the peak is not retained as a true peak either. While commonly used, the assumption behind this strategy is much stronger than the SNR hypothesis. Indeed, even a small peak in a region with very little noise can be a feature of interest, corresponding to a biomarker. A better threshold can be set by using local threshold, depending on local properties of the signal. The R package PROcess [29] offers to compute the threshold $t_{PROcess} = 1.64 \cdot mad(spectrum)$ where mad is the median absolute deviation. Coombes et al. [19] use a threshold on slope of the spectrum at peak locations. If right or left slope is not higher than noise, they filter out the corresponding peaks.

The information contained in multiple spectra can be used to select true peaks. Indeed, an interesting biomarker must appear in a *large enough* proportion of collected samples to be detected as such. This proportion has to be choosed arbitrarly. We might for instance consider that a biomarker must be found in at least n spectra in the study, or equivalently in a least $p\%$ of the samples. Morris et al. [16] describe an alternative to this arbitrary threshold by identifying peaks on the *mean spectrum*. This has two advantages: it filters out some instrument noise, but also some noisy peaks that appear in a too small number of spectra. Identification of peaks only has to be performed on the mean spectrum, which avoids peak alignment and saves computing time.

Fushiki et al. [30] offer an interesting point of view on peaks. They build a kind of histogram of original peaks along the m/z axis. The peak finding problem is moved from the intensity space to the space of number of peaks, thus taking into account the information carried by multiple spectra. This allows to detect even a very small peak if it is present in most spectra.

A very important issue concerning the number of peaks retained for differential analysis (section 3) is related to the statistical power of the study. As was shown for transcriptomic studies [31], this power decreases with the number of *non-differentially expressed* genes considered in the study. This applies to proteomics as well.In general, there is a trade-off between strict filtering, that can achieve high power but miss important features, and relaxed criteria for peak selection, when features are less likely missed but power decreases.

2.1.3 Not using local maxima

Local-maxima-based methods are very popular because they are very intuitive in nature: they mainly rely on the natural concept of peak as a visible feature emerging from noise. However, there are other possible approaches that still focus on the concept of peaks but do not consider local maxima. The important concept is to look for features that *look like* peaks, and not only for features that reach higher intensities than their neighborhood. Tan et al. [21] also point out these two main strategies for peak detection, calling them *intensity-threshold-based methods* (from previous section) and *spectral matching approaches*.

Du et al. [17] describe a wavelet-based method. The basic idea is quite simple: since identifying peaks in the traditional spectrum space is a complicated work, it can prove very attractive to move the problem to another space. The Wavelet Transform (WT) offer such a space in which features of interest, thus including peaks, correspond to wavelet coefficients. Many different wavelet function exist and can be used (*Haar* wavelet, *Daubechies* wavelets...). Du et al. move the peak finding problem to the wavelet coefficients space: they look for high coefficients that cluster together for different scales, on a same position. Therefore, they offer a very original

peak searching algorithm.

Randolph et al. [32] develop a very similar idea, also based on wavelet transforms. However, while Du et al. use different wavelet scales for peak detection, they use only one scale of the wavelet decomposition. In this new wavelet space, the idea is to look for local maxima as would be done in the original spectrum.

2.2 Mass-jittering: *where* do we locate peaks ?

For all peak identification methods, the localization of the peak on the m/z axis (or equivalently on the TOF axis) is an important issue. Due to random phenomena occurring in the mass spectrometer, a peak corresponding to a given peptide will not always be associated to the same m/z, even by using alignment procedure as described in section 1.3. This is usually referred to as *mass jittering*. When comparing different spectra, peaks that are thought to correspond to the same protein must be associated with the same label. Several strategies were developed to do so.

Yasui et al. [33] develop the concept of *window of potential shift*. They base their idea on the m/z measurement uncertainty in MS experiments. A shift of about $\pm[0.1-0.2]\%$ is expected for SELDI-TOF spectra. The same kind of shift exists for MALDI-TOF experiments. A window can be defined by using this shift as a width. All peaks in the same window will have the same label and will therefore be associated with the same protein, provided spectra are well aligned. This method can be thought of as a majority vote to decide which label to use for a set of peaks. It is based on known instrument limitations. Alternatively, the use of the mean spectrum gives a reference for peak locations: Morris et al [16] propose to find all local maxima in the mean spectrum and record them as labels. A search for local maxima in a *support window* (so called *peak bins* in [16]) around these reference peaks is then performed through all individual spectra. The width of the support window should be set according to the uncertainty of the instrument.

The method of peaks average proposed by Fushiki et al. [30], described in the previous section, needs to take into account mass-jittering. To do so, the authors use gaussian kernels centered at peaks. The width of the kernel sets an *influence width* for peaks. Peaks representing the same protein, even if not at the exact same m/z position, will add up to some extent.

Another convenient approach was first proposed by Tibshirani et al. [34]. The basic idea is to group peaks that are close on the m/z axis (or the *log* of it) using classical hierarchical clustering methods. These methods output a tree which can be cut at a given depth to automatically obtain clusters of peaks associated with a unique m/z label (e.g. mean m/z in the cluster).

2.3 Estimating intensity: *how much* signal is that ?

Using peaks as features of interest requires to estimate the intensities (as a function of number of molecules hitting the detector) associated to these peaks. The most simple approach is to use peak heights as intensities. This is a very quick way to access intensities. However, as can easily be verified on a spectrum, peak heights tend to decrease with m/z, while peak widths increase. In other words, peaks are narrow for low m/z but get broader with increasing m/z values. It should be kept in mind that intensities are later compared *peak-wise*: the difference between intensities is what matters. The important notion is therefore that intensities are comparable for given m/z positions.

The fact that peak shape changes with m/z suggests another approach to evaluate intensities. Peaks are actually two-dimensional features with a height and a width. Computing the area under the peak is therefore another proxy to intensity. For a given peak, this is not a complicated step, but it requires to evaluate the peak width, adding some complexity to this step.

Both approaches are used in publications concerned with profiling, with no clear advantage for one or the other. However, it should be pointed out that no comparison study has ever been published in this field. A consistent choice for all spectra in an experiment is necessary to avoid further sources of noise.

It is also worth noting that a non-linear transformation should be applied to intensities in order to make them homoscedastic. Examples of possible transformations are cube-root or log functions.

For close peaks (on the m/z axis) corresponding to different proteins, estimating the intensity can be a really hard step. If such overlapping peaks are visible in the spectrum, several relatively recent methods [35][22][36] use mixtures of distribution and offer the possibility to deconvolve features, enhancing resolution for each peak and enabling a better intensity reading.

2.4 Non-peak based local features

Tan et al. [21] develop an original approach that do not make hypothesis on the shape of local features. The hypothesis is that a region containing information for differential analysis will display more inter-spectra variability than intra-spectrum variability. This is in direct connection with ANOVA F tests. It is then possible to detect *regions of significance* in spectra. These regions have a strong potential to contain peptides that are differentially expressed between spectra and might therefore be interesting local features. This use of F tests along spectra establishes a bridge between local-feature-based methods, as developed in this article, and functional approaches.

3 Differential analysis

3.1 Identification of biomarkers

Spectra processing outputs a list of local features. The matrix X of intensities of the MS signal for each of these features is a matrix of potential predictors, e.g. for a disease prognosis. These predictors are the *biomarkers* we look for among all identified features. In this context, a biomarker can be defined as a protein that differentiates samples from different groups. A classical grouping separates samples of individuals with a disease and samples of healthy individuals. Several methods have been used or developed in order to identify these biomarkers. A lot of methods in transcriptomics are adaptable to proteomics. We review them first. This search of biomarkers is tightly linked with the well-known *curse of dimensionality* that exist in all high-throughput methods: it will be developed in a second time.

3.1.1 Different methods

Levner et al. [37] compare several features selection methods. They distinguish *filter-based* feature selection (e.g. t-test), working in a *feature-wise* fashion, and *wrapper based* methods. While the former of these methods only focus on the identification of features that exhibit differences in intensities between groups, the latter evaluate importance of features embedded in a classification algorithm: biomarker identification is based on their ability to classify samples. Classification algorithms will be developped in section 3.2.

1. The most simple approaches consider each feature (i.e. protein) individually, for two groups (e.g. healthy VS diseased). The aim is to compare intensities between groups. Comparing mean intensities is the most intuitive tool. This corresponds to a univariate statistical test, as for instance the classical *t-test*. Each univariate test uses the H0 hypothesis: "levels of expression are the same in both groups". Each test tells whether two groups can

be distinguished based solely on information from one particular feature. Other univariate tests can be used, that use a similar logic (as pointed out in [37]). Since features are never considered simulaneously, correlations can not be taken into account and information from a set of features might be redundant.

2. Another class of methods has also been in use in high-throughput settings: dimension reduction techniques. Since we have too many features to work with regarding the number of samples, an intuitive strategy is simply to reduce the dimension. Generally, if we call Y the vector of disease states for all patients, we want to explain Y with information from X. X-based methods only use information from X, while X&Y-based methods also use information from Y. A typical example of X-based methods is Principal Component Analysis. Partial Least Square is a good example of X&Y-based methods.

3.1.2 A common issue

The curse of dimensionality, outlined in transcriptomic research, applies to proteomics. Indeed, the number of features to be tested (whether genes or proteins) is much larger than the number of samples available for the experiment. This justifies a selection of variables of interest. The number of tests to be performed is an important issue. Proteomic studies, as transcriptomic studies, confront us with the *multiple testing setting*.

In this setting, the control of the number of false positives is essential. Performing several successive univariate tests increases the global type I error rate (α risk). To deal with this issue, several strategies were set up. The basic idea is to adjust p-values of usual test statistics in order to control the global error rate. These global error rates are nicely reviewed by Dudoit et al. [38] for transcriptomic studies, with a natural extension to proteomics. The most used strategy in current papers is to control the *False Discovery Rate* (FDR), introduced by Benjamini and Hochberg [39]. FDR allows to control the proportion of false biomarkers among all proteins we label as such.

Statistical power is also an important aspect to bear in mind. The idea behind power, in this context, is our ability to detect a true biomarker, i.e. a protein that indeed is a marker for a disease. If power is too low, our ability to detect a true biomarker will be equivalently low and we will fail in our endeavor to discover new biomarkers. Power depends on type I error level that is also linked to global error rate control (e.g. FDR). While the number of non-differential (H0) genes is defined by the size of the microarray in transcriptomic studies, we should keep in mind that it is not set in advance in proteomics. A trade-off must be found between false discoveries and ability to detect a true biomarker.

At this step, one might be satisfied with the obtained list of biomarkers. These biomarkers can be identified to match specific peptides or proteins. Further biological confirmations can be sought through other experiments (focusing on one protein at a time) *via* classical (low-throughput) biological methods (ELISA, tandem MS...). Nevertheless, these biomarkers can be used to perform classification.

3.2 Classification

3.2.1 Classical classification methods

In this section, we want to stress an essential application of MS. Classification refers to the assignment of a group for each spectrum, i.e. for each sample or equivalently individual. It can be tightly integrated with biomarker discoveries, as we saw with wrapper based identification

methods. For example, given a serum sample, we want to know whether the donor belongs to a group of healthy patients or to a group of patients who are likely to develop a cancer.

A new strategy for classification appears with mass-spectrometry: *profiling*, similar to the concept of signature in transcriptomics. The idea is to use a *protein profile* instead of a unique protein concentration. A profile is basically a list of intensities for different m/z positions. Such a profile can be used in two different ways i) as a set of proteins that have to (or could later) be identified individually, ii) as a set of distinguishing features without searching biological support for these features. This can be an efficient approach in clinical applications.

In either case, a classification method is required. After a first shift from mass spectra to a matrix of intensities, a second shift from local features to biomarkers, we encounter a third shift from a matrix of intensities to a simple vector of labels identifying a group for each sample in the study (see figure).

Evaluation of a classifier performance usually requires a *learning set* and a *test set*. The classifier is *trained* on the learning set and later *validated* on the test set. Such an evaluation procedure is required to avoid *overfitting*, i.e. avoid the use of characteristics of the learning set that are not of interest but rather specific to this particular set. It is essential to evaluate the classifier on new data (test set). Classifiers performances usually will be shifted downward when applied on a test set, even when trying to avoid overfitting. This shift must also be evaluated to tell whether a classifier yields good overall performances or not. *Sensitivity* as well as *specificity* are computed to evaluate the classifier performances, based on the learning set. Then, the classifier is used on the test set. This time, the classifier is not provided with group labels. Again, classifier performances are evaluated. Data scarcity sometimes requires the use of *Leave n Out Cross Validation* (LnOCV), as *Internal validation* method. In a good review of classification in the field of proteomics, Hilario et al.[40] draw attention to the potential misemployment of such cross validation techniques. They mainly warn the use to perform part of the pretreatment process on the full dataset before a split between training and test set is decided. This can lead to artificially increased performances.

Wu et al. [41] proposed a comparison of some classification algorithms on MS data. They compare Linear Discriminant Analysis (LDA), Quadratic Discriminant Analysis (QDA), k-Nearest Neighbors (KNN), Bagging, Boosting, Random Forest (RF) and Support Vector Machines (SVM). All these methods are supervised clustering methods. LDA and QDA are equivalent to maximal likelihood discriminant rules: samples are classified so that the likelihood of such a clustering is maximal. KNN uses a vote of nearest samples (in a space defined by intensities for different features) to choose a class for a new sample. Bagging, boosting and RF are methods to aggregate classifiers as Classification And Regression Trees (CART). Finally, SVM is a projection of the n-dimensional space (where n is the number of retained biomarkers) in a higher dimensional space where a hyperplane successfully separates groups. Wu et al. show that SVM perform the best on MS data in terms of prediction error. Random Forest can also perform really well but prediction errors are larger than with SVM. LDA, though a simple method, performs quite well with a prediction error below 20%. All the methods presented here are compared on the basis of a common predictor set. However, the retained number of predictors is not optimal for all methods. In the same flavor, a comparison of classifiers in the field of transcriptomics is provided by Dudoit et al.[42].

As mentioned in section 3.1, some classifiers simultaneously perform feature selection and classification: the so-called wrapper-based identification methods. With these classifiers, a large set of potential biomarkers is used as predictors. Among these predictors, those that most help in classification are retained as biomarkers and used in the classifier, while the other predictors are discarded. Classification And Regression Trees are an example of a wrapper-based method, as well as their combination using bagging, boosting or random-forest.

3.2.2 Probability prediction for group assignment

Logistic regression is an interesting tool in the context of classification. It does not provide a group repartition for samples, but predicts the probability of belonging to a specific group among two. By setting a threshold on this probability, it is possible to infer a grouping structure from probability predictions. Interesting is the fact that several methods were developed for predictor selection in logistic regression. The simplest approach is to perform *Forward Stepwise Regression* (FSR). The idea is to sequentially select the best predictor at each adjustment step, further adjusting on residuals of the previous step. A similar approach is adapted by Yasui et al. [33]. More sophisticated methods for selection of predictors exist, jointly known as *penalized regression*. The idea behind these techniques is to constrain the regression coefficients. LASSO was proposed by Tibshirani[43] and allows to *shrink* coefficient, thereby filtering out weak predictors. LARS enhances FSR by optimizing adjustement steps and is best described by Efron et al. [44].

3.2.3 Non peak based classification

Some classification methods were developed that do not rely on the peak concept. Instead, these methods use every single point of the spectra to look for regions that can distinguish different groups of samples. Indeed, the search for part of the spectrum containing a useful information can be performed under the guidance of variables related to clinical events. This is in close connection with wrapper-based methods described above.

The pioneering study by Petricoin et al. [45] is a typical example of this strategy using every points in a spectra as potential predictors. They use a genetic algorithm (GA) that eventually selects a few points on the m/z axis in spectra as the best set of group predictors. Other authors [46][47] used GA to identify biomarkers and perform classification. GA are an example of *bioinspired approaches*Artificial Neural Networks (ANN) are another example. They offer an interesting approach as a wrapper-based method, combining predictors weighting (i.e. feature selection) and classification. ANN were adopted by Ball et al. [48] in a preliminary study to identify biomarkers of astroglial tumors.

Tong et al. [49] use the same initial idea. They consider each point of a spectrum as a potential feature of interest, in the same way as Petricoin et al., but select m/z positions using *Classification And Regression Trees* (CART) associated in a *Decision Forest* (a specific tree aggregator). A Decision Forest is a collection of a few trees with constraints on trees heterogeneity and quality. It is anticipated that the use of several trees instead of one will allow better predictions under these two constraints.

Conclusion

In this review, we divided MS data analysis into three parts: i) Pre-processing comprises several different steps and ends up with a clean signal to work with: we believe this first part will remain important even with new approaches for MS data; ii) identification of local features is a convenient way to handle MS data and ends up with a list of numerical values, maybe more tractable than an analogical signal; iii) classification and group prediction turn MS into a powerful tool for diagnosis or prognosis. There again, this part will remain useful for later researches.

Analysis of MS data still confronts the researcher with some open questions, whether in the field of pre-processing (the order or pre-processing steps, the choice of an optimal method for each step), or in the field of local features identification (the relevance of the very concept of

peak might be questionable) or in the field of the actual result analysis (e.g. the choice of a classification algorithm). Furthermore, challenging questions appear for the analysis of large datasets. Although they are not investigated in this paper, they deserve a special emphasis.

As an answer to the sometimes problematic definition of *local feature*, the functional approach shows great promises and has started to be investigated. As previously mentioned, step 1 and 3 described in this review might apply in this frame. Research on these topics therefore still deserves attention.

Acknowledgements We thank C. Truntzer, D. Pecqueur and C. Mercier for helpful discussions regarding this paper.

References

[1] Baggerly KA, Morris JS, Coombes KR. Reproducibility of SELDI-TOF protein patterns in serum: comparing datasets from different experiments. *Bioinformatics* 2004;20:777–785.

[2] Coombes KR, Morris JS, Hu J, Edmonson SR, Baggerly KA. Serum proteomics profiling–a young technology begins to mature. *Nat Biotechnol* 2005;23:291–292.

[3] Diamandis EP. Mass spectrometry as a diagnostic and a cancer biomarker discovery tool: opportunities and potential limitations. *Mol Cell Proteomics* 2004;3:367–378.

[4] Hu J, Coombes KR, Morris JS, Baggerly KA. The importance of experimental design in proteomic mass spectrometry experiments: some cautionary tales. *Brief Funct Genomic Proteomic* 2005;3:322–331.

[5] Grizzle W, Semmes O, Bigbee W, Zhu L, Malik G, Oelschlager D, Manne B, Manne U. The Need for Review and Understanding of SELDI/MALDI Mass Spectroscopy Data Prior to Analysis *Cancer Informatics* 2005;1:86–97.

[6] Albrethsen J. Reproducibility in protein profiling by MALDI-TOF mass spectrometry. *Clin Chem* 2007;53:852–858.

[7] Semmes OJ, Feng Z, Adam BL, Banez LL, Bigbee WL, Campos D, Cazares LH, Chan DW, Grizzle WE, Izbicka E, Kagan J, Malik G, McLerran D, Moul JW, Partin A, Prasanna P, Rosenzweig J, Sokoll LJ, Srivastava S, Srivastava S, Thompson I, Welsh MJ, White N, Winget M, Yasui Y, Zhang Z, Zhu L. Evaluation of serum protein profiling by surface-enhanced laser desorption/ionization time-of-flight mass spectrometry for the detection of prostate cancer: I. Assessment of platform reproducibility. *Clin Chem* 2005;51:102–112.

[8] West-Nørager M, Kelstrup CD, Schou C, Høgdall EV, Høgdall CK, Heegaard NHH. Unravelling in vitro variables of major importance for the outcome of mass spectrometry-based serum proteomics. *J Chromatogr B Analyt Technol Biomed Life Sci* 2007;847:30–37.

[9] Lin Z, Jenson SD, Lim MS, Elenitoba-Johnson KSJ. Application of SELDI-TOF mass spectrometry for the identification of differentially expressed proteins in transformed follicular lymphoma. *Mod Pathol* 2004;17:670–678.

[10] Antoniadis A, Lambert-Lacroix S, Letue F, Bigot J. Nonparametric Pre-Processing Methods and Inference Tools for Analyzing Time-of-Flight Mass Spectrometry Data. *Current Analytical Chemistry* 2007;3:127–147.

[11] Morris JS, Brown PJ, Herrick RC, Baggerly KA, Coombes KR. Bayesian Analysis of Mass Spectrometry Proteomics Data using Wavelet Based Functional Mixed Models Working Paper 22UT MD Anderson Cancer Center Department of Biostatistics 2006.

[12] Arneberg R, Rajalahti T, Flikka K, Berven FS, Kroksveen AC, Berle M, Myhr KM, Vedeler CA, Ulvik RJ, Kvalheim OM. Pretreatment of mass spectral profiles: application to proteomic data. *Anal Chem* 2007;79:7014–7026.

[13] Mantini D, Petrucci F, Pieragostino D, Boccio PD, Nicola MD, Ilio CD, Federici G, Sacchetta P, Comani S, Urbani A. LIMPIC: a computational method for the separation of protein MALDI-TOF-MS signals from noise. *BMC Bioinformatics* 2007;8:101.

[14] de Noo ME, Mertens BJA, Ozalp A, Bladergroen MR, van der Werff MPJ, van de Velde CJH, Deelder AM, Tollenaar RAEM. Detection of colorectal cancer using MALDI-TOF serum protein profiling. *Eur J Cancer* 2006;42:1068–1076.

[15] Coombes KR, Tsavachidis S, Morris JS, Baggerly KA, Hung MC, Kuerer HM. Improved peak detection and quantification of mass spectrometry data acquired from surface-enhanced laser desorption and ionization by denoising spectra with the undecimated discrete wavelet transform. *Proteomics* 2005;5:4107–4117.

[16] Morris JS, Coombes KR, Koomen J, Baggerly KA, Kobayashi R. Feature extraction and quantification for mass spectrometry in biomedical applications using the mean spectrum. *Bioinformatics* 2005;21:1764–1775.

[17] Du P, Kibbe WA, Lin SM. Improved peak detection in mass spectrum by incorporating continuous wavelet transform-based pattern matching. *Bioinformatics* 2006;22:2059–2065.

[18] Malyarenko DI, Cooke WE, Adam BL, Malik G, Chen H, Tracy ER, Trosset MW, Sasinowski M, Semmes OJ, Manos DM. Enhancement of sensitivity and resolution of surface-enhanced laser desorption/ionization time-of-flight mass spectrometric records for serum peptides using time-series analysis techniques. *Clin Chem* 2005;51:65–74.

[19] Coombes KR, Fritsche HA, Clarke C, Chen JN, Baggerly KA, Morris JS, Xiao LC, Hung MC, Kuerer HM. Quality control and peak finding for proteomics data collected from nipple aspirate fluid by surface-enhanced laser desorption and ionization. *Clin Chem* 2003;49:1615–1623.

[20] Fung ET, Enderwick C. ProteinChip clinical proteomics: computational challenges and solutions. *Biotechniques* 2002;Suppl:34–8, 40-1.

[21] Tan CS, Ploner A, Quandt A, Lehtiö J, Pawitan Y. Finding regions of significance in SELDI measurements for identifying protein biomarkers. *Bioinformatics* 2006;22:1515–1523.

[22] Dijkstra M, Roelofsen H, Vonk RJ, Jansen RC. Peak quantification in surface-enhanced laser desorption/ionization by using mixture models. *Proteomics* 2006;6:5106–5116.

[23] Jeffries N. Algorithms for alignment of mass spectrometry proteomic data. *Bioinformatics* 2005;21:3066–3073.

[24] Wong JWH, Cagney G, Cartwright HM. SpecAlign–processing and alignment of mass spectra datasets. *Bioinformatics* 2005;21:2088–2090.

[25] Sauve AC, Speed TP. Normalization, baseline correction and alignment of high-throughput mass spectrometry data 2004.

[26] Pratapa PN, Patz EF, Hartemink AJ. Finding diagnostic biomarkers in proteomic spectra *Pac Symp Biocomput* 2006:279–290.

[27] Meuleman W, Engwegen JY, Gast MCW, Beijnen JH, Reinders MJ, Wessels LF. Comparison of normalisation methods for surface-enhanced laser desorption and ionisation (SELDI) time-of-flight (TOF) mass spectrometry data. *BMC Bioinformatics* 2008;9:88.

[28] Yasui Y, McLerran D, Adam BL, Winget M, Thornquist M, Feng Z. An Automated Peak Identification/Calibration Procedure for High-Dimensional Protein Measures From Mass Spectrometers. *J Biomed Biotechnol* 2003;2003:242–248.

[29] Li X. *PROcess: Ciphergen SELDI-TOF Processing* 2005. R package version 1.12.0.

[30] Fushiki T, Fujisawa H, Eguchi S. Identification of biomarkers from mass spectrometry data using a "common" peak approach. *BMC Bioinformatics* 2006;7:358.

[31] Lee MLT, Whitmore GA. Power and sample size for DNA microarray studies. *Stat Med* 2002;21:3543–3570.

[32] Randolph TW, Mitchell BL, McLerran DF, Lampe PD, Feng Z. Quantifying peptide signal in MALDI-TOF mass spectrometry data. *Mol Cell Proteomics* 2005;4:1990–1999.

[33] Yasui Y, Pepe M, Thompson ML, Adam BL, Wright GL, Qu Y, Potter JD, Winget M, Thornquist M, Feng Z. A data-analytic strategy for protein biomarker discovery: profiling of high-dimensional proteomic data for cancer detection. *Biostatistics* 2003;4:449–463.

[34] Tibshirani R, Hastie T, Narasimhan B, Soltys S, Shi G, Koong A, Le QT. Sample classification from protein mass spectrometry, by 'peak probability contrasts'. *Bioinformatics* 2004;20:3034–3044.

[35] Lange E, Gropl C, Reinert K, Kohlbacher O, Hildebrandt A. High accuracy peak-picking of proteomics data using wavelet techniques *Pac. Symp. Biocomput* 2006;11:243–254.

[36] Noy K, Fasulo D. Improved model-based, platform-independent feature extraction for mass spectrometry. *Bioinformatics* 2007;23:2528–2535.

[37] Levner I. Feature selection and nearest centroid classification for protein mass spectrometry. *BMC Bioinformatics* 2005;6:68.

[38] Dudoit S, Shaffer J, Boldrick J. Multiple Hypothesis Testing in Microarray Experiments *Statistical Science* 2003;18:71–103.

[39] Benjamini Y, Hochberg Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing *Journal of the Royal Statistical Society. Series B (Methodological)* 1995;57:289–300.

[40] Hilario M, Kalousis A, Pellegrini C, Mueller M. Processing and classification of protein mass spectra. *Mass Spectrom Rev* 2006;25:409–449.

[41] Wu B, Abbott T, Fishman D, McMurray W, Mor G, Stone K, Ward D, Williams K, Zhao H. Comparison of statistical methods for classification of ovarian cancer using mass spectrometry data. *Bioinformatics* 2003;19:1636–1643.

[42] Dudoit S, Fridlyand J, Speed T. Comparison of Discrimination Methods for the Classification of Tumors Using Gene Expression Data. *Journal of the American Statistical Association* 2002;97:77–88.

[43] Tibshirani R. Regression Shrinkage and Selection via the Lasso *Journal of the Royal Statistical Society. Series B (Methodological)* 1996;58:267–288.

[44] Efron B, Hastie T, Johnstone I, Tibshirani R. Least angle regression *Annals of Statistics* 2004;32:407–499.

[45] Petricoin EF, Ardekani AM, Hitt BA, Levine PJ, Fusaro VA, Steinberg SM, Mills GB, Simone C, Fishman DA, Kohn EC, Liotta LA. Use of proteomic patterns in serum to identify ovarian cancer. *Lancet* 2002;359:572–577.

[46] Koomen JM, Shih LN, Coombes KR, Li D, chun Xiao L, Fidler IJ, Abbruzzese JL, Kobayashi R. Plasma protein profiling for diagnosis of pancreatic cancer reveals the presence of host response proteins. *Clin Cancer Res* 2005;11:1110–1118.

[47] Li L, Umbach DM, Terry P, Taylor JA. Application of the GA/KNN method to SELDI proteomics data. *Bioinformatics* 2004;20:1638–1640.

[48] Ball G, Mian S, Holding F, Allibone RO, Lowe J, Ali S, Li G, McCardle S, Ellis IO, Creaser C, Rees RC. An integrated approach utilizing artificial neural networks and SELDI mass spectrometry for the classification of human tumours and rapid identification of potential biomarkers. *Bioinformatics* 2002;18:395–404.

[49] Tong W, Xie Q, Hong H, Shi L, Fang H, Perkins R, Petricoin EF. Using decision forest to classify prostate cancer samples on the basis of SELDI-TOF MS data: assessing chance correlation and prediction confidence. *Environ Health Perspect* 2004;112:1622–1627.

2.3 Méthodologie statistique

La recherche de biomarqueur constitue un deuxième temps d'analyse en spectrométrie de masse. Le prétraitement du signal étant réalisé, l'analyse de données à proprement parler peut être menée. Ce temps d'analyse intervient alors que les données changent de format. La spectrométrie de masse délivre en effet les spectres décrits précédement, l'ensemble des étapes du prétraitement a permis d'obtenir une matrice d'intensités.

La détection des pics dans les spectres permet d'associer à chaque individu i, pour chaque pic j, une intensité I_{ij}. Avec n individus et m pics identifiés, on construit donc une matrice de travail $\boldsymbol{I}$ de taille $n \times m$. Chaque ligne contient les intensités lues pour un individu i. Chaque colonne contient les intensités lues pour un pic j. Les pics identifiés sont censés correspondre chacun à un peptide, lui-même correspondant à une protéine. La donnée supplémentaire du statut Y_i de l'individu vis à vis d'une variable d'intérêt est aussi requise. Classiquement, on s'intéresse à la situation $Y_i = 0$ pour un individu sain et $Y_i = 1$ pour un individu atteint d'une maladie étudiée. Le vecteur $\boldsymbol{Y}$ contient les statuts de tous les individus de l'étude. Le couple $(\boldsymbol{I}, \boldsymbol{Y})$ constitue la base de travail pour l'analyse statistique et la recherche de biomarqueurs.

Parmi l'ensemble des pics trouvés dans les spectres, la question centrale est de mettre en évidence les pics présentant un différentiel d'expression entre groupes. Ces pics correspondent à des protéines différentielles (Differentially Expressed Proteins, DEP), contrairement aux pics ne présentant pas de différentiel d'expression, correspondant aux protéines non différentielles (Non-Differentially Expressed Proteins , NDEP). La mise en évidence d'un *pic différentiel* est fondée sur un *test statistique*.

L'ensemble des notions utiles à la bonne compréhension du travail présenté dans les chapitres 3 et 4 est présenté ici, afin de préciser le contexte de ces différentes notions et d'en présenter une rapide revue.

2.3.1 Généralités et notations

L'hypothèse d'un différentiel d'expression entre groupes pour une protéine est testée grâce à un test statistique arbitraire. Ce test doit permettre de tester une hypothèse binaire, en fournissant une réponse binaire. La plus fréquente hypothèse est l'égalité de la moyenne par groupe des concentrations d'une protéine. Pour une protéine j, si on note μ_0 la concentration moyenne dans le groupe 0 et μ_1 la concentration moyenne dans le groupe 1, on peut alors définir deux hypothèses exclusives comme dans l'équation (2.1), où H_0 représente l'hypothèse nulle et H_1 l'hypothèse alternative.

$$\begin{cases} H_0 & : \mu_0 = \mu_1 \\ H_1 & : \mu_0 \neq \mu_1 \end{cases} \tag{2.1}$$

Tout test doit permettre de conclure au rejet ou au non rejet de H_0 (comme toujours en statistique, le rejet de l'hypothèse nulle ne donne pas de renseignement sur l'hypothèse alternative valide). La réalité étant elle aussi binaire (protéine différentielle ou non), quatre situations peuvent être envisagées, comme présenté dans le tableau 2.1. Une conclusion juste peut être portée, donnant lieu soit à un vrai positif (VP) soit à un vrai négatif (VN). A l'inverse, une conclusion

Réalité	Acceptation de H_0	Rejet de H_0
Protéine non différentielle	VN	FP
Protéine différentielle	FN	TP

TAB. 2.1 – Situations possibles dans le cas d'un test binaire. VP = Vrai Positif, FP = Faux Positif, VN = Vrai Négatif, FN = Faux Négatif

fausse peut être portée, donnant lieu soit à un faux positif (FP) soit à un faux négatif (FN). On définit la *sensibilité* comme le rapport $\frac{VP}{VP+FN}$ et la *spécificité* comme le rapport $\frac{VN}{VN+FP}$. La sensibilité traduit la capacité du test à mettre en évidence les protéines différentielles. La spécificité traduit la capacité à mettre en évidence les protéines non différentielles. On définit aussi la valeur prédictive positive (VPP) comme le rapport $VPP = \frac{VP}{VP+FP}$, et la valeur prédictive négative $VPN = \frac{VN}{VN+FN}$, qui sont des propriétés des tests eux-mêmes et moins des marqueurs dans la population.

La conclusion du test est faite pour un niveau de risque de première espèce fixé, correspondant à la probabilité de rejeter H_0 alors que cette hypothèse est vraie. De façon formelle, ce risque est défini comme la probabilité $p(FP) = p(rH0|H0)$, où $rH0$ représente l'événement *rejet de* H_0. Ce risque est traditionnellement noté α, conduisant à l'acceptation de H_0 tant que la valeur du test est inférieure au $(100-\alpha)$ème percentile de la distribution cumulée du test, pour un test unilatéral, ou au $(100-\frac{\alpha}{2})$ème percentile pour un test bilatéral. On a donc rejet de l'hypothèse nulle si la valeur t prise par le test est telle que $p_{value} = 1 - F_0(t) < \alpha$, avec F_0 la fonction de densité cumulée de la statistique du test sous H_0.

Ce risque est un paramètre statistique essentiel d'un test, puisqu'il précise le risque de faire une fausse découverte, c'est-à-dire de conclure erronément à un différentiel d'expression pour une protéine qui présente le même profil de concentration chez tous les individus. Etant nécessaire à la conclusion du test, il est systématiquement contrôlé. Un risque α de 5% est traditionnellement accepté.

L'autre risque encouru lors de la réalisation d'un test, c'est-à-dire la probabilité d'accepter erronément l'hypothèse nulle, dénommé risque de deuxième espèce, est bien moins souvent contrôlé. Il s'agit formellement de $p(FN) = p(aH_0|H_1)$ où aH_0 désigne l'événement *accepter* H_0. Ce risque est traditionnellement noté β. Il est en pratique plus difficile à définir et calculer, dans la mesure où il dépend de l'hypothèse alternative H_1, non univoque.

La puissance statistique se définit comme le complément à 1 de ce risque de deuxième espèce. Elle s'écrit formellement $p(VP) = p(rH_0|H_1) = 1-\beta$ et représente donc la probabilité de détecter une protéine différentielle. Elle est nécessairement aussi dépendante de l'hypothèse alternative, qui doit être définie pour permettre son calcul. L'équation 2.2 met en évidence la place de cette hypothèse alternative et le lien avec le risque α accepté, avec F_0^{-1} la distribution cumulée inverse (telle que $F_0^{-1}(x)$ donne le x-ème quantile de F_0). La fonction F_0 est obligatoirement définie pour la réalisation du test et les conclusions s'interprètent en fonction de α. En revanche, la réalisation d'un test ne demande pas de spécifier la distribution F_1 du test sous l'hypothèse alternative, ce qui rend en pratique la puissance (ou le risque de deuxième espèce) largement moins étudiée.

$$1 - \beta = 1 - F_1(F_0^{-1}(1-\alpha)) \tag{2.2}$$

L'abord le plus fréquent de la puissance consiste à considérer non pas la puissance elle-même mais la taille d'expérience nécessaire pour garantir une puissance minimum fixée arbitrairement, en général 80%, étant donnée une hypothèse alternative choisie et un risque de première espèce. Ce travail de calibration expérimentale implique donc un regard sur l'expérience dans lequel la puissance statistique est une donnée accessoire à contrôler.

La calibration expérimentale demande de fixer des paramètres essentielles de l'expérience, qui sont en fait des déterminants majeurs de la puissance statistique. Le plus classique est la taille d'expérience, c'est-à-dire le nombre de sujets n à inclure. Les expériences de comparaison de groupes font aussi intervenir un paramètre de répartition entre groupes. Au delà de ces paramètres directement contrôlables par l'expérimentateur, deux autres paramètres propres aux variables étudiées interviennent comme déterminants de la puissance. Il s'agit de la différence de moyenne de concentrations entre groupes Δ et de l'écart type de la distribution de concentrations σ, pour chaque protéine (on fait implicitement l'hypothèse d'égalité des écarts types pour la distribution des concentrations sous H_0 et H_1). Pour une valeur de σ fixée, plus Δ est grand et plus il est facile de détecter ce différentiel d'expression. Réciproquement, pour une valeur de Δ fixée, plus σ est grand et plus il est difficile de mettre en évidence un différentiel d'expression. Ce raisonnement à une quantité fixée et l'autre variable conduit à préférer l'utilisation du rapport $\frac{\Delta}{\sigma}$, appelé ici *effet différentiel*. Celui-ci n'est pas contrôlable par l'expérimentateur mais conditionne le schéma expérimental.

A titre d'exemple, la figure 2.2 illustre un test basé sur une loi normale et les quantités associées aux risques de première et de deuxième espèce. La distribution du test sous H_0 est ici une loi normale $N(0, 2)$, la distribution du test sous H_1 est aussi une loi normale $N(4, 2)$. Dans ce cas de figure, on a $\Delta = \mu_1 - \mu_0 = 4$ (avec μ_k la concentration moyenne sous l'hypothèse k) et $\sigma = 2$, ce qui correspond donc à un effet différentiel $\frac{\Delta}{\sigma} = 2$. Cet ampleur d'effet différentiel, bien que trop grande pour être réaliste dans le contexte, permet de bien visualiser les différents risques et leur lien avec Δ et *sigma*. La taille d'expérience n et la répartition entre groupes interviennent surtout au niveau de l'estimation de σ.

Les tests classiques de comparaison de tendance centrale (typiquement, la moyenne) peuvent être utilisés, de type test de Student ou de Wilcoxon pour une comparaison de deux groupes paramétrique ou non paramétrique, ou ANOVA pour une comparaison de plus de deux groupes. Ces tests de comparaison de tendance centrale sont les plus fréquemment utilisés, mais quelques approches visant à comparer les variances entre groupes ont aussi été développées [27]. Celles-ci ne peuvent pas être appliquées dans le cadre d'un test diagnostique, mais elles peuvent contribuer à la détection de nouveaux biomarqueurs diagnostiques.

Les données de spectrométrie de masse requièrent la réalisation de nombreux tests, en pratique un pour chaque pic trouvé dans les spectres. Ces tests peuvent être réalisés dans le cadre d'une approche *univariée* ou *multivariée*. L'approche univariée consiste à réaliser chaque test indépendamment des autres, ce qui ne permet pas de prendre en compte les corrélations existant entre les variables testées. On réalise implicitement la modélisation de la variable réponse ($\boldsymbol{Y}$) en fonction d'une unique variable prédictive (X_j) pour chaque test. L'approche multivariée modélise la variable réponse en fonction de l'ensemble des variables prédictives ($\boldsymbol{X}$). Elle permet ainsi de prendre en compte les corrélations entre variables.

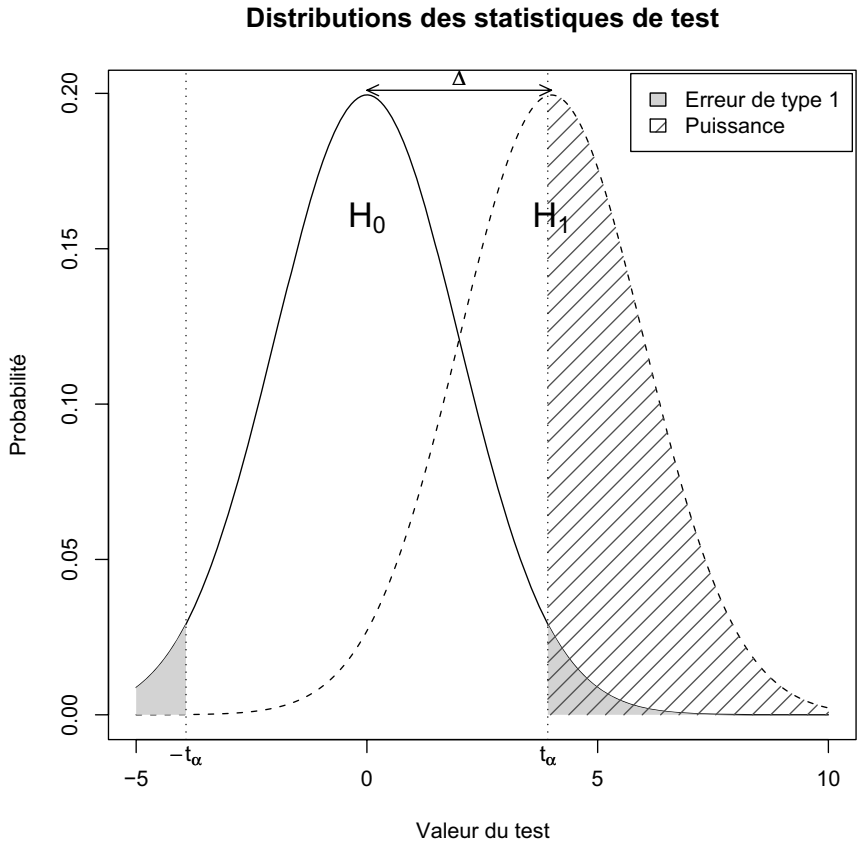

FIG. 2.2 – Fonctions de densité des deux hypothèses gaussiennes. Toute valeur du test supérieure à la limite t_α (ou inférieure à $-t_\alpha$) conduit au rejet de H_0. L'aire sous la courbe de l'hypothèse nulle à droite de t_α et à gauche de $-t_\alpha$ correspond au risque de première espèce ; l'aire sous la courbe de l'hypothèse alternative à droite de t_α (et à gauche de $-t_\alpha$, ici négligeable) correspond à la puissance.

Le choix entre l'approche univariée et l'approche multivariée n'est pas évident. La meilleure prise en compte des corrélations entre variables prédictives suggère *a priori* l'usage d'une approche multivariée. Toutefois, les modèles multivariées ne sont le plus souvent pas ajustables lorsque le nombre de variables m est supérieur au nombre d'observations n (i.e. d'individus inclus dans l'étude), alors que le nombre de pics identifiés en spectrométrie de masse est classiquement bien plus grand que le nombre de spectres obtenus (en pratique, m est de l'ordre de quelques centaines, alors que le nombre d'individus est de l'ordre de quelques dizaines). Plusieurs solutions ont été proposées :

- la solution couramment rencontrée consiste à réaliser initialement un ensemble de tests univariés pour ne garder que les variables ayant un test univarié positif dans une étape ultérieure basée sur une approche multivariée [28]. Cette approche ne résout pas le problème des corrélations,
- l'Analyse en Composante Principale permet une réduction du nombre de variables, mais implique la construction de nouvelles variables composites peu intuitives pour la compréhension des phénomènes biologiques,
- une solution plus élégante est apportée par les modèles dits *pénalisés*, de type Lasso [29], ou plus généralement dits de *Least Angle Regression* (LAR) [30]. Toutefois, ces modèles rendent difficile le contrôle du risque de première espèce, comme illustré dans la section 3.3.3.

Un choix est donc nécessaire entre ces différentes approches, en gardant à l'esprit les avantages et limitations de chacune. En pratique, deux approches ont été adoptées dans le cadre de cette thèse. L'étude de puissance présentée dans le chapitre 3 utilise une approche univariée basée sur la répétition d'un même test (test de nullité du coefficient de régression d'un modèle logistique à une variable) pour les différentes variables et une approche multivariée basée sur un modèle de régression logistique pénalisée, de type Lasso. Chacune de ces approches exige en pratique la réalisation de nombreux tests.

2.3.2 Particularités des tests multiples

Les données de spectrométrie de masse sont typiquement représentées par le couple $(\boldsymbol{I}, \boldsymbol{Y})$, décrit dans la section 2.3.1. Le nombre de tests à réaliser est égal au nombre de pics trouvés m. Cette multiplicité des tests est compliquée par la taille des expériences en spectrométrie de masse, avec $n < m$ comme mentionné dans la section précédente. Cette situation, qu'on retrouve dans les données issues de la transcriptomique, a conduit la communauté à définir le concept de *malédiction de la dimension* (*curse of dimensionality*). Cette expression traduit les difficultés rencontrées dans l'analyse statistique lorsque le nombre d'individus inclus dans l'étude est largement inférieur au nombre de variables étudiées.

Cette situation s'intègre dans la problématique statistique des *tests multiples*. Cette problématique n'est pas apparue avec les technologies haut-débit, mais elle a été renouvelée avec leurs données. Le problème principal rencontré est de contrôler les risques de première et deuxième espèce lorsqu'on réalise plus d'un test. La réflexion développée dans la littérature s'articule autour du tableau 2.2. Ce tableau généralise le tableau 2.1.

L'habituelle correction de Bonferroni stipule le besoin de corriger le risque de première espèce accepté par le nombre de tests, afin de garantir le risque de première espèce de l'ensemble des

	Conclusion		
	Accepter H0	Rejeter H0	Total
Réalité			
NDEP	U	V	m_0
DEP	T	S	m_1
Total	m-R	R	m

TAB. 2.2 – Répartition des conclusions en situation de tests multiples

tests réalisés pour un même jeu de données. Cette correction définit ainsi un nouveau risque par test α_1, qui permet de garantir que le risque global de l'expérience α_m (pour m tests) est bien contrôlé au niveau souhaité, selon la relation de l'équation (2.3).

$$\alpha_1 = \frac{\alpha_m}{m} \tag{2.3}$$

Dans le formalisme des tests multiples, cette stratégie contrôle le *Family Wise Error Rate* (FWER) à un niveau α_m, selon la relation de l'équation 2.4.

$$FWER = p(V \geq 1) = 1 - p(V = 0) \leq \alpha_m \tag{2.4}$$

Toutefois, les expériences de transcriptomique initialement, puis de protéomique, ont mis en évidence la séverité d'une telle approche lorsque plusieurs centaines (voire milliers) de tests sont réalisés. Le risque α individuel acceptable est alors rendu tellement faible que le rejet de H_0 devient difficile, ce qui implique une réduction de puissance. L'équation 2.2 traduit ce problème en terme de puissance (F_0 et F_1 étant des fonctions de densité cumulées, elles sont croissantes monotones). Cette approche FWER est aussi rendue peu adaptée par la problématique même de l'expérience. La spectrométrie de masse (tout comme les puces à ADN), dans la recherche de biomarqueurs, est utilisée pour réaliser un criblage (le *screening* anglo-saxon). Les conclusions des expériences de spectrométrie de masse doivent donc ensuite être validées. Elles suggèrent des pistes de réflexion et des hypothèses de recherche, ce qui rend le nombre de faux positifs acceptables plus importants que dans les expériences classiques. Des approches alternatives au FWER ont donc été développées.

Les premiers à avoir proposé une réflexion différente dans ce domaine sont Benjamini et Hochberg [31]. Leur constat initial est simple : dans une expérience haut débit, on s'intéresse avant tout aux rejets de l'hypothèse nulle. La population de tests d'intérêt est donc celle pour laquelle on rejette H_0. Ce constat a conduit Benjamini et Hochberg à utiliser la proportion appelée *False Discovery Rate*, décrite dans l'équation 2.5.

$$FDR = \begin{cases} E(\frac{V}{R}) & R > 0 \\ 0 & R = 0 \end{cases} \tag{2.5}$$

Benjamini et Hochbert décrivent une procédure pratique pour obtenir un FDR contrôlé à un niveau arbitraire α_m, c'est-à-dire permettant de garantir un FDR inférieur à α_m. Cette procédure, notée BH95, est basée sur la liste ordonnée des p-values classiques des tests réalisés, notées $p_{(1)}, \ldots, p_{(j)}, \ldots, p_{(m)}$. Le principe est de trouver le nombre k défini dans l'équation (2.6).

On rejette l'hypothèse nulle pour les tests $j \leq k$.

$$k = \max_{j \in [1;m]} \left\{ p_{(j)} \leq \frac{j}{m} \cdot \alpha_m \right\} \tag{2.6}$$

De façon analogue à l'approche FWER, il s'agit de trouver le risque α_1 pour chaque test individuel garantissant un risque de première espèce global α_m contrôlé. Cette procédure de contrôle du FDR fournit des valeurs de α_1 supérieures à celles fournies par les procédures FWER, permettant ainsi d'obtenir une meilleure puissance. Contrairement à la correction de Bonferroni, il faut toutefois noter que l'approche de Benjamini et Hochberg ne prend pas en compte la dépendance entre les tests, i.e. la structure de corrélation des variables.

En pratique, Benjamini, Hochberg et Kling [32] ont montré que la simple procédure BH95 permet de prendre en compte des structures de corrélation simples de type corrélation positive égale entre variables sous H_0. Les auteurs introduisent en outre des méthodes de rééchantillonage pour estimer la distribution empirique des p-values et en déduire des estimations plus précises du FDR. Une approche supplémentaire a été proposée par Benjamini et Yekutieli, avec une procédure spécifique pour les situations comprenant des corrélations négatives [33]. Une synthèse des procédures de rééchantillonage est proposée dans un article de Dudoit et al. [34].

Storey et al.[35, 36] ont par la suite développé le concept de *positive FDR*, en introduisant deux points importants :

- une définition du FDR conditionnée par l'existence d'au moins un rejet de H_0, *i.e.* par l'événement $R > 0$
- une estimation de la proportion π_0 de variables (gènes ou protéines par exemple) sous H_0, c'est à dire non différentiellement exprimées

Les auteurs montrent que l'approche BH95 utilise implicitement $\pi_0 = 1$, ce qui la rend trop conservative. Ils proposent aussi le concept de *q-value*, comme une mesure du pFDR associé au rejet d'une hypothèse j avec une p-value p_j.

Les multiples approches proposées pour le FDR et son contrôle ont conduit à des études comparatives [37, 38, 39, 40] visant à juger la qualité de ces méthodes en terme de contrôle effectif de la proportion de faux positifs et en terme de conservatisme (et donc de puissance). L'approche de la q-value, basée sur le pFDR de Storey et al., se révèle ainsi la plus sensible parmi les méthodes évoquées ici, mais affiche en contrepartie un FDR légèrement sous-estimé par rapport à la valeur cible. Cette approche q-value a été utilisée dans l'article présenté dans le chapitre 3.

Toutes ces approches ont pour but de déterminer le niveau de risque individuel α_1 acceptable pour garantir un risque global contrôlé α_m. Elles définissent ainsi toutes (implicitement ou non) un lien entre ces risques individuel et global. De façon générique, on peut définir la fonction l comme dans l'équation (2.7). La bonne compréhension de ce lien non évident est un point important de l'article présenté en chapitre 4. Il permet en effet aussi le lien entre niveau de contrôle du FDR et puissance individuelle, sujet de l'étude de ce chapitre.

$$\alpha_1 = l(\alpha_m) \tag{2.7}$$

En utilisant $l(\alpha_m)$ à la place de α dans l'équation (2.2), on peut évaluer la puissance indi-

viduelle en situation de test multiple. La fonction l a toutefois indirectement fait l'objet de nombreux travaux. Pawitan et al. [41] ont proposé une écriture du FDR reprise dans l'équation 2.8 en fonction de la valeur limite c du test. Cette valeur est par définition liée à α_1 selon la relation $\alpha_1 = 1 - F_0(c)$, avec toujours F_0 la fonction de densité cumulée du test sous H_0. F se définit comme un mélange de F_0 et de F_1, fonction de densité cumulée sous l'hypothèse alternative H_1, selon la relation $F = \pi_0 \cdot F_0 + (1 - \pi_0) \cdot F_1$, où π_0 est la proportion de variables sous H_0.

$$FDR_{Pawitan} = \frac{\pi_0 \cdot (1 - F_0(c))}{1 - F(c)} \tag{2.8}$$

De façon similaire, d'autres écritures du FDR comme une fonction du risque de première espèce ont été proposées [42, 43, 44]. Dans chaque cas, cette écriture demande de spécifier la distribution du test sous l'hypothèse alternative. L'idée est ainsi d'estimer le nombre de vrais positifs S, afin d'inclure cette estimation dans le calcul du FDR. On remarque ainsi deux choses :

- la quantité FDR dépend fondamentalement de l'hypothèse alternative et n'a de sens qu'avec la donnée de cette hypothèse, comme on le commentera dans le chapitre 4,
- le lien l entre α_1 et FDR est basé sur une hypothèse alternative, rendant son expression non triviale.

Les approches FWER ou FDR permettent de généraliser la notion d'erreur de type 1. La notion de puissance individuelle fait à son tour face aux mêmes critiques en situation de tests multiples. Une contrainte sur cette puissance représente en effet un haut niveau d'exigence. De la même façon que le risque de première espèce peut être étendu en situation de tests multiples, une nouvelle définition de la puissance en situation de tests multiples peut être proposée. La question statistique posée et le type d'étude peuvent guider le choix de cette nouvelle définition. Par exemple, une expérience menée pour identifier un biomarqueur quelconque parmi un ensemble supposé n'exige pas l'identification de tous les biomarqueurs de cet ensemble. Inversement, l'utilisation de la spectrométrie de masse pour la détection d'un biomarqueur connu fait apparaître la nécessité de contrôler précisément la puissance individuelle.

Cette nouvelle définition de la puissance n'est pas évidente. De la même façon qu'en situation de test unique, la puissance n'est pas aussi étudiée que le risque de première espèce. Il n'existe donc pas de définition consensuelle de la puissance lorsque plus d'un test est réalisé. En reprenant le tableau 2.2, on peut envisager plusieurs perspectives sur la probabilité de déclarer comme telle une protéine différentiellement exprimée :

1. une approche symétrique à celle du FDR, avec la quantité $1 - \frac{T}{m-R}$
2. une approche complémentaire du FDR, avec la quantité $\frac{S}{R}$
3. une approche centrée sur les biomarqueurs et non les découvertes, avec la quantité $\frac{S}{m_1}$
4. différentes approches visant à contrôler la quantité S

L'approche symétrique du FDR, notée dans la littérature False non-Discovery Rate (FNDR) [41] ou Miss Rate (MR) [45], est très différemment appréciée. Pawitan [41] utilise un exemple pour argumenter contre l'utilisation du FNDR, alors que Taylor [45] argumente en faveur du Miss Rate. En considérant une valeur limite du test c, le Miss Rate tel que défini par Taylor est calculé uniquement dans une fenêtre de valeurs du test $[c_0; c]$ et non pour l'ensemble des valeurs $[0, c]$, rendant d'après les auteurs la notion plus pertinente. Le contrôle conjoint du FDR et du MR

permet de fournir une idée grossière de la puissance, par l'intermédiaire d'un contrôle de V et T à marges fixées. La traduction concrète du MR en terme de vrais positifs ne permet toutefois pas une utilisation pertinente comme généralisation de la puissance.

Une quantité complémentaire du FDR est aisément manipulable lorsque le FDR lui-même est contrôlé. Toutefois, le concept de puissance est assimilable au concept de sensibilité et non au concept de valeur prédictive positive ($\frac{TP}{TP+FP}$). Il paraît donc plus pertinent de définir une puissance globale ayant pour référence le nombre total de biomarqueurs potentiels m_1.

L'approche centrée sur les biomarqueurs semble la plus pertinente. Elle prend pour ensemble de tests d'intérêt les tests appliqués aux protéines différentielles, c'est à dire m_1 tests. Le False Negative Rate, défini comme $\frac{T}{m_1}$, a été discuté par Norris [46] et fournit une extension naturelle du risque de deuxième espèce en situation de tests multiples. Son complément à m_1, la quantité $\frac{S}{m_1}$, est la quantité la plus proche du concept de sensibilité ou de puissance individuelle. L'espérance de cette quantité $\frac{E(S)}{m_1}$ est communément appelée *puissance moyenne*. Elle fournit une estimation de la capacité moyenne à rejeter l'hypothèse nulle pour une protéine différentielle. Sous conditions de puissances individuelles identiques pour toutes les protéines différentielles et d'indépendance de ces puissances individuelles, l'espérance de S s'écrit $E(S) = m_1 \cdot (1 - \beta)$, comme suggéré par Lee [47]. On a donc, sous ces conditions, une équivalence entre la puissance individuelle et la puissance moyenne.

Ces trois premières approches abordent le problème de la puissance de façon collective, ou moyenne, en fournissant des estimations de proportions de protéines dans une classe ou une autre. La question même de l'expérience, dans les études haut-débit, suggère toutefois d'autres approches. Comme mentionné plus haut, l'objectif d'une étude peut être la mise en évidence d'*au moins un biomarqueur*. La quantité d'intérêt est alors $p(S > 0)$, appelée ici puissance relaxée. Son contrôle fournit des garanties sur la capacité de l'étude à découvrir au moins un biomarqueur (étant données les caractéristiques attendues de ce biomarqueur). On peut généraliser cette définition avec l'expression $p(S > k)$, où k représente un nombre minimal de découvertes à faire, ou encore $p(S > \frac{k}{m_1})$ afin de raisonner en proportion de biomarqueurs attendus. Tsai [48] propose une écriture théorique de la quantité $p(S > k)$, fournissant un lien entre puissance individuelle et cette définition relaxée.

On retient dans ce travail deux notions de puissance : la puissance individuelle, avec sa généralisation possible à la puissance moyenne $\frac{E(S)}{m_1}$, et la puissance relaxée $p(S > 0)$.

Le mauvais contrôle de la puissance, qu'elle qu'en soit la définition, dans la majorité des études haut-débit réalisées, a des répercussions importantes. D'une part, la capacité de découverte des technologies haut-débit est limitée, non par défaut des technologies, mais par insuffisance des schémas expérimentaux mis en place. D'autre part, la reproductibilité apparente des études, en terme de biomarqueurs identifiés, est faible. Ein-Dor [49, 50] a mis en lumière ce phénomène, fournissant des arguments solides en faveur de schémas expérimentaux pensés pour un niveau de puissance voulu.

Chapitre 3

Simulations pour l'étude de la puissance

3.1 Problématique

La spectrométrie de masse offre une rapidité d'analyse du protéome inédite. Les études suggérées par une telle technologie sont séduisantes. Malgré leur nombre dans la littérature actuelle, aucune application clinique n'a encore vu le jour. Des raisons de ce relatif échec actuel ont été suggérées par Ein-Dor[49, 50]. Un manque de puissance fait partie des raisons invoquées. Les études classiques reposent en effet sur quelques dizaines d'individus et ne sont pas conçues pour garantir une puissance voulue. Or, la notion même de puissance statistique se traduit comme la capacité à détecter les biomarqueurs dans les jeux de données.

La puissance statistique, dans le contexte de la détection d'effets différentiels entre protéines, est une notion majeure à contrôler pour garantir la qualité et la pertinence des études menées. Ses déterminants classiques sont notamment la taille d'expérience et la répartition entre les groupes expérimentaux, deux paramètres fondamentaux de tout plan d'expérience. Toutefois, on ne peut pas parler de puissance statistique sans faire référence, même implicite, à une hypothèse alternative. Ces trois paramètres ont un impact important sur la puissance, mais aucune étude de spectrométrie de masse ne les a jusqu'ici pris en compte. Une étude de la puissance attendue dans les expériences de spectrométrie de masse apparaît donc comme essentielle pour permettre une validation des études déjà menées, ainsi que pour aider à la construction des plans d'expérience.

L'abord statistique des données a déjà fait apparaître des difficultés dans l'analyse des données de spectrométrie. A ces difficultés d'ordre méthodologique s'ajoutent des difficultés liées aux échantillons biologiques, dues à l'absence de référence disponible lorsque sont manipulés des fluides biologiques complexes. On ne dipose donc pas d'échantillon dans lequel serait précisées les listes de protéines différentielles et non différentielles. L'évaluation sur des données expérimentales classiques n'est donc pas possible. Pour surmonter cet obstacle, deux approches sont possibles :

- l'approche *in vitro* demande de manipuler des échantillons biologiques pour y introduire des peptides de référence de façon différentielle, afin de pouvoir évaluer la capacité des méthodes disponibles à les détecter,
- l'approche *in silico* demande quant à elle de générer des spectres de masse, c'est à dire de

simuler une expérience de spectrométrie de masse, dans laquelle sont aussi introduites des variables différentielles.

Les deux approches offrent avantages et inconvénients. L'étude *in silico* est détaillée dans ce chapitre. Les données d'une étude *in vitro* basée sur le même schéma expérimental devraient être disponibles prochainement et permettront à terme une validation des conclusions de l'étude *in silico*. L'étude des deux types d'expérience repose sur les mêmes étapes. La méthodologie développée ici sera ainsi étendue aux nouvelles données *in vitro*.

Afin de permettre de bien comprendre les déterminants de la puissance statistique en spectrométrie de masse, on propose ici une approche séquentielle avec plusieurs niveaux de lecture de la puissance. Deux aspects principaux ont été envisagés : l'impact de la méthode de mesure et l'impact du contrôle de l'erreur de type 1. La combinaison de ces deux aspects donnent quatre niveaux de lectures pour la puissance. Cette approche permet d'investiguer les différentes étapes de l'évolution de la puissance au cours de l'analyse, permettant ainsi de préciser les étapes causant la plus grande perte.

3.2 Méthodes

Les approches de simulation sont fréquentes dans le domaine de la biologie haut-débit, particulièrement en transcriptomique [51, 52, 53, 40, 54, 55]. Les données de cette technologie sont souvent simulées à partir de distributions simples (gaussiennes le plus souvent), en se basant sur des données réelles pour fixer les paramètres de ces distributions. Les jeux de données simulés permettent alors d'investiguer des questions nécessitant un *gold standard*, c'est à dire une connaissance idéale du contenu d'un échantillon.

En spectrométrie de masse, les simulations doivent répondre à trois objectifs particuliers : la définition d'une référence (gold standard) ; le réalisme de la distribution des mesures ; la prise en compte des techniques de prétraitement.

Le premier et principal objectif est de fournir une référence, comme en transcriptomique. Les données haut-débit décrivent par nature des échantillons complexes, dont le contenu en protéines n'est pas précisément connu. Les travaux en protéomique sont donc classiquement menés en aveugle, sans qu'il existe une technologie validée fournissant une confirmation de ce contenu protéique. La simulation est pour cela basée sur des échantillons biologiques virtuels, dont le contenu est parfaitement connu, afin de pouvoir estimer les qualités et défauts des méthodes utilisées.

Ensuite, la simulation doit naturellement fournir des données réalistes. En spectrométrie de masse, ces données sont issues d'une chaîne de prétraitement (détaillée dans la section 2.2.2) plus complexe que celle de la transcriptomique, avec notamment deux aspects propres :

- la liste des pics d'un spectre n'est pas définie *a priori*, avec possible inclusion de faux pics et oubli de vrais pics,
- la mesure faite est une intensité lue dans un spectre, signal analogique échantillonné. Elle repose de fait sur de nombreuses hypothèses.

Une stratégie plus complexe de simulation que celle utilisée en transcriptomique doit donc être utilisée, prenant en compte le processus même d'acquisition du spectre de masse à partir des données, afin de permettre une étude des propriétés du signal induites par le prétraitement. La

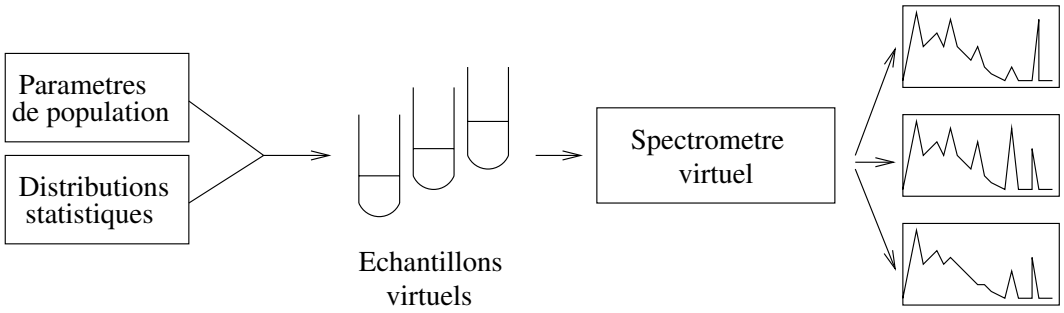

FIG. 3.1 – Schéma générale de la statégie de simulation : des hypothèses de simulation aux spectres de masse

prise en compte de ce processus doit en outre permettre de répondre à l'objectif de simulation de données réalistes.

Ces simulations rendent aussi possible la distinction de la responsabilité du plan d'expérience d'une part et de la responsabilité de la technologie expérimentale d'autre part. En revanche, une telle stratégie de simulation fait nécessairement un plus grand nombre d'hypothèses que la simulation directe d'intensités. Ces hypothèses sont souvent difficilement vérifiables avec les expériences actuellement disponibles. De plus, le choix des propriétés des biomarqueurs simulés est arbitraire et dépend vraisemblablement du contexte de l'étude (notamment de l'utilisation faite du biomarqueur).

Les résultats de simulation doivent donc être considérés de façon conjointe avec les hypothèses faites pour la simulation, ainsi que les objectifs de l'étude virtuelle (notamment en terme de type de biomarqueurs recherché, c'est à dire en terme d'hypothèse alternative). Les hypothèses de simulation devront être adaptées aux nouvelles données expérimentales à venir.

3.2.1 Stratégie de simulation

Toute expérience de spectrométrie de masse repose sur deux composants : un échantillon biologique et un spectromètre. Deux étapes sont donc nécessaires en pratique pour la simulation de spectres *in silico*. Il faut dans un premier temps générer des échantillons biologiques virtuels, comprenant la variabilité biologique souhaitée, avant de transformer ces échantillons virtuels en spectres virtuels, comprenant à la fois cette variabilité biologique et la variabilité instrumentale. Le schéma générale de la stratégie de simulation est présenté figure 3.1.

Génération d'échantillons virtuels Un échantillon biologique peut-être décrit comme une liste de peptides associés chacun à une concentration. Deux éléments doivent donc être précisés : une distribution permettant la génération d'étiquettes pour les peptides, en pratique leur masse, et un ensemble de distribution permettant de générer les concentrations.

Les paramètres de ces distributions ont été estimés à partir d'un jeu de spectres réels. Ces spectres sont issus d'une étude ayant pour thème la maladie de Hodgkin. L'objectif de cette étude était l'identification de facteurs pronostiques de la réponse au traitement dans le plasma d'individus atteints par la maladie. 24 individus rechuteurs et 24 individus non rechuteurs ont été inclus dans l'étude, avec 4 répétitions par individu. L'analyse de ces spectres permet d'obtenir des distributions d'intensité. Les distributions d'intensité ont été utilisées pour estimer les paramètres

des distributions de concentration. La variabilité de ces distributions inclue déjà une part de variabilité instrumentale, ce qui peut conduire *in fine* à des paramètres de variabilité surestimés pour les concentrations. Il n'existe toutefois pas d'étude évaluant la distribution quantitative des concentrations pour une large gamme de protéine.

Génération de spectres Un effort de recherche important a été fait par l'équipe de Morris et collègues, du M.D. Anderson Hospital, Houston, Texas, aux Etats-Unis, quant à la génération de spectres virtuels. En effet, cette équipe a introduit un simulateur de spectres de masse en 2005 [24]. Ce simulateur, développé pour comprendre les caractéristiques des données de spectrométrie de masse [56], permet de générer un spectre de masse à partir d'une liste de concentrations (en unité arbitraire, qu'on peut par exemple associée à des mmol.L^{-1}). Il décrit les étapes physiques de la spectrométrie de masse type *temps de vol*, plus spéciquement SELDI (mais en pratique, le processus reste globalement valable pour les technologies temps de vol dans leur ensemble, notamment MALDI). Il s'agit en pratique d'écrire les équations prédisant le temps de vol d'une particule ionisée, selon les lois de la dynamique et de l'éléctricité. Plus en détails, le spectromètre virtuel réalise les opérations suivantes :

- lecture d'une liste de peptides avec une étiquette de masse et un nombre de particules (assimilable à une concentration, en pratique),
- attribution d'une vitesse initiale et d'une direction à chaque particule, comme l'acquièrent les particules lors de l'impulsion laser dans un spectromètre réel,
- à partir de ces vitesses et positions initiales, calcul du temps de vol pour chaque particule,
- construction d'un histogramme du nombre de particule par intervalle de temps de vol,
- éventuellement, transformation de l'espace des temps de vol à l'espace des rapports masse sur charge m/z.

Au terme de cette expérience virtuelle, on obtient donc un spectre de masse avec des pics pour les différentes molécules simulées. Toutefois, les deux variabilités de mesure indiquées en chapitre 2.2.2, à savoir le bruit aléatoire et la ligne de base, ne sont pas générées lors de cette expérience virtuelle. Il est donc nécessaire de les inclure pour terminer l'expérience virtuelle et obtenir des spectres réalistes.

La procédure complète de simulation de spectres s'écrit donc :

- Génération de concentrations pour des protéines différentiellement exprimées (Differentially Expressed Proteins, DEP)
- Génération de concentrations pour des protéines non différentiellement exprimées (Non-DEP, NDEP)
- Spectrométrie virtuelle
- Ajout du bruit aléatoire
- Ajout de la ligne de base

Cette stratégie de simulation ne prend toutefois pas en compte le processus physique d'ionisation, comme remarqué par les auteurs du simulateur. Toute protéine introduite dans l'analyse aura donc un pic correspondant, ce qui ne correspond pas à la réalité. En effet, différentes capacités à être ionisés, la compétition entre peptides et la suppression d'un peptide par un autre limitent le nombre de peptides perçus sur le détecteur. En pratique, ces phénomènes font l'object de travaux sur la préparation des échantillons en amont de l'étape de spectrométrie, avec des études basées

sur un sous-protéome choisi.

3.2.2 Plan d'expérience de l'étude expérimentale *in silico*

L'objectif des simulations réalisées dans ce travail est d'investiguer l'effet de trois paramètres sur la puissance statistique :

- l'effet différentiel noté $\frac{\Delta}{\sigma}$
- la répartition entre groupes 0 et 1, notée p_{rep} pour la proportion d'individus du groupe 1
- la taille d'expérience notée n

L'unité fondamentale de l'étude expérimentale est l'*expérience*. Une expérience comprend un ensemble de spectres, avec un spectre par individu, permettant l'étude de m protéines détectées dans les spectres. Chaque expérience fournit donc une matrice $n \times m$ et permet de tester l'hypothèse nulle H_0 pour chacune des m protéines.

Chaque jeu de données doit permettre une estimation de la puissance dans les conditions qu'il définit. Une expérience ne donnant qu'une notion binaire (rejet ou non rejet de H_0) pour chaque protéine, il est nécessaire de répéter les expériences pour parvenir à une estimation de la puissance. La proportion des rejets de H_0 pour les protéines différentiellement exprimées (DEP) est donc évaluée sur un ensemble de 400 expériences. Ces 400 expériences constituent une *simulation*.

Trois niveaux d'effet différentiel ($\frac{\Delta}{\sigma} = \{0.3, 0.5, 0.75\}$), trois proportions d'individus du groupe 1 ($p_{rep} = \{0.15, 0.33, 0.5\}$) et trois tailles d'expérience ont été choisies ($n = \{100, 500, 1000\}$). En pratique, 9 simulations ont donc été réalisées, pour chaque couple de paramètres ($\frac{\Delta}{\sigma}; p_{rep}$), avec $n = 1000$ échantillons par expérience. L'effet du paramètre n a été investigué en tirant sans remise des sous-échantillons de taille $< n$ dans les échantillons de taille $n = 1000$.

Ces différentes situations expérimentales ont été choisies pour répondre à trois questions :

- quel est le gain en puissance possible lorsque la taille d'expérience n passe de 100 (comme beaucoup d'études courantes) à 1000 ? Une taille d'expérience $n = 1000$ est potentiellement envisageable, mais son coût doit être contrebalancé par une plus grande certitude d'identifier les biomarqueurs
- quel est l'impact d'une non équirépartition entre groupes, variable facile à ajuster ?
- quel différentiel d'expression peut espérer être vu dans une expérience à n et p_{rep} fixé ?

L'effet différentiel $\frac{\Delta}{\sigma}$ est choisi inférieur à 1, alors qu'il vaut 1 à 2 dans la majorité des simulations et études de puissance réalisées en transcriptomique. Le choix fait dans ce travail vise à focaliser l'étude sur des biomarqueurs présentant un faible différentiel d'expression. En pratique, les données collectées au sein de l'équipe pour une étude concernant la maladie de Hodgkin montre que l'albumine, marqueur pronostique bien connu de la maladie en question, présente un différentiel d'expression $\frac{\Delta}{\sigma} = 0.3$. Les nouveaux biomarqueurs présenteront certainement des différentiels d'expression du même ordre de grandeur. Il ne paraît donc pas raisonnable de choisir, comme dans la majorité des études, un effet différentiel plus grand que 1.

3.2.3 Stratégie d'analyse adoptée

Les données de spectrométrie utilisées pour l'analyse résultent de l'ensemble des étapes de prétraitement décrites dans la section 2.2.2. La lecture d'intensité a été réalisée par mesure de

l'aire sous le pic. Ce choix est discuté dans la section 3.3.1.

Approche univariée Dans le cadre de ce travail, des modèles univariés de la relation entre le statut de l'individu et l'intensité d'un pic ont été construits. Les données initiales sont un vecteur de statuts binaires $\boldsymbol{Y}$ des individus vis-à-vis de la pathologie d'intérêt et la matrice $\boldsymbol{X}$ d'intensités mesurées par pic et par individu, de dimension $n \times m$.

Un ensemble de modèles logistiques a été utilisé, permettant de modéliser la probabilité $p = P(Y_i = 1)$ selon l'équation (3.1) pour chaque variable j, où *logit* désigne la fonction logistique ($logit(p) = log(\frac{p}{1-p})$), X_j désigne la i-ème covariable et i désigne l'individu i.

$$logit(p_i) = \alpha + \beta_j X_{ij} + \epsilon_i \tag{3.1}$$

Ce modèle permet de tester l'implication des différentes covariables dans la prédiction du statut Y, en testant la nullité des coefficients de régression β_j. Plusieurs tests existent à cette fin, dont le test paramétrique de Wald. Dans ce test, l'hypothèse nulle H_0 correspond à la nullité du coefficient de régression. Le test s'écrit comme dans l'équation (3.2). L'hypothèse nulle est alors définie comme $H_0 : \beta_j = 0$, avec l'hypothèse alternative $H_1 : \beta_j \neq 0$. Sous l'hypothèse nulle, les valeurs du test suivent une distribution χ^2 à 1 degré de liberté.

$$\frac{\beta^2}{var(\beta)} \sim \chi^2(1) \tag{3.2}$$

On retient ici deux notions de puissance :

- la puissance individuelle $1 - \beta$, qu'on peut aisément étendre à la puissance moyenne $\frac{E(S)}{m_1}$ lorsque la puissance individuelle est identique pour tous les biomarqueurs potentiels et lorsque ces puissances individuelles sont indépendantes,
- la puissance relaxée $p(S > 0)$, qui fournit une estimation de la probabilité de rentabiliser l'étude menée par la découverte d'au moins un biomarqueur.

Ces deux notions de puissance présentent des degrés d'exigence différents, offrant deux perspectives complémentaires sur la capacité de découverte des études.

Le calcul de la puissance a été réalisé sur deux types de données :

- les données de concentrations, avant spectrométrie de masse, simulées et connues parfaitement, contenant seulement la variabilité biologique
- les données issues des spectres, après spectrométrie, comprenant à la fois la variabilité biologique et la variabilité instrumentale.

Ces deux niveaux de calcul permettent d'évaluer la perte de puissance liée à l'erreur de mesure et plus généralement au prétraitement. De plus, l'application d'une stratégie de contrôle du FDR ou la non application de cette stratégie permet de faire émerger, sur chaque niveau de calcul, la part de perte de puissance due à la situation de tests multiples. On peut ainsi suivre l'évolution de la puissance sur quatre étapes :

- données de concentration, pas de contrôle du FDR,
- données de concentration, contrôle du FDR,
- données de spectrométrie, pas de contrôle du FDR,
- données de spectrométrie, contrôle du FDR.

Approche multivariée L'approche statistique utilisée pour l'évaluation de la puissance statistique dans le cadre du travail soumis repose sur des modèles logistiques univariés. Cette approche ne permet pas d'appréhender les corrélations existant entre les variables, comme précisé dans la section 2.3.1. La prise en compte de ces corrélations requiert l'utilisation d'un modèle multivarié.

Dans le cadre de ce travail, le modèle Lasso logistique a donc été choisi et offre une approche multivariée de la question de la puissance développée ici. Le Lasso fournit en effet des coefficients *exactement* nuls pour les variables non incluses dans le modèle, permettant ainsi de réaliser une sélection de variables. En pratique, on construit un modèle de type $logit(p_i) = \alpha + \sum_{j=1}^{j=m} \beta_j X_{ij} + \epsilon_i$, avec une contrainte sur $\sum_{j=1}^{j=m} |\beta_j|$ (contrainte basée sur une norme L_1.), qui s'écrit sous la forme de la fraction des coefficients de la régression moindre carrés ordinaire, avec un paramètre $\lambda = \frac{\sum_{j=1}^{n} |\beta_j|}{\sum_{j=1}^{n} |\beta_j^0|} \in [0; 1]$, où les coefficients β_j^0 représentent les coefficients moindre carré usuels. Cette contrainte de pénalisation des coefficients de régression implique un contrôle du risque de première espèce. Ce contrôle est difficile à chiffrer en terme de risque de première espèce et rend difficilement interprétable une correction ultérieure des p-values pour effectivement contrôler le FDR. Le niveau effectif de contrôle du FDR a été investigué pour ces modèles Lasso. En outre, une étude du niveau de contrôle du FDR en fonction de la contrainte de pénalisation est proposée en section 3.3.

3.3 Résultats annexes

Dans cette section, des résultats complémentaires et annexes à ceux présentés dans l'article joint sont proposés. L'objectif est de fournir une meilleure compréhension des choix réalisés et des résultats obtenus. Trois points principaux sont présentés :

- le choix d'une méthode de lecture d'intensité,
- l'effet de la spectrométrie sur les intensités au travers des effets différentiels,
- l'approche multivariée.

3.3.1 Méthodes de lecture d'intensités

La question de la lecture de l'intensité associée à un pic est souvent considérée comme triviale et passée sous silence. Pourtant, plusieurs possibilités existent. On peut classiquement utiliser le maximum d'intensité atteint par un pic, ou une transformée de ce maximum, mais on peut aussi utiliser l'aire sous le pic. On oppose ainsi une approche à unidimensionnelle du pic, à une approche bidimensionnelle. La comparaison des approches *logmax* (logarithme du maximum d'un pic) et *AUC* (aire sous le pic) est proposée ici en terme de puissance. Le tableau 3.1 permet cette comparaison, avec la puissance individuelle lue par position en $\frac{m}{z}$ pour les deux stratégies de lecture envisagées. Une étude interne réalisée dans l'équipe, construite sur un modèle mixte avec utilisation de répétitions par individu, a permis d'estimer la variance résiduelle pour chacune des méthodes de mesure (C. Mercier, rapport technique interne). Cette variance résiduelle est plus faible pour la méthode de lecture AUC, confortant encore le choix de cette méthode.

$\frac{m}{z}$	3571	4857	6143	7429	8714	10000
AUC	0.16	0.17	0.18	0.17	0.19	0.20
logmax	0.14	0.12	0.13	0.12	0.14	0.13

TAB. 3.1 – Comparaison des puissances par position pour les deux méthodes de lecture d'intensité logmax et AUC. Stratégie d'analyse identique pour le reste du prétraitement. Simulation de paramètres $\frac{\Delta}{\sigma} = 0.3$, $p_{rep} = 0.5$, $n = 1000$.

3.3.2 Comparaison de l'effet différentiel pré- et post-spectrométrie

Afin de mieux comprendre les résultats de puissance donnés ici, une comparaison des effets différentiels avant (sur les données de concentration simulées) et après spectrométrie (sur les spectres générés) a été réalisée. Les résultats pour les trois niveaux d'effet différentiel simulés, pour les protéines différentiellement exprimées, sont présentés dans la figure 3.2, avec une taille d'expérience $n = 1000$ et une répartition égale entre les groupes. Les effets différentiels simulés sont bien conformes aux attentes, avec par exemple une moyenne égale à 0.3008 pour un effet différentiel fixé à 0.3. Les effets différentiels lus dans les spectres sont par contre plus petits. Un test de Wilcoxon pour la comparaison des tendances centrales des distributions de ces effets différentiels conclut à une différence significative de moyenne ($p < 2.2 \cdot 10^{-16}$). A titre de comparaison, les effets différentiels pré- et post-spectrométrie pour les protéines non différentiellement exprimées sont présentés en figure 3.3. La moyenne pré-spectrométrie des effets différentiels pour ces NDEP vaut -0.0004, contre -0.014 pour les DEP, correspondant dans les deux cas à des effets différentiels trop petits pour être intéressants (la différence entre moyennes est toutefois statistiquement significative par test de Wilcoxon). L'ensemble des étapes de la mesure et son traitement n'induit pas de biais : on ne crée pas d'effet différentiel important là où aucun effet différentiel n'est *a priori* simulé.

3.3.3 Approche multivariée

L'article présenté dans cette section ne fait pas apparaître le travail réalisé avec le Lasso, approche multivariée retenue pour cette étude de puissance. Les résultats concernant ce modèle sont donc présentés ici.

La puissance constatée avec un modèle de type Lasso est présentée dans le tableau 3.2. La comparaison avec les résultats obtenus pour les modèles univariés avec contrôle du FDR fait émerger plusieurs points :

- le Lasso présente une meilleure puissance pour beaucoup de cas difficiles ($p_{rep} = 0.15$ notamment, mais aussi avec $n = 500$ dans beaucoup de cas)
- en revanche, le Lasso ne dépasse pas 56% de puissance ($n = 1000$, $\frac{\Delta}{\sigma} = 0.30$, $p_{rep} = 0.15$), alors que l'approche univariée peut atteindre 78%, bien que dans un autre schéma expérimental ($n = 1000$, $\frac{\Delta}{\sigma} = 0.75$, $p_{rep} = 0.5$)

Toutefois, le Lasso ne permet pas un contrôle explicite de l'erreur de type 1 généralisé (type FDR). Une évaluation du FDR constaté dans ces expériences, lors de l'utilisation du Lasso, a donc été réalisée en parallèle.

Le niveau de FDR atteint pour les différentes situations présentées ci-dessus est détaillé dans le tableau 3.3. Une constante de pénalisation $\lambda = 0.5$ est utilisée par défaut dans ces modèles.

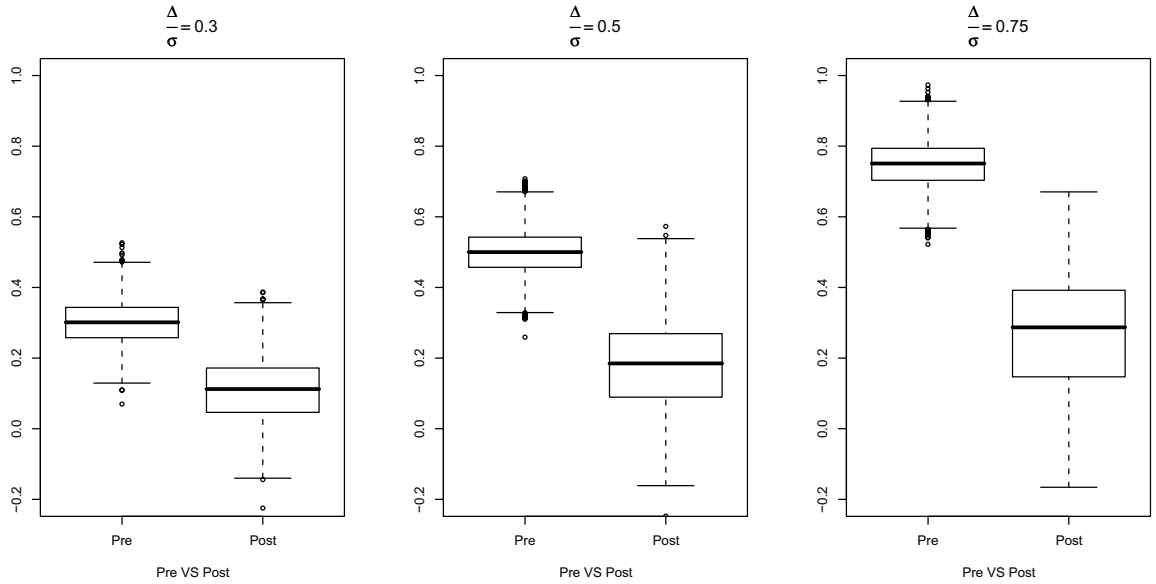

FIG. 3.2 – Comparaison des effets différentiels en pré- et post-spectrométrie pour les protéines différentiellement exprimées. Les effets différentiels en pré-spectrométrie (Pre) sont mesurés dans les concentrations générés, alors que les effets différentiels en post-spectrométrie (Post) sont mesurés dans les intensités lues dans les spectres. Dans chaque cas, les expériences de taille $n = 1000$ avec $p_{rep} = 0.5$ sont utilisées.

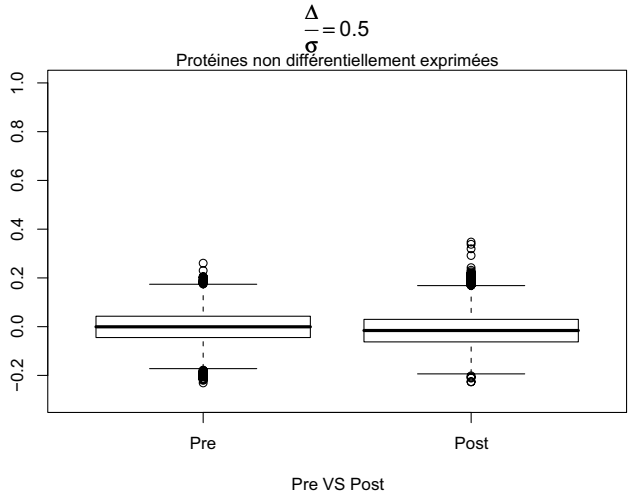

FIG. 3.3 – Comparaison des effets différentiels en pré- et post-spectrométrie pour les protéines non différentiellement exprimées. Expériences de taille $n = 1000$ avec $p_{rep} = 0.5$

TAB. 3.2 – Puissance évaluée sur les données issues des spectres - approche multivariée LASSO logistique.

Puissance	n	100			500			1000		
	$\frac{\Delta}{\sigma} \downarrow$ $p_{rep} \rightarrow$	0.15	0.33	0.50	0.15	0.33	0.50	0.15	0.33	0.50
	0.30	0.14	0.07	0.07	0.45	0.46	0.48	0.56	0.55	0.51
Individual	0.50	0.12	0.15	0.18	0.46	0.42	0.41	0.51	0.46	0.44
	0.75	0.29	0.28	0.24	0.45	0.41	0.39	0.47	0.42	0.41

FDR	n	100			500			1000		
	$\frac{\Delta}{\sigma} \downarrow$ $p_{rep} \rightarrow$	0.15	0.33	0.50	0.15	0.33	0.50	0.15	0.33	0.50
	0.30	0.80	0.89	0.89	0.71	0.70	0.67	0.69	0.68	0.77
Individual	0.50	0.88	0.83	0.80	0.75	0.73	0.73	0.75	0.75	0.73
	0.75	0.67	0.67	0.71	0.67	0.67	0.70	0.69	0.70	0.70

TAB. 3.3 – FDR empirique mesuré pour les différentes simulations réalisés, approche Lasso avec contrainte de pénalisation $\lambda = 0.5$ par défaut.

Ce FDR empirique dépasse clairement les 5% exigés pour l'approche univariée. Les résultats des deux approches ne sont donc pas comparables. Ce non contrôle du risque de première espèce suggère l'utilisation d'une contrainte de pénalisation plus importante, c'est à dire un λ plus petit, afin notamment de rendre comparables les résultats entre approche univariée et approche multivariée. Un choix plus adapté de la contrainte λ est possible ici, puisqu'un contrôle de la puissance et du FDR empiriques est possible. En pratique, l'ajustement de λ doit toutefois être fait en aveugle.

Une analyse de la puissance du modèle Lasso avec une contrainte de pénalisation variable a été réalisée dans un premier temps afin d'évaluer le comportement du Lasso en terme de FDR lorsque la contrainte de pénalisation évolue. Dans ces essais, on évalue la puissance et le FDR empiriques en fonction de la contrainte de pénalisation. On a donc une valeur de puissance et de FDR par valeur de λ. La figure 3.4 représente l'évolution de ces quantités pour différents paramètres expérimentaux.

Comme attendu, plus λ est petit (*i.e.* plus la pénalisation est importante) et plus le contrôle du FDR est efficace. Le lien entre pénalisation et FDR n'est toutefois pas linéaire. Selon la simulation, on a en effet des comportements non superposables, signifiant qu'il existe un lien complexe entre λ et FDR. Le choix de λ dépend notamment de l'effet différentiel, rendant difficile le choix d'une valeur initial pour la pénalisation. Cette évolution et ces différences conduisent à rechercher la plus grande valeur de λ telle que le FDR soit contrôlé à un niveau souhaité, pour un schéma expérimental donné. Sur les données simulées, cette recherche d'une contrainte compatible avec un FDR souhaité n'aboutit pas toujours. Il existe des simulations pour lesquelles il n'est pas possible de trouver une valeur de λ contrôlant le FDR à 5%, par exemple. La figure 3.5 illustre cette impossibilité. Elle présente la recherche d'un λ contrôlant le FDR sous 5%, par un algorithme de descente de gradient jusqu'à ce niveau de FDR. On constate un FDR plancher pour chaque effet différentiel, ne permettant pas d'atteindre la valeur FDR=5%.

3.4 Discussion

Les résultats de l'étude de puissance réalisée ici font clairement ressortir le besoin de réaliser des études comprenant plus d'individus, lorsque la spectrométrie de masse est utilisée pour la détection de nouveaux biomarqueurs. La majorité des études menées jusqu'à maintenant ne permet pas la détection de biomarqueurs avec une probabilité suffisante pour être rentable.

Deux niveaux de perte de puissance sont mis en évidence : une perte due à la variabilité instrumentale, et une perte due au contrôle du risque de première espèce en situation de tests

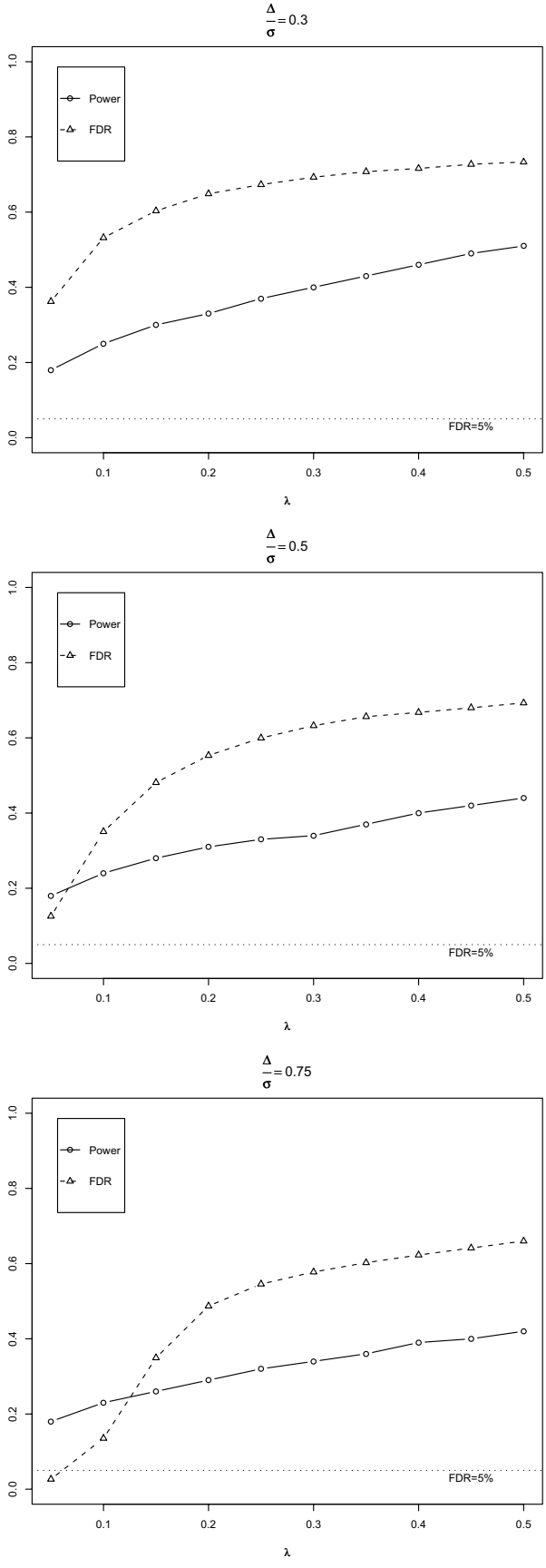

FIG. 3.4 – Evolution du FDR empirique et de la puissance en fonction de la contrainte de pénalisation λ

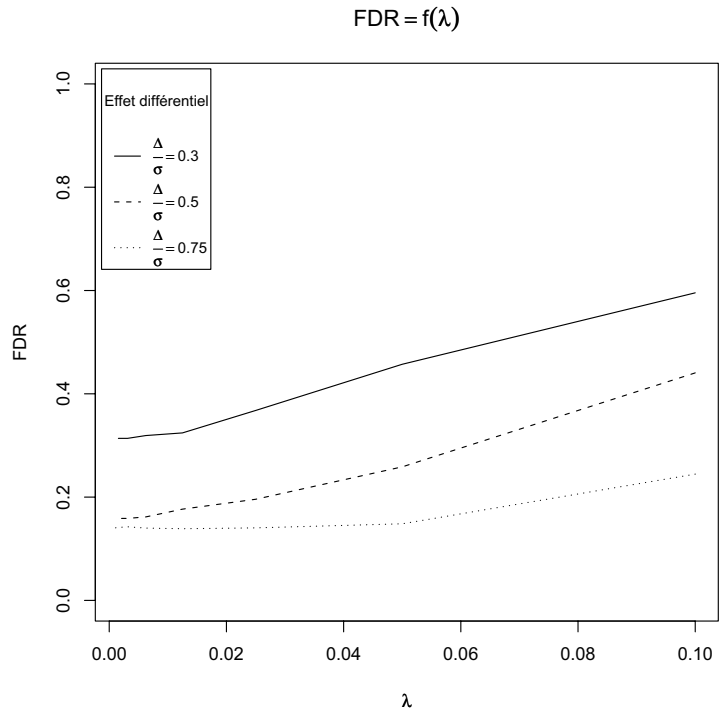

FIG. 3.5 – Evolution du FDR pour les petites valeurs de λ. Recherche de λ permettant de contrôler le FDR à 5%.

multiples. Une synergie est mise en évidence entre ces deux pertes, encourageant à la réduction de la variabilité instrumentale pour améliorer la puissance des études. L'impact de cette variabilité instrumentale est bien mis en évidence par la comparaison des effets différentiels attendus et mesurés.

La variabilité instrumentale considérée ici ne peut distinguer la variabilité directement due à l'instrument de mesure de la variabilité induite par les algorithmes de prétraitement. Des améliorations sont ainsi à rechercher tant sur le plan expérimental, avec une amélioration des méthodes même de mesure, que sur le plan analytique, avec une amélioration des algorithmes de prétraitement et des procédures statistiques utilisées. Le choix d'une lecture d'intensité basée sur l'aire sous le pic se montre ainsi plus pertinent. L'utilisation d'un modèle multivarié sophistiqué Lasso, en revanche, ne donne pas lieu à de meilleurs résultats que l'approche basique univariée. La difficulté à contrôler le risque de première espèce pour un modèle Lasso encourage toutefois à une étude plus poussée du lien existant entre constante de pénalisation λ et FDR, ou toute autre mesure du risque de première espèce.

Afin de décortiquer plus avant les déterminants de la perte de puissance, La piste de l'amélioration des méthodes de prétraitement conduit naturellement à se poser la question de leur effet sur la détectabilité des peptides différentielles. Des études comparatives ont été proposées [57, 58], faisant à nouveau état du manque crucial d'un jeu de données de référence. Les différences mises en évidence en terme d'efficacité dans la détection des pics posent des questions quant à l'impact d'une détection imparfaite sur les résultats, notamment de puissance. Cette question est investiguée dans le chapitre suivant.

Statistical power in mass-spectrometry proteomic studies

Thomas Jouve[1,2,3]*, Delphine Maucort-Boulch[1,2,3], Patrick Ducoroy[4], Pascal Roy[1,2,3]

Abstract

Background Time-Of-Flight mass-spectrometry (MS) (MALDI or SELDI) is a promising tool for the identification of new biomarkers, or differentially expressed proteins, in various clinical situations. The ability to recognize a biomarker as such, or statistical power, can not be accessed through classical experiments due to the lack of true protein content information. **Methods** We provide a simulation study to investigate the statistical power of MS experiments, under FDR control. Virtual mass spectra are created from virtual individuals belonging to one of two groups. We examine the effect of i) the sample size n, ii) the group repartition p_{rep}, iii) the differential expression level, three critical parameters in any clinical setting. The power study is led before and after inclusion of instrumental variability. Also, we evaluate an alternative power measure that can prove useful in MS experiments, called relaxed power. **Results:** We show that small sample sizes of about $n = 100$ spectra offer a power equivalent to the power seen in group-label permuted data, about 5-10%. Increasing n, p_{rep} or the differential effect allows to reach a much better power. The power loss incurred through FDR control is high when instrumental variability is high (as in MS data), while this power loss is negligible on data free of instrumental variability. **Conclusion** The high instrumental variability encountered in MS, together with FDR control, builds a detrimental synergy leading to a low statistical power for usual MS studies sample sizes. This detrimental synergy is a proper issue in MS studies and should be compensated for by increasing sample sizes in MS-based biomarker discovery experiments, but also by lowering MS instrumental variability.

Background

Mass-spectrometry (MS), as a member of the *high-throughput* methods, allows to explore a wide feature space, with simultaneous measurements for hundreds of proteins. This technology therefore allows to screen for new biomarkers of human diseases by analyzing easily sampled biological fluids (e.g. blood). Numerous studies with a strong emphasis on oncology [1] were published, aiming at detecting new cancer biomarkers. Some potential biomarkers were put into light, though still requiring disease-specificity assessment [2]. While these studies show great promises, no assessment of the very biomarker detection ability was performed.

The biomarker detection ability translates into the concept of statistical power. Sample size calculations represent the most common perspective on this statistical quantity. Some power studies were already performed in the field of transcriptomics [3, 4], reminding us of the need for improved experimental designs. The added value of new transcriptomic biomarkers started to be investigated as well [5].

*corresponding author, thomas.jouve@chu-lyon.fr
[1]Hospices Civils de Lyon, Service de Biostatistique, Lyon, France
[2]Université Claude Bernard Lyon 1, Université de Lyon, Villeurbanne, France
[3]Laboratoire Biostatistique Santé, UMR CNRS 5558, Pierre-Bénite, France
[4]Clinical and Innovation Proteomic Platform, CHU Dijon, France

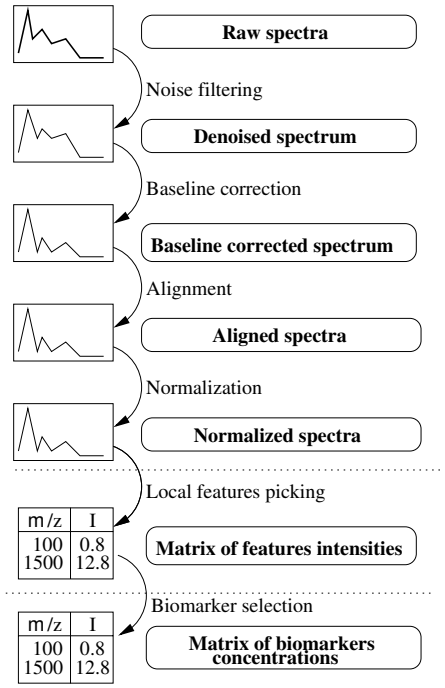

Figure 1: Preprocessing of spectra: from raw data to usable intensity matrix

In proteomics, however, especially in mass-spectrometry studies, power considerations are still poorly investigated, with one recent exception [6] providing a sample size calculation method based on theoretical considerations.

Statistical power is tightly linked to the concept of variability. In a simple situation with inference for one single biomarker, the experimental scheme of MS data integrates different levels of variability. Biological variability exists and define biomarkers with molecules showing different expression patterns between groups of individuals, but this variability also hides biomarkers at the time of analysis. Indeed, expression levels within groups of individuals also account for biological variability. Instrumental variability adds to this first component of variability and hides further true biomarkers. This instrumental variability also depends on the pretreatment process, presented in figure 1, which should be optimized. Variability in mass spectrometry studies already gave rise to reproducibility studies[7, 8], showing relatively low instrumental variability, i.e. good reproducibility. However, this aspect offers no guarantee for statistical power.

Our ability to detect biomarkers fundamentally depends on these two components of variability. The large feature space encountered in proteomics also adds to the complexity of biomarkers detection. Indeed, it requires the use of statistical strategies adapted to the multiple testing setting.

Transcriptomic studies offered some insights on statistical power when considering biological variability and multiple testing issues. Results from these studies can not be applied to proteomics, since they do not take instrumental variability into account. A proper power study in proteomics is therefore required. In the absence of a gold standard providing true samples contents, the first step of this power study must rely on simulations. These simulations indeed encompass biological and instrumental variability, as well as multiple testing issues, the three defining determinants of statistical power in proteomics.

In this article, we assess the statistical power of MALDI-TOF or SELDI-TOF MS studies for the identification of new biomarkers, when controlling the number of false discoveries at a given level. To address this problem, we use a simulation strategy to generate spectra mimicking real MS experiments,

as described in a first part of the methods. Using a single arbitrary preprocessing scheme to handle these simulated spectra, we investigate the effect of different parameters on statistical power at different analysis levels, under FDR control, as described in a second part of the methods. Our results are summarized in a third section, followed by reading guidelines and commentaries.

Methods

The methods used in this article take into account the different issues of MS data analysis: biological and instrumental variabilities, described in the *Simulations* part, as well as multiple-testing specificities, described in the *Analysis* part.

Simulations

To reflect the two different sources of variability, two simulation steps are defined: generation of (virtual) samples and subsequent transformation into (virtual) mass spectra.

A *sample*, from a *subject*, is a labeled list of m concentrations for different proteins (no repetition was used in this study, so each subject corresponds to one sample). Each subject belongs either to the risk group 1 or to the reference group 0. For simplification purposes, the term protein will refer to proteins themselves or protein fragments (peptides). Labels are anticipated mass-to-charge ratio $\frac{\mathrm{m}}{z}$. A group of n samples builds an *experiment*. Experiments contain $n \times m$ concentration measures. An experiment is the data usually available for a biomarker identification study. In order to measure power, collections of experiments are built. A collection of experiments is referred to as a *simulation*.

Table 1 summarizes the different parameters that must be specified in order to fully describe the virtual biological samples, i.e. parameters for biological variability. They were chosen through inference on a set of 192 real spectra obtained from a Bruker TOF-SIM MALDI spectrometer. These spectra come from a prognostic study for Hodgkin's disease, led by the CLinical and Innovation Proteomic Platform (Clipp). Spectra were acquired from plasma of 32 different individuals, each with 6 replicates, in the mass window extending from 1 to 10 kDa.

Using these parameters, samples description requires the definition of protein concentrations. A strategy to generate these concentrations is therefore required, summarized in figure 2. We call *fundamental parameters* the parameters describing the means and standard deviations of all proteins in the whole population. Using Gaussian distributions, 6 parameters are required. Parameters μ_M and μ_S represent the mean parameters for the mean and standard deviation of the protein concentrations, respectively. Similarly, σ_M and σ_S represent the standard deviation parameters for the mean and standard deviation of the protein concentrations, respectively. Finally, p_{rep} represents the repartition between group 0 and group 1 subjects, and Δ_j sets the difference between mean concentrations of protein j for group 0 and group 1 subjects. For *Non-Differentially Expressed Proteins* (NDEP), we have $\Delta_j = 0$, while $\Delta_j \neq 0$ for *Differentially Expressed Proteins* (DEP). There are m_0 NDEP and m_1 DEP, with $m_0 + m_1 = m$. We used $m_0 = 60$, $m_1 = 8$ and the same $\frac{\Delta_j}{S_j}$ value for all given simulation. The 6 previously defined parameters allow to define protein j concentration mean M_j and intra-group standard deviation S_j parameters for each of the m proteins, as shown in equation 0.1, where N denotes the Gaussian distribution. These equations define *protein concentration distributions*, at the population level.

$$\begin{cases} M_j \sim (1 - p_{rep}) \cdot N(\mu_M, \sigma_M) + p_{rep} \cdot N(\mu_M + \Delta_j, \sigma_M) \\ S_j \sim N(\mu_S, \sigma_S) \end{cases} \tag{0.1}$$

Using these distributions, concentrations c_{ij} for all samples i and protein j within an experiment were generated, thereby defining the whole $n \times m$ matrix describing the experiment. Each set of concentrations

	Distribution	Parameter	Model	Spectral component
Biological variability				
		m/z	Arbitrary defined	
DEP	Gaussian	Mean	$N(\mu_M, \sigma_M) + \delta_i \cdot \Delta$	$\phi(t)$
		Standard deviation	$N(\mu_S, \sigma_S)$	
		$log(\mathsf{m}/z)$	$N(\mu_{\mathsf{m}/z}, \sigma_{\mathsf{m}/z})$	
NDEP	Gaussian	Mean	$N(\mu_M, \sigma_M)$	$\phi(t)$
		Standard deviation	$N(\mu_S, \sigma_S)$	
Instrumental variability				
Baseline	Gamma	Shape	$N(\mu_{shape}, \sigma_{shape})$	$b(t)$
		Scale	$N(\mu_{scale}, \sigma_{scale})$	
		Maximum	$N(\mu_{max}, \sigma_{max})$	
		Maximal position	$N(\mu_{argmax}, \sigma_{argmax})$	
Random noise	Gaussian	Standard deviation	$N(0, \sigma_{rn})$	$r(t)$
Total concentration	Log-normal	Standard deviation	$logN(0, \theta)$	a

Table 1: Elements of a virtual experiment: models for biological variability and instrumental variability. DEP=Differentially Expressed Proteins, NDEP=Non-DEP, δ_i describes the group of subject i. Note that the link between proteins and their spectral component ϕ is made by the virtual mass-spectrometer.

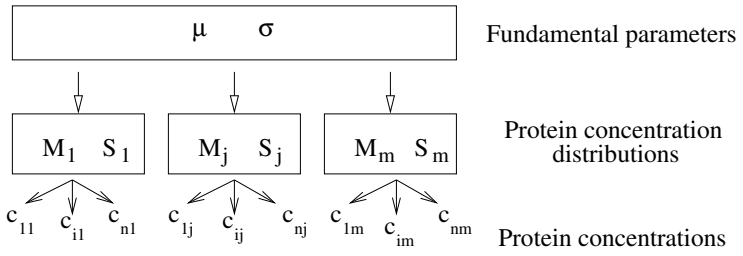

Figure 2: Overview of the process used to generate concentration values for each protein, for each sample. μ and σ are chosen parameters, M_j and S_j are random variables extracted from the distributions based on the fundamental parameters, defining protein concentration distributions, at the population level, c_{ij} values are random values drawn from the protein concentration distributions, defining concentrations within each sample.

$c_{i\cdot}$ defines a sample. The gaussian variability structure of these concentrations might be simple but clustering of spectra in previous experiments showed good performances for simple algorithms (e.g. Linear Discriminant Analysis) [9], making it a reasonable choice.

Mass labels for DEP are arbitrary defined once and for all simulations within the [1;10]kDa window, spanning the whole interval. Mass labels for NDEP come from a Gaussian distribution of the mass logarithm, truncated in the same mass window as DEP, and are defined once for each experiment, using two fundamental parameters, as in equation (0.2).

$$log(\frac{\mathsf{m}}{z}) \sim N(\mu_{\mathsf{m}/z}, \sigma_{\mathsf{m}/z}) \tag{0.2}$$

A spectrum is derived from a sample using a virtual mass spectrometer described by Morris et al. [10], implemented in the R software [11]. Given a list of protein concentrations with associated masses, it outputs a spectrum $\phi(t)$ free of the two classical MS instrument-noises. The virtual mass-spectrometer makes the link between a simulated sample and $\phi(t)$. Chemical noise $b(t)$ and random noise $r(t)$ are added to this spectrum to obtain a realistic spectrum. In mathematical terms, we consider the model of equation (0.3) to describe a spectrum.

$$I(t) = a \cdot (\phi(t) + b(t) + r(t)) \tag{0.3}$$

In this equation, t is the *time-of-flight*, $\phi(t)$ represents the actual signal of interest, generated by the virtual mass-spectrometer by using the previously defined samples, $b(t)$ and $r(t)$ correspond respectively to the baseline and a random noise. The sum of these three components, with a multiplicative coefficient a accounting for the variability in sample deposit and ionization, builds the total signal of interest $I(t)$, as the output of a real MS assay. Baseline $b(t)$ is generated as a scaled gamma distribution. Its parameters are drawn from Gaussian distributions with parameters μ_{shape}, σ_{shape}, μ_{scale}, σ_{scale}. Parameters μ_{max}, σ_{max}, μ_{argmax}, σ_{argmax} are used to adapt the gamma density to the position of a classical baseline, both on the m/z axis ($argmax$ parameters) and on the intensity axis (max parameters). Baseline definition is performed by sampling from distributions using these parameters, once for each spectrum. Similarly, coefficient a is generated through a Gaussian distribution with standard deviation θ, once for each spectrum. Noise $r(t)$ is generated through a Gaussian distribution as well, with standard deviation parameter σ_{rn}.

Simulation parameters linked to instrumental variability were inferred from a set of real spectra, in the same way as biological variability parameters described above, from a Bruker TOF-SIM MALDI spectrometer. Both are summarized in table 1. Noise intensity was compared to the mean signal intensity

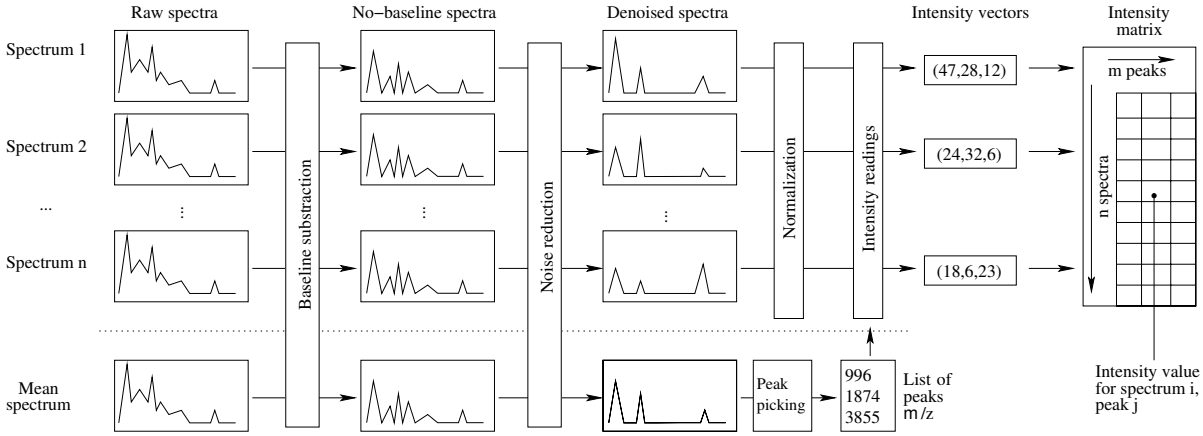

Figure 3: Denoising strategy for virtual experiments: from spectra to matrix of peak intensities

to calibrate $r(t)$. Baseline properties (maximum intensity and position on the m/z axis) were inferred in the same way to calibrate $b(t)$. Physical parameters (e.g. drift tube length, voltages, and ion focus delay, not described here but required for the virtual MS) were set according to the properties of this instrument as well.

Nine different simulations of 400'000 spectra (400 experiments with 1000 spectra each) were performed. Two parameters define a simulation: p_{rep} (describing the proportion of group 1 samples) and $\frac{\Delta_j}{S_j}$ (describing the systematic differential effect for DEP, later referred to as $\frac{\Delta}{\sigma}$ for simplification purposes). Sub-experiments of sizes $n = 100$ and $n = 500$ were drawn from experiments of size $n = 1000$, finally allowing to study the effect of n, p_{rep}, and $\frac{\Delta}{\sigma}$. These simulations were performed with the R software on a AMD Athlon X2 5000+ computer with 2GB RAM. The 9 simulations were generated in a mean time of 20 hours each, occupying an average space of 30 GB each.

Analysis

Preprocessing

The overall scheme of the preprocessing strategy used here is described in figure 3. It shows how all steps described below integrate to build a preprocessing strategy.

Smoothing of successive local minima was used for baseline substraction, as a fast and efficient approach to the problem of modeling the baseline and substract it. The PROcess[12] package implementation was used. An *undecimated wavelet transform* (UDWT) was used to perform random noise $r(t)$ reduction in the simulated spectra [13], using the *Rice Wavelet Toolbox* (rwt) R-package. UDWT only requires one thresholding parameter, specifying which scales of the signal should be filtered out. This thresholding parameter was chosen as 3.6 times the mean absolute deviation (MAD) of the signal. Hard thresholding was used, i.e. zeroing all coefficients smaller than the thresholding parameter.

Peak localization was performed on the preprocessed mean spectrum, as recommended by Morris et al.[10]. The mean spectrum was computed as the mean of all raw spectra and then preprocessed. Noise filtering on the mean spectrum was performed using a high filtration threshold. Since the focus of this study was not set on preprocessing, all local maxima found in the pretreated mean spectrum were initially considered as peaks. This approach makes the least hypotheses on peaks shape. A lot of these initially identified peaks had very low intensities, at the level of random noise intensity fluctuations. The list of detected peaks was therefore filtered, using a signal-to-noise ratio threshold on intensity (noise was

defined as the part of the signal filtered out by the rwt algorithm). The position of the local maxima and of their neighboring minima on both sides were kept, allowing to define the width of a peak.

A tolerance on m/z positions was introduced to match found peaks and simulated peaks since the exact position of a peak in low-resolution spectra (like MALDI-TOF spectra) is not precisely defined. A $\pm 0.3\% \cdot \mathsf{m/z}$ tolerance was used, defining a window into which a found peak and a simulated peak were considered one single entity, corresponding to a single protein. A count of found peaks was kept in all experiments, allowing to derive an average measure of the Peak Detection Ability (PDA).

To normalize peak intensities, all intensities in a spectrum were divided by the *Total Ion Current* (TIC), estimated by the area under the spectrum, as is commonly used in SELDI / MALDI pretreatment. Furthermore, peak intensities were estimated through the *Area Under the Peak* (AUP), following the chosen normalization strategy.

In order to give an insight on the choice of variability parameters of the simulations, a comparison of the standard deviation within a spectrum with the maximum intensity of this spectrum was performed both for the real spectra (used previously for calibration) and the simulated spectra. This stdev/max ratio allows to compare variability between different spectra with arbitrary units. Similarly, a variability comparison for extracted intensities was performed, through the coefficient of variation (CV) defined by the standard deviation of a peak intensities divided by the mean intensity for this peak. The same pre-processing strategy was applied to the real spectra and the simulated spectra and the CV were compared.

Statistical analysis

The logistic regression model is well-suited for diagnostic or prognostic studies, giving a probability e.g. of disease for diagnostic studies, given the value of some covariate(s). Univariate logistic models were used in this study to predict the probability of belonging to group 1 given the intensity measure for a protein. The Wald test on the regression coefficient of the model was used to test for differential expression between groups. The control of the number of false positive conclusions was performed using the q-value [14] (controlling the positive False Discovery Rate). The q-value is the multiple-testing equivalent of the classical p-value, expressing the collective type 1 error as the highest q-value associated to a protein for which H_0 is rejected.

Statistical power, tightly linked to the type 2 error, is the other issue of the statistical analysis. Calling R the event *H0 rejection*, power can be written as $pr(R|DEP)$. Power estimation in the setting of multiple testing needs to be carefully considered since this setting does not offer a trivial definition of power. Two measures of power were used and are described thereafter: the individual power and the relaxed power. Values for both measures are presented in the results.

The classical definition, here referred to as *individual power*, can be written $1 - \beta_1$, where β_1 is the individual type 2 error (with respect to a protein, or potential biomarker). It still holds in this context and has a natural interpretation when the differential effects are the same for all DEP. Indeed, it is easy to show that $(1 - \beta_1) = \frac{E(S)}{m_1}$, where S is the number of DEP for which H_0 is rejected. The ratio $\frac{E(S)}{m_1}$ is referred to as average power and has the same value as the usual individual power in this precise situation. In more concrete terms, this average power evaluates the average number of true discoveries among all DEP. It quantifies the ability offered by the study to indeed detect a proportion of the biomarkers, as expected from a power definition.

Nevertheless, new questions arise with high-throughput studies. It can be useful to estimate the probability of detecting all biomarkers. Lee et al. [3] proposed to use the type 2 error β_F, defined as $(1 - \beta_F) = (1 - \beta_1)^{m_1}$ (under independence hypotheses), for this probability estimation. This can be a too strong constraint for study calibration. Depending on the study, it can be interesting to find at least some of the biomarkers. This can be restated as $pr(S > k)$ where we require more than k true discovery.

Power	n	100			500			1000		
	$\frac{\Delta}{\sigma} \downarrow$ $p_{rep} \rightarrow$	0.15	0.33	0.50	0.15	0.33	0.50	0.15	0.33	0.50
Individual	0.30	0.00	0.01	0.02	0.24	0.59	0.68	0.70	0.95	0.98
	0.50	0.03	0.14	0.19	0.88	0.99	1.00	0.99	1.00	1.00
	0.75	0.24	0.68	0.77	1.00	1.00	1.00	1.00	1.00	1.00
Relaxed	0.30	0.03	0.10	0.12	0.76	0.98	0.99	1.00	1.00	1.00
	0.50	0.19	0.54	0.61	1.00	1.00	1.00	1.00	1.00	1.00
	0.75	0.71	0.98	1.00	1.00	1.00	1.00	1.00	1.00	1.00

Table 2: Power evaluated on sample data, FDR controlled at 5%, n is the total number of subjects, $\frac{\Delta}{\sigma}$ is the differential effect; p_{rep} is the proportion of cases.

We call *Relaxed Power* (RP) the special case when $k = 0$, i.e. $RP = p(S > 0)$. Its algebraical writing is $(1 - \beta_{RP}) = 1 - \beta_1^{m_1}$. It is the probability of at least one biomarker discovery (provided there is at least one to detect).

The number of H0 rejection (with a q-value threshold) for each covariate was calculated among all experiments, within a simulation. This allowed to easily derive statistical power. A count of H0 rejection without type 1 error control strategy was also kept in order to estimate power before any multiple testing error control strategy. It was estimated by using a 5% individual type 1 error, defining a *naive approach* (its lack of consideration of the number of the number of tests makes it virtual). The experimental design developed here allows to study the impact of n, p_{rep} and $\frac{\Delta}{\sigma}$ on these different power levels. The precision of power values is provided as the 95% confidence interval for a binomial distributions $B(400, P)$, where P is the empirical power over the 400 experiments of each simulation.

Permutations of group labels on spectra data were also performed in order to create a generalized H_0 data, in which no protein is truly differentially expressed. The apparent power on this permuted data gives an insight on the level of power that can be expected by chance. Comparisons of power estimated on true and permuted data help in finding truly informative simulations.

Results

Table 2 shows individual power $1 - \beta_1$ and relaxed power $pr(S > 0)$ when using sample data (as opposed to spectral data), that is to say concentration values. This data contains biological variability only, ignoring instrumental variability. Good individual power values ($\geq$ 80%) can be reached when the sample size is $n = 1000$. When dealing with fewer samples, power drops largely. Group repartition does also substantially affects power: there is a power decrease with decreasing p_{rep} from 0.5 to 0.15, with 10 to 50% losses. Finally, the differential effect size impacts power severely. This is best seen for $n = 500$, $p_{rep} = 0.15$, with a power drop of 76% from $\frac{\Delta}{\sigma} = 0.75$ to $\frac{\Delta}{\sigma} = 0.3$. Altogether, it should be noticed that the amplitudes of power variation for one given parameter are different for the individual values of another parameter.

On this sample data, we see that relaxed power reaches 100% in many settings. The same pattern of changes as in the individual power case is seen: sample size decreases power most severely, but group repartition also has an impact. The differential effect has an impact about the same amplitude as the sample size. However, relaxed power is better than individual power by definition. Experiments with a

total sample size of $n = 500$, a group repartition $p_{rep} = 0.5$ and a differential effect $\frac{\Delta}{\sigma} = 0.3$ will almost surely lead to a biomarker discovery, while individual power only showed a deceptive 68%.

Power values without type 1 error control strategy, in the naive approach, was also evaluated (results not shown) on sample data. The general pattern for power loss is found here as well: power increases with sample size n, group repartition p_{rep} and differential effect $\frac{\Delta}{\sigma}$. This naive individual power reaches 14% for the most difficult setting ($n = 100$, $p_{rep} = 0.15$ and $\frac{\Delta}{\sigma} = 0.3$). Using $n = 500$ samples instead allows to attain 68% individual power, while the use of a balanced design gives a power of 29%. Here again, an increase in n has a larger effect on power than an increase of p_{rep}, but the increase of power with p_{rep} still allows to double power results. Increasing the differential effect to $\frac{\Delta}{\sigma} = 0.5$, power achieves 41% with $n = 100$ and $p_{rep} = 0.15$, increasing to 67% by using $p_{rep} = 0.5$. All other cases accomplish power $\geq 80\%$.

The stdev/max ratio, comparing variability for raw intensities, ranged from 0.029 to 0.191 for real spectra, with a median at 0.062, while it ranged from 0.083 to 0.152, with a median at 0.099, for simulated spectra. The standard deviation for real spectra, to the scale of the maximum intensity on these spectra, is therefore slightly smaller than for simulated spectra. When it comes to peak intensity values, however, the CV ranges from 1.968 to 7.682, with a median at 5.918 for real spectra, while it ranges from 0.360 to 0.424, with a median at 0.390 for simulated spectra. Summarizing, the variability for full spectra appears somewhat higher in simulated spectra for raw intensities, but detected peaks in simulated spectra have a smaller variability than real spectra. The Peak Detection Ability (PDA) was also evaluated to investigate our preprocessing strategy (results not shown). The PDA reaches a level of 97% when using 100 samples. When using 10 times more samples, the PDA increases to 99%. Comparing this PDA to the total number of simulated proteins (about 70), this means that peaks are most often all detected. Varying the parameter p_{rep} does not impact PDA (PDA=98% for $n = 500$ and $p_{rep} = 0.15$, PDA=98% as well for $n = 500$ and $p_{rep} = 0.5$). This confirms the mean spectrum properties described in [10].

Table 3 displays the evolution of power for spectral data (i.e. our simulated MS data) with sample size n, repartition between groups p_{rep} and differential effect $\frac{\Delta}{\sigma}$. This data contains both biological and instrumental variability. An important power loss is incurred when going through MS and its statistical analysis. Where a 100% individual power was found for sample data, an individual power of 78% at best can be reached when dealing with spectral data, for 1000 samples, an equal repartition between group and a differential effect $\frac{\Delta}{\sigma} = 0.75$. The general pattern is the same as in table 2. Individual power declines with decreasing n, p_{rep} and $\frac{\Delta}{\sigma}$, reaching values as low as 5% for the worst case scenario ($\frac{\Delta}{\sigma} = 0.3$, $p_{rep} = 0.15$, $n = 100$). Using group-label permutations on this data, the apparent individual power reaches 5-10% in the same setting, showing that this level of power corresponds essentially to a non-informative simulation.

Focusing now on relaxed power, we see that this definition of power does not provide better results for $n = 100$, while it is largely better for $n = 1000$, allowing to double power results. However, this shows that the return on investment for small experiments ($n = 100$) is very low (less than 20% of chances to detect at least one biomarker).

Figure 4 compares the approach controlling the FDR with the *naive approach*, putting into perspective results from table 2 and 3 (here presented as *FDR control power*). The FDR control leads to a loss of power, when compared to the naive approach. This loss is very different depending on the differential effect. Of particular importance is the data-dependent power loss magnitude. When dealing with sample data, controlling for the FDR only slightly lowers power results. The introduction of instrumental variability leads to a situation where the FDR correction severely degrades power. For a small differential effect ($\frac{\Delta}{\sigma} = 0.3$), the FDR power loss is almost negligible on concentration data, while power is divided by three (from 50% to 18%) on spectral data.

Power	n	100			500			1000		
	$\frac{\Delta}{\sigma} \downarrow$ $p_{rep} \rightarrow$	0.15	0.33	0.50	0.15	0.33	0.50	0.15	0.33	0.50
Individual	0.30	0.05	0.08	0.11	0.09	0.12	0.11	0.12	0.20	0.18
	0.50	0.07	0.07	0.10	0.20	0.29	0.36	0.31	0.47	0.54
	0.75	0.09	0.11	0.10	0.41	0.59	0.64	0.63	0.72	0.78
Relaxed	0.30	0.06	0.09	0.16	0.14	0.23	0.28	0.26	0.47	0.46
	0.50	0.09	0.10	0.14	0.39	0.64	0.74	0.73	0.93	0.96
	0.75	0.14	0.18	0.21	0.84	0.90	0.93	0.90	0.90	1.00

Table 3: Power evaluated on spectral data, FDR controlled at 5%, n is the total number of subjects $\frac{\Delta}{\sigma}$ is the differential effect; p_{rep} is the proportion of cases.

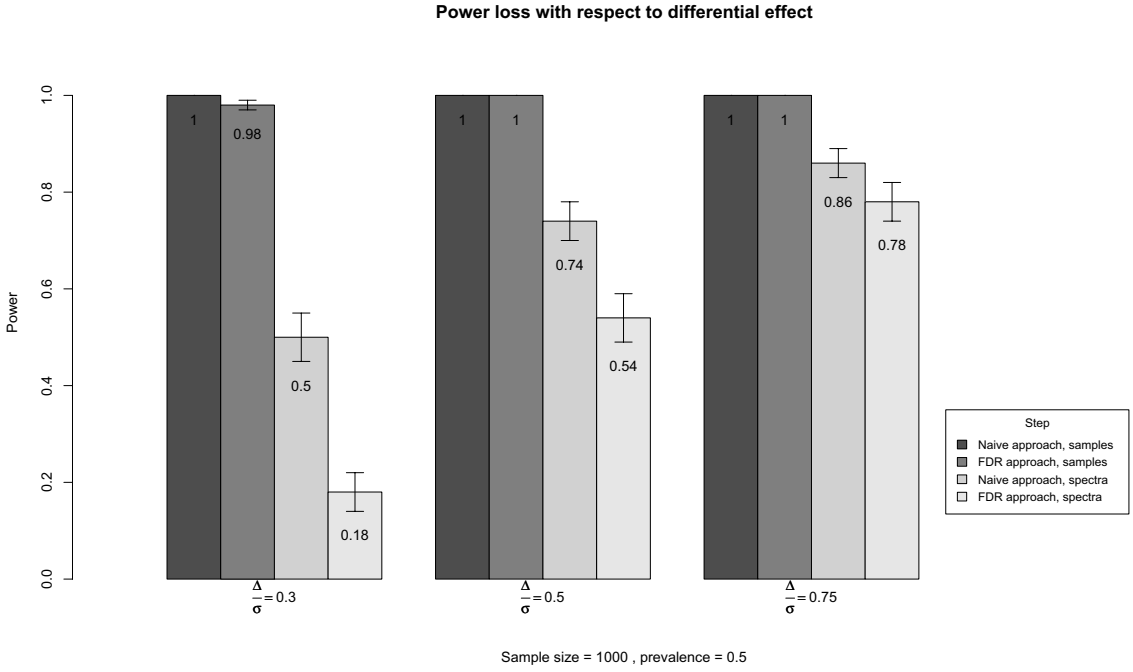

Figure 4: Power comparison for different analytical steps. "Naive" refers to the lack of consideration of any multiple testing type 1 error control (using a classical 5% cutoff), samples refers to the use of sample data (perfect knowledge of concentrations) and spectra refers to the use of spectral data. 95% confidence interval are shown.

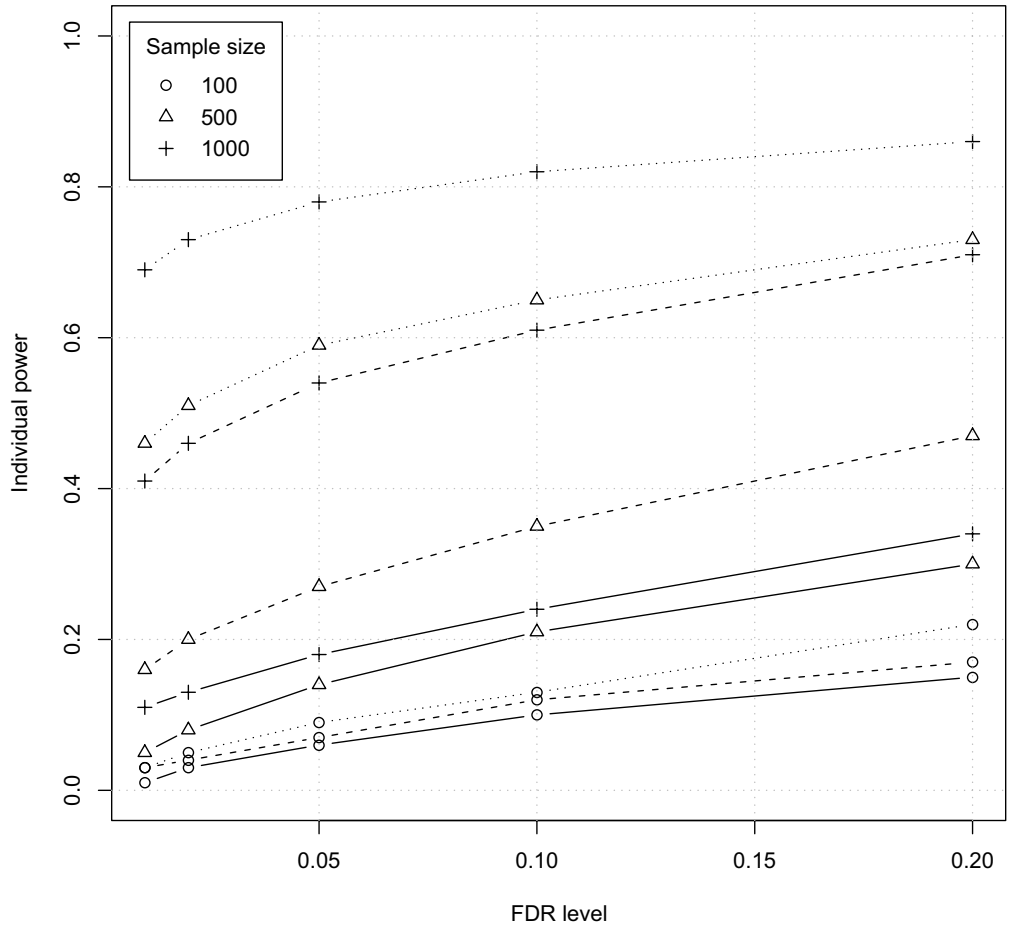

Figure 5: Evolution of individual power with FDR level (q-value correction), for different sample sizes n (different point shapes) and different differential effect $\frac{\Delta}{\sigma}$. $p_{rep} = 0.5$. Solid lines correspond to $\frac{\Delta}{\sigma} = 0.3$, dashed lines correspond to $\frac{\Delta}{\sigma} = 0.5$ and dotted lines correspond to $\frac{\Delta}{\sigma} = 0.75$

Figure 5 examines the link between power and type 1 error. This plot represents the variation of power with different FDR control levels. As expected, the more we accept false positive conclusions, the better the power. The behavior of the different curves is very similar from one differential effect to the other. Furthermore, this plot shows the major effect of the sample size n and the differential effect $\frac{\Delta}{\sigma}$ on power results, regardless of the accepted FDR level. Considering a change of parameter n or of the accepted FDR level, the net result on power is always better through an increase of n than through an increase of the accepted FDR.

Discussion

The simulation study presented here gives insight into the ability to identify biomarkers in the frame of mass-spectrometry proteomics. Results focus on the statistical power of MS experiments at two analysis levels (before and after mass-spectrometry) and for different experimental settings (number of subjects

n and subjects' repartition between groups p_{rep}), with or without FDR control.

The key message of our results is that statistical power in mass-spectrometry biomarker identification experiments is very low. Classical experiments led up to now dealt with about one hundred of subjects at most. In this setting, we see that the chance to identify individual biomarkers is low (around 10% at most). Comparison to the individual power seen in group-label permuted data showed the same level of power. Experiments with such difficult settings (low sample sizes), for low differential effects, are not informative. When trying to generalize this result to the detection of all 8 simulated DEP, we end up with a value of $(0.10)^8 = 10^{-8}$, meaning it is virtually impossible to detect all 8 DEP in such setting. Sample sizes clearly must be increased, but we can also see that an equal repartition between groups allows a gain in power that is most likely less costly for the experiment than the inclusion of more subjects. Balanced designs must be sought in order to achieve the best power available given all other parameters. This is coherent with results from non-omics studies, offering the best power results with balanced designs, for a fixed sample size n.

Our results put into light a sequential power loss explaining the low power. This sequential loss provides insight into the strategies that can be set up in order to gain power. Indeed, power is impacted at several levels. If we ignore the instrumental variability, power results are shown to be good, reaching values greater than 80% in many of the settings we considered. FDR control, in that situation, does not lead to a high power loss. The ratio between the naive and the FDR approaches is close to 1. The necessary consideration of instrumental variability, however, induces a major power drop. This is the crucial point of this power investigation, e.g. with results divided by 2 from 98% to as low as 50% for $\frac{\Delta}{\sigma} = 0.3$, $n = 1000$ and $p_{rep} = 0.5$. In this situation incorporating both components of variability, the FDR control induces a much larger power drop than with the naive approach. Power falls to 18% for the same simulation as previously considered, with a ratio between the naive and the FDR approach of less than 0.4. There exists a novel *detrimental synergy* between instrumental variability and FDR control. To sum up, taking into account the three important determinants of power devised previously (biological and instrumental variabilities, multiple testing strategies) yields a power collapse. This collapse must be carefully considered for later MS studies.

These results point out the general strategy to improve power, reminding us that variabilities should be minimized. Increasing sample sizes is a possible way to reduce variabilities, but a complementary approach can also be found in improved experimental designs. The use of matched designs would help in reducing the biological variability and should be considered when studies offer this possibility. As for the instrumental variability, including technical repetitions should allow to reduce it. This would also allow to use mixed models, as in [6], thereby allowing to discriminate between biological and instrumental variabilities. Finally, improving the mass spectrometry experiment itself will also lead to a reduced instrumental variability. This also provides a way to improve power.

The very definition of statistical power is an issue when it comes to MS data. While this definition is straightforward in the single testing setting, it raises new question in multiple testing settings. In this study, we decided to use two simple measures of power: individual power and relaxed power. The first one is the most stringent. It is also equivalent to $\frac{E(S)}{m_1}$ in our setting where all differential effects are equal within a simulation. These two measures taken together give a general overview of the possibility to extract some positive knowledge from a MS dataset. We can see from our results that experiments with low individual power are still able to identify at least one biomarker, but not every single one. However, further research are required to compare power measures in multiple testing contexts. This issue, exemplified here in proteomics, also extends to other omics, as Tsai et al. [15] showed for the field of transcriptomics by showing the dependence between power measures and sample size estimations.

The differential effect, defining the alternative hypothesis, affects the statistical power by definition.

We here chose to use a differential effect $\frac{\Delta}{\sigma} = 0.3$. We do not expect newly identified biomarkers to show large effect sizes. We therefore think this small value should be taken as a reference to design studies. This assumption also explains the low power level shown in this simulation study, contrasting with more optimistic results from a recent study[6], where differential effects are assumed larger (typically with a $\frac{\Delta}{\sigma} > 1$). We do not investigate the same class of biomarkers. Furthermore, while results in [6] interestingly introduce the two components of variability, the respective effect of each component on power is not investigated. We here show how this investigation can lead to strategies to improve power.

While these power questions are a major concern for all MS technologies, the focus of this article was put on MALDI and SELDI-TOF instruments. A large body of studies have indeed relied on MALDI or SELDI-TOF instruments, making it critical to investigate their results, as presented here. Electro Spray Ionization (ESI) combined with Liquid-Chromatography MS (LC-MS) is increasingly more used in biomedical proteomics and offers an alternative to the technologies under focus in this study. The general frame of the presented work could be used to investigate the potential added value of the LC-MS technology.

Simulations imply the assumption of various hypotheses concerning samples and mass-spectrometry. Spectra with properties similar to real mass-spectrometry assays were simulated. However, the simulations described here deliberately do not contain a variability on peak positions within an experiment. As a consequence, we do not have to use alignment algorithms in the preprocessing steps. While alignment is an essential step of the preprocessing, spectra alignment depends primarily on data quality, which can be improved at the data acquisition step. The power results seen in this study therefore apply to well-aligned data and degraded power should be expected from studies including poorly-aligned spectra, reinforcing the importance of better experimental designs.

Conclusions

The power results presented here allow to understand better our inability to identify common sets of biomarkers from one experiment to the other. Proposed experimental designs simply do not offer enough power to make sure we identify the majority of (or even some) biomarkers in each experiment. If we hypothesize that there is more than one biomarker to find in a study, then we can only see part of the list each time and miss what a previous experiment found. This issue was already pointed out in the field of transcriptomic [16, 17]. Similarly, the common identification of inflammation phase proteins[18, 19, 20] is not surprising. These proteins must exhibit very large differential effect between groups, thus requiring lower sample sizes for their detection. Experiments led up to now are not calibrated to identify potential specific biomarkers, very likely to have small differential effect about $\frac{\Delta}{\sigma} = 0.3$, because of the aforementioned detrimental synergy between instrumental variability and FDR control. Sample sizes should be increased and instrumental variability should be lowered in order to alleviate the effects of this detrimental synergy.

Authors contributions

PR and PD designed the study; TJ performed the simulations, analyzed the data and drafted the manuscript; PR, DMB and TJ participated in the writing of the manuscript. All authors read and approved the final manuscript.

Acknowledgements

The authors wish to thank Caroline Truntzer for valuable discussions.

References

[1] L. C. Whelan, K. A R Power, D. T. McDowell, J. Kennedy, and W. M. Gallagher. Applications of seldi-ms technology in oncology. *J Cell Mol Med*, 12(5A):1535–1547, 2008.

[2] M. A. Karpova, S. A. Moshkovskii, I. Y. Toropygin, and A. I. Archakov. Cancer-specific maldi-tof profiles of blood serum and plasma: Biological meaning and perspectives. *J Proteomics*, Sep 2009.

[3] Mei-Ling Ting Lee and G. A. Whitmore. Power and sample size for dna microarray studies. *Stat Med*, 21(23):3543–3570, Dec 2002.

[4] Yudi Pawitan, Stefano Calza, and Alexander Ploner. Estimation of false discovery proportion under general dependence. *Bioinformatics*, 22(24):3025–3031, Dec 2006.

[5] Caroline Truntzer, Delphine Maucort-Boulch, and Pascal Roy. Comparative optimism in models involving both classical clinical and gene expression information. *BMC Bioinformatics*, 9:434, 2008.

[6] David A Cairns, Jennifer H Barrett, Lucinda J Billingham, Anthea J Stanley, George Xinarianos, John K Field, Phillip J Johnson, Peter J Selby, and Rosamonde E Banks. Sample size determination in clinical proteomic profiling experiments using mass spectrometry for class comparison. *Proteomics*, 9(1):74–86, Jan 2009.

[7] Jakob Albrethsen. Reproducibility in protein profiling by maldi-tof mass spectrometry. *Clin Chem*, 53(5):852–858, May 2007.

[8] Catherine Mercier, Caroline Truntzer, Delphine Pecqueur, Jean-Pascal Gimeno, Guillaume Belz, and Pascal Roy. Mixed-model of anova for measurement reproducibility in proteomics. *J Proteomics*, 72(6):974–981, Aug 2009.

[9] Bart J A Mertens. Proteomic diagnosis competition: design, concepts, participants and first results. *J Proteomics*, 72(5):785–790, Jul 2009.

[10] Jeffrey S Morris, Kevin R Coombes, John Koomen, Keith A Baggerly, and Ryuji Kobayashi. Feature extraction and quantification for mass spectrometry in biomedical applications using the mean spectrum. *Bioinformatics*, 21(9):1764–1775, May 2005.

[11] R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2008. ISBN 3-900051-07-0.

[12] Xiaochun Li. *PROcess: Ciphergen SELDI-TOF Processing*, 2005. R package version 1.12.0.

[13] Kevin R Coombes, Spiridon Tsavachidis, Jeffrey S Morris, Keith A Baggerly, Mien-Chie Hung, and Henry M Kuerer. Improved peak detection and quantification of mass spectrometry data acquired from surface-enhanced laser desorption and ionization by denoising spectra with the undecimated discrete wavelet transform. *Proteomics*, 5(16):4107–4117, Nov 2005.

[14] JD Storey. A direct approach to false discovery rates JR Stat. *Journal of the Royal Statistical Society. Series B (Methodological)*, 64:479, 2002.

[15] Chen-An Tsai, Sue-Jane Wang, Dung-Tsa Chen, and James J Chen. Sample size for gene expression microarray experiments. *Bioinformatics*, 21(8):1502–1508, Apr 2005.

[16] Liat Ein-Dor, Itai Kela, Gad Getz, David Givol, and Eytan Domany. Outcome signature genes in breast cancer: is there a unique set? *Bioinformatics*, 21(2):171–178, Jan 2005.

[17] Liat Ein-Dor, Or Zuk, and Eytan Domany. Thousands of samples are needed to generate a robust gene list for predicting outcome in cancer. *Proc Natl Acad Sci U S A*, 103(15):5923–5928, Apr 2006.

[18] Eleftherios P Diamandis. Mass spectrometry as a diagnostic and a cancer biomarker discovery tool: opportunities and potential limitations. *Mol Cell Proteomics*, 3(4):367–378, Apr 2004.

[19] Eleftherios P Diamandis and Da-Elene van der Merwe. Plasma protein profiling by mass spectrometry for cancer diagnosis: opportunities and limitations. *Clin Cancer Res*, 11(3):963–965, Feb 2005.

[20] Glen L Hortin. The maldi-tof mass spectrometric view of the plasma proteome and peptidome. *Clin Chem*, 52(7):1223–1237, Jul 2006.

Chapitre 4

Détection des pics et erreurs statistiques

4.1 Problématique

Les déterminants de la puissance classiques ont été investigués dans le chapitre 3. Ces déterminants ne sont pas uniques, mais ils représentent un ensemble de paramètres importants à contrôler lors de l'élaboration du plan d'expérience et de la détermination des objectifs. Toutefois, les données de spectrométrie de masse font apparaître de nouveaux déterminants de la puissance, qui ne sont pas liés au plan d'expérience.

Le prétraitement des spectres, comme toute technique de prétraitement, influence la qualité des données et par conséquent la puissance. Si cette remarque paraît triviale dans un contexte général, elle prend un sens très précis en spectrométrie de masse. En effet, le prétraitement y a pour particularité de comprendre une étape de détection des variables d'intérêt. Cette étape de détection peut conduire à deux types d'erreur : on peut soit inclure des variables en trop (qui n'ont pas d'existence réelle et sont des artefacts du signal), soit omettre d'inclure des variables réelles (qui correspondent bien à des peptides mais n'ont pas été retenues comme telles). Cela conduit respectivement à la création du groupe des *faux pics* et du groupe des *pics fantômes*. Ces groupes nouveaux résultent de l'imperfection des algorithmes de prétraitement.

Cette situation particulière ne rend pas directement applicable le classique paradigme des tests binaires, développés dans la section 2.3.1. Les variables en trop, correspondant à de faux pics, n'ont pas place dans le classique tableau 2x2 (tableau 2.2) issu de ce paradigme, de même que les variables omises. Or, les stratégies de contrôle du risque de première espèce, notamment du FDR, sont construites autour de ce tableau. Sa modification va donc engendrer des perturbations dans l'estimation du FDR et de la puissance.

L'effet des faux pics et des pics fantômes sur le contrôle du FDR et sur la puissance est investigué dans cette partie. La meilleure compréhension de cet effet doit permettre d'estimer l'effet de l'algorithme de prétraitement pour optimiser son choix d'une part et éventuellement améliorer ces algorithmes d'autre part.

4.2 Méthodes

L'exploration de l'effet des pics poubelles et pics fantômes sur les risques de première et deuxième espèce repose sur l'étude analytique des expressions du FDR et de la puissance modifiées lorsqu'on intègre les faux pics et les pics fantômes.

4.3 Résultats annexes

Le travail sur une redéfinition du FDR mené pour l'article joint conduit à une formule du FDR présenté dans l'équation (4.1) (ne prenant ici en compte ni les faux pics, ni les pics fantômes). Cette équation fait apparaître le lien ϕ entre le risque de première espèce et la puissance, avec $\phi(\alpha) = 1 - \beta$. Cette fonction ϕ dépend du test et de l'hypothèse alternative choisis.

$$\frac{\pi_0 \cdot \alpha}{\pi_0 \cdot \alpha + (1 - \pi_0) \cdot \phi(\alpha)} \tag{4.1}$$

Ecrire le FDR sous cette forme fait toutefois apparaître une problématique non investiguée dans la littérature, précisément le lien entre FDR et hypothèse alternative. Le contrôle du FDR se définit alors par rapport à une hypothèse alternative choisie. Une valeur de FDR n'a pas le même sens selon cette hypothèse.

On considère ici le simple test de t pour la comparaison de moyennes μ_1 et μ_2 entre deux groupes, avec $\Delta = \mu_1 - \mu_2$ et un écart type σ de la mesure identique pour les deux groupes. En utilisant l'effet différentiel $\frac{\Delta}{\sigma}$, on peut définir l'hypothèse alternative comme une distribution de Student décentrée en $\lambda = \frac{\Delta}{\sigma} \cdot \sqrt{\frac{n}{2}}$, où n est le nombre d'individus inclus dans l'étude. En pratique, le choix de la valeur limite de l'erreur de type 1 individuelle, guidé par le FDR, est très différent selon l'hypothèse alternative choisie. La figure 4.1 présente l'évolution de α, valeur limite de l'erreur de type 1 individuelle, en fonction de l'effet différentiel $\frac{\Delta}{\sigma}$. Pour un effet différentiel en dessous de 0.4, le contrôle du FDR devient très difficile et oblige à accepter des risques de première espèce individuels en décroissance exponentielle (la courbe est quasi-linéaire en échelle logarithmique autour d'un effet différentiel de 0.2).

En outre, l'efficacité du test statistique utilisé (traduit par exemple par sa courbe ROC) doit être pris en compte. Le test de t, utilisé ici, est réputé puissant. D'autres tests présentent des efficacités différentes, induisant vraisemblablement des risques de première espèce individuels plus petits encore à effet différentiel fixé, pour un FDR contrôlé à un niveau voulu. La courbe présentée en figure 4.1 peut ainsi avoir une pente plus forte encore.

La puissance individuelle (équivalente à la puissance moyenne si les puissances individuelles sont identiques et indépendantes) est présentée parallèlement, montrant la perte de puissance attendue pour un même niveau de contrôle du FDR. Lorsque l'effet différentiel devient petit, le contrôle du FDR implique un contrôle du risque de première espèce individuel beaucoup plus fort que l'approche Bonferonni, pourtant réputée plus conservative. Cet effet apparaît car la puissance individuelle associée décroit de façon importante.

L'écriture du FDR sous la forme $\frac{1}{1+\frac{S}{V}}$ met en évidence l'origine de cet effet étonnant. Contrôler le FDR sous la traditionnelle limite de 5% implique en effet un rapport $\frac{S}{V} \geq 19$. En se remémorant les courbes de la figure 2.2, on traduit ce rapport sous la forme du rapport

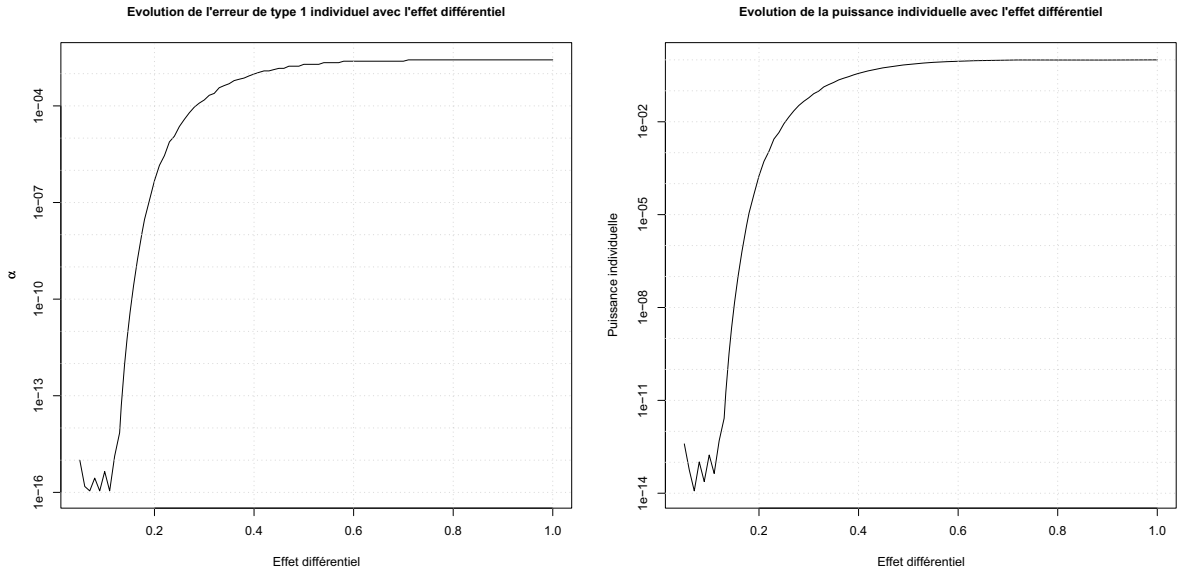

FIG. 4.1 – Niveau d'erreur de type 1 individuel accepté pour un FDR fixé à 5%, pour différents effets différentiels définissant l'hypothèse alternative du test. Echelle logarithmique sur l'axe des ordonnées.

des intégrales à droite de la limite définie par le risque de première espèce individuelle, pour les courbes de l'hypothèse nulle et de l'hypothèse alternative. Or, ce rapport décroit avec la décroissance de l'effet différentiel, c'est à dire avec le rapprochement des courbes des hypothèses nulle et alternative, comme le montre la figure 4.2, toujours dans le cas particulier du test de t

Ce problème émerge comme on l'a vu pour des valeurs d'effet différentiel inférieure à 0.4. En spectrométrie de masse, comme on l'a vu précédemment (voir notamment la figure 3.2), l'effet différentiel peut être de cet ordre de grandeur du fait notamment de l'erreur de mesure. Le contrôle du FDR pour de petits effets différentiels, dans les études de spectrométrie, est donc un réel facteur de perte de puissance.

4.4 Discussion

La transcriptomique fournit bien souvent une base de travail pertinente pour l'étude des données protéomiques. De nombreux concepts de traitement du signal et d'analyse statistique ont été étendus à l'analyse de données de spectrométrie de masse. Les résultats présentés ici mettent en évidence le besoin constant de s'interroger sur la validité de telles extensions.

En pratique, les imperfections des algorithmes de détection de pics conduisent à limiter la puissance statistique des études. Une partie de cette limitation est triviale et reflète juste l'effet de la non détection d'un pic, ne permettant évidemment pas de conclure quant à un différentiel d'expression pour un tel pic fantôme. L'autre partie est plus subtile et fait intervenir les méthodes de contrôle du risque de première espèce (FDR, en l'occurence). Les faux pics interviennent dans l'expression de ce risque. Leur effet rejoint celui des pics non différentiels mis en évidence par Lee [47]. Dans les deux cas, l'approche proposée ici permet de quantifier l'effet des imperfections de détection des pics.

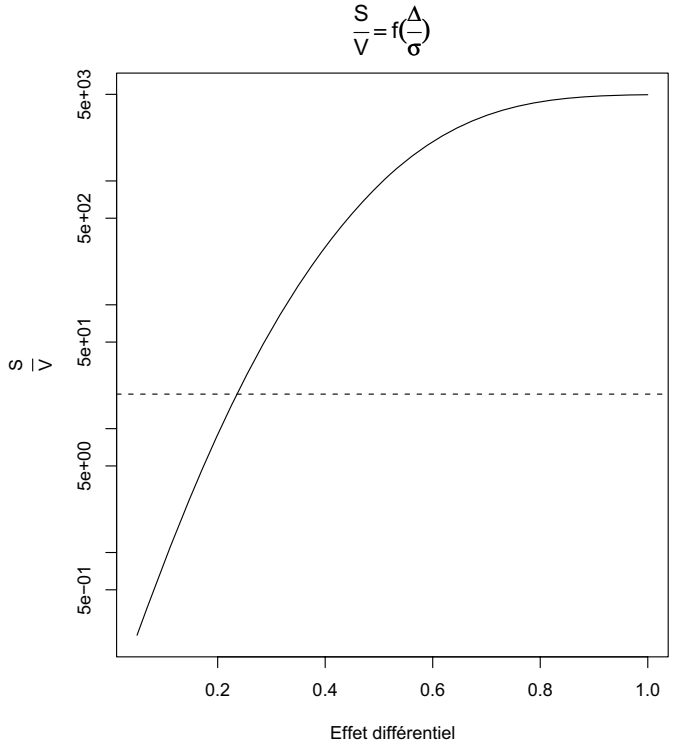

FIG. 4.2 – Evolution du rapport $\frac{S}{V}$ en fonction de l'effet différentiel, dans une situation avec $m_0 = 200$ soit 200 protéines non différentielles, $m_1 = 10$ soit 10 protéines différentielles, $\alpha = 10^{-4}$, pour $n = 200$ sujets inclus dans l'étude (100 dans chaque groupe).

Cette étude fait ressortir un niveau supplémentaire de perte de puissance pour les études de spectrométrie de masse. Les algorithmes de prétraitement eux-mêmes ont un effet important, ici par le biais d'interactions avec le contrôle du FDR. Ce contrôle du FDR, en dehors des imperfections dans la détection des pics, semble aussi pouvoir jouer un rôle négatif sur la puissance. Le concept même de FDR est donc à valider lorsque les effets différentiels sont petits, typiquement inférieurs à 0.4 pour le simple test de t.

Effects of garbage and ghost peaks in mass-spectrometry

Thomas Jouve [*], Delphine Maucort-Boulch [†], Patrick Ducoroy [‡], Pascal Roy [§]

Abstract

Mass-spectrometry data belong to the omics data, thereby requiring close attention to multiple-testing problems. These problems were widely investigated for transcriptomic data, although many fields of research remain open. Mass-spectrometry (MS) data face the experimenter with the new question of peaks detection. False peaks can be picked (*garbage*) and true peaks can be missed (*ghost*). In this article, true spectra are used to show the reality of a trade off between garbage and ghost peaks. The combined effect of garbage and ghost peaks on type 1 and type 2 errors, using new analytical writings of these errors, is investigated. The inclusion of garbage and ghost peaks artificially leads to an increase of the FDR and a decrease of statistical power. These results offer new insights on MS data analyses, showing the need to take into account peak detection issues in order to properly design future studies.

1 Introduction

The so-called field of *omic* technologies has considerably developed in the last ten years. New high-throughput technologies have led to new fields of investigation. In particular, biomarker discovery benefits from these high-throughput data. It aims at discovering new molecular markers of a biological state, e.g. a pathogenic process or a pharmacological response, in order to enhance diagnosis or prognosis, among other applications.

Biomarker research in omics often refers to the classical binary outcome table presented in table 1. The key idea is to distinguish Differentially Expressed Genes or Peptides (DEP) from Non Differentially Expressed Genes or Peptides (NDEP) between two (or more) groups of individuals, using a binary conclusion test (either accepting or rejecting the null hypothesis denoted as H_0) [1].

[*]thomas.jouve@chu-lyon.fr
[†]delphine.maucort-boulch@chu-lyon.fr
[‡]patrick.ducoroy@clipproteomic.fr
[§]pascal.roy@chu-lyon.fr

Given the test result and the actual expression status of each candidate gene or protein, it is then possible to sort out these candidates in the different cells of the binary outcome table.

Furthermore, omic data sets face researchers with the so-called *small n large p* problem, where n refers to the number of subjects in the experiment and p refers to the number of covariates under study. Statistical tools were therefore developed in order to deal with this multiple testing situation. The classical False Discovery Rate (FDR) can be easily defined using table 1, as in equation (1). The FDR focuses on the second column of table 1.

$$FDR = E(\frac{V^0}{R^0}) \tag{1}$$

The FDR is widely used as a controled quantity to limit the number of false positives in multiple testing settings. This control can be achieved through different algorithms developed in the past years[2, 3, 4, 5]. Similarly, statistical power can be defined in different ways using quantities from table 1, e.g. average power as in equation (2), used in this article. Statistical power focuses on the second line of table 1.

$$P = \frac{E(S^0)}{m_1} \tag{2}$$

	Conclusion		
	Accept H0	Reject H0	Total
Truth			
NDEP	U^0	V^0	m_0
DEP	T^0	S^0	m_1
Total		R^0	m

Table 1: Classical outcome of a binary decision (e.g. transcriptomic data)

These statistical procedures were all developed for transcriptomic studies, mostly based on microarrays. Microarrays are designed using specific probes that allow the study of a predefined set of genes. Table 1 was introduced in this context, where the list of genes under study is available. Proteomics, contrary to transcriptomics, uses technologies (2D-gel, mass spectrometry) that do not set features of interest *a priori*. In proteomics, the output is a set of features (e.g. spot on a gel, or peak in a mass spectrum) that remain to be labeled as a peptide, or as a protein (the term

protein will be used hereafter to refer to any of both). Truly high-throughput methods like mass-spectrometry (MS) require automated feature detection algorithms. Many preprocessing algorithms were developed [6] that i) locate peaks in spectra ii) associate a mass-to-charge ratio (m/z) label and an intensity to each of these peaks. The m/z label serves as a biological identifier. These two steps were collectively referred to as *feature detection*. In this context, it corresponds to the finding of a peak and its association to a m/z label. As good as the preprocessing algorithms might be, room for imperfections remains in this detection process. These imperfections are the motivation of this article.

In this article, we investigate the effect of an imperfect peak detection algorithm on FDR and on statistical power in mass-spectrometry data. A real MS data set is used to investigate the characteristics of the detection algorithms, as described in section 2.1. A new paradigm for MS data binary testing when feature detection algorithms are imperfect is outlined in section 2.2, leading to new writings of FDR and power presented in section 2.3 and 2.4. Data supporting the imperfect feature detection algorithms are presented in section 3, together with changes implied by these new writings. Reading guidelines and implications for further research are exposed in section 4.

2 Material and methods

When dealing with mass spectrometry data, two categories of features are implicitly taken into account:

- false peaks, not corresponding to a peptide, are part of the found peaks and can either be said differentially expressed or not. We call these peaks *garbage peaks* as their number should be as limited as possible.
- some true peptides effectively present in fluid samples fail to be detected; these unidentified peptides can either be truly differentially expressed or not, but the possible outcome of a test on the corresponding peaks is necessarily unknown. We call these peaks *ghost peaks*, as they are not found using the chosen preprocessing algorithms.

The preprocessing algorithms can either be made very sensitive to small peaks or very specific and retain only a fraction of true peaks. A trade-off between sensitivity (more garbage peaks) and specificity (more ghost peaks) must therefore be found and considered carefully.

2.1 Real data analysis

The imperfections of feature detection algorithms are not new: Emanuele et al.[7] already pointed out that only a fraction of all true peaks were detected by the best available algorihms. Two real data sets were analysed in order to investigate further the imperfections of these algorithms. The Hodgkin data set contains 192 spectra from blood samples of 48 subjects suffering from Hodgkin's lymphoma. A follow up study focused on survival of each subject. The calibration data set is a set of 44 calibration spectra, based on samples containing only 10 known proteins recommended by the mass spectrometer's manufacturer for instrument calibration. An example spectrum from each data set is provided in figure 1. The same mass spectrometer was used for both data acquisitions.

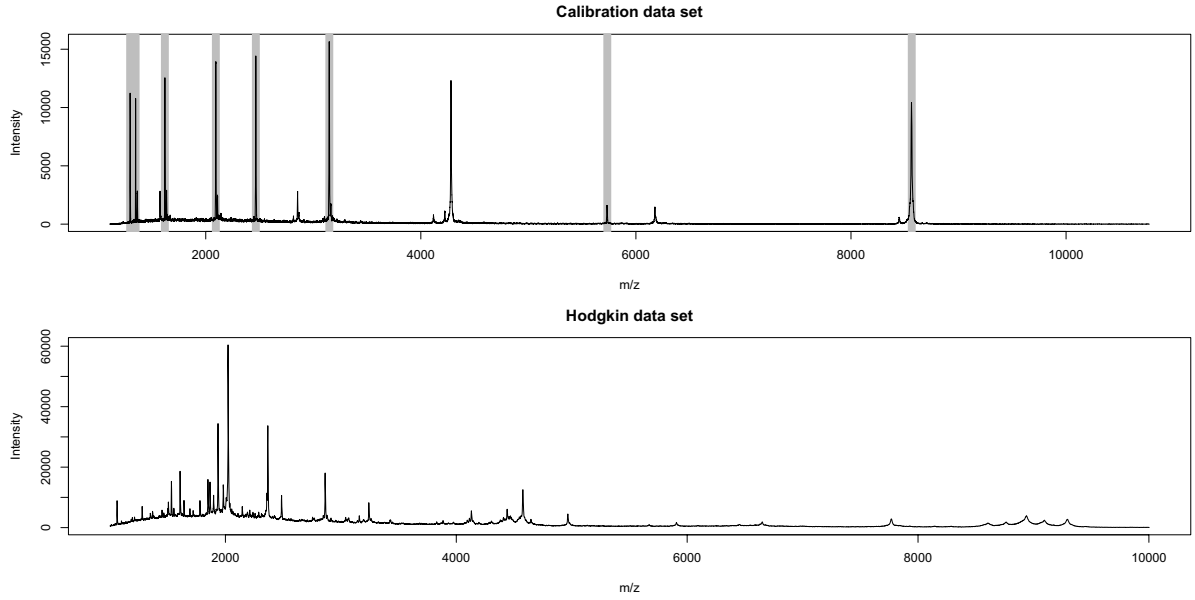

Figure 1: Spectrum examples from calibration and Hodgkin data sets. The true protein content for the calibration data set is outlined by peaks with a grey box. True protein content for Hodgkin data set is unknown.

Preprocessing was performed as follows :

- Baseline substraction *via* loess regression from the PROcess package [8]
- Smoothing using the Undecimated Wavelet Transform, as recommended by Coombes et al. [9]. Noise is defined as the filtered signal from this procedure.
- Normalization by the Total Ion Current (TIC)

- Peak detection on the mean of raw spectra, as recommended by Morris et al. [10], using local maxima whose intensitiy lies over a chosen signal-to-noise (SNR) threshold.
- Intensity measurement by the Area Under the Curve (AUC)

The calibration data provides the knowledge of true and false peaks. The evolution of the number of detected peaks with the SNR threshold was therefore investigated, together with the number of garbage peaks. This parallel investigation of the number of ghost and garbage peaks can help in the choice of an optimal SNR threshold. Furthermore, the comparison of intensity profiles is possible between the calibration and Hodgkin data sets, putting into perspective this choice of a SNR threshold.

2.2 Redefining the paradigm

As was already mentioned, changes in the classical table 1 are required, defining a new paradigm for test outcomes. Table 2 is introduced and presents the new test outcomes. The original table 1 only takes into account S, T, U and V from this new table. The output of any preprocessing algorithm is limited to the subtable outlined in gray, while the desired table is limited to the two first lines. Tables 1 and 2 are equivalent if no peak is forgotten (no ghost peak) and no false peak is kept for the analysis (no garbage peak).

	Found peaks		Ghost peaks		Total
	Accept H0	Reject H0	Accept H0	Reject H0	
NDEP	U	V	U'	V'	m_0
DEP	T	S	T'	S'	m_1
Garbage peaks	G_0	G_1			G
Total		R			m+G

Table 2: Modified outcome of a binary decision for *a priori* unknown features.

In table 2, garbage peaks will add to the true peaks, both in the category of H_0 acceptance and rejection. The total for each category (or margin) is therefore different when garbage peaks are included. These garbage peaks appear at a hardly definable rate, depending both on data quality and on the preprocessing algorithms. The more peaks are kept in the analysis, the more garbage peaks will be included as well, but the exact link can not be written *a priori*. The total number of garbage peaks, G, will therefore be used thereafter and should be set according to the user's preliminary knowledge of these garbage peaks in the chosen experimental setting.

Similarly, unidentified peptides must be taken into account, introducing an adjacent ghost table. Line and column margins are influenced by the number of ghost peaks. Repartition between S, T, U, V and their primed counterparts is defined by the peak detection ability of the preprocessing algorithms. This peak detection ability, here again, depends on data quality and on the pretreatment algorithms. The total number of potential true discoveries can be written as $S^0 = S + S'$, and so on for T, U and V. The peak detection ability p_d is then written as $p_d = \frac{V}{V^0} = \frac{S}{S^0} = \frac{T}{T^0} = \frac{U}{U^0}$, assuming an identical detection ability for all true peaks, regardless of its test outcome or differential status. This probability p_d is also one minus the ghost proportion.

2.3 Inclusion of garbage peaks

In a first step, no ghost peak is considered. This is what can be expected from a sensitive peak-picking algorithm, not filtering out any of the true peaks but possibly with a high number of garbage peaks. In this situation, all primed quantities equal zero. The most extreme situation of this kind is to consider as peaks all local maxima in a (pretreated) spectrum. Although no true peak is likely to be missed, the number of false peaks can be fairly high.

In this setting, the number of false H_0 rejections is now $V^0 + G_1$ since all rejected garbage peaks will necessarily be false rejections. The total number of H_0 rejections therefore amounts to $R = V^0 + S + G_1$. FDR must be redefined as in equation (3).

$$FDR_{garbage} = E(\frac{V^0 + G_1}{V^0 + S^0 + G_1}) \tag{3}$$

Statistical power can then be written as in equation (4) (this writing does not differ from equation (2)).

$$P = \frac{E(S^0)}{m_1} \tag{4}$$

In order to simplify these expressions for FDR and power, V^0, S^0 and G_1 can be written using individual type 1 error α and individual statistical power $1-\beta$. With these quantities, it is possible to write $E(V^0) = m_0 \cdot \alpha$ and $E(S^0) = m_1 \cdot (1-\beta)$, assuming identical type 1 error and statistical power for each test. G_1 is also linked to the type 1 error. It amounts to the proportion of all garbage peaks for which H_0 is rejected. Garbage peaks are expected to represent white noise, with a risk of H_0 rejection therefore equal to the type 1 error, allowing to write $G_1 = \alpha \cdot G$. This leads to equations (5) and (6).

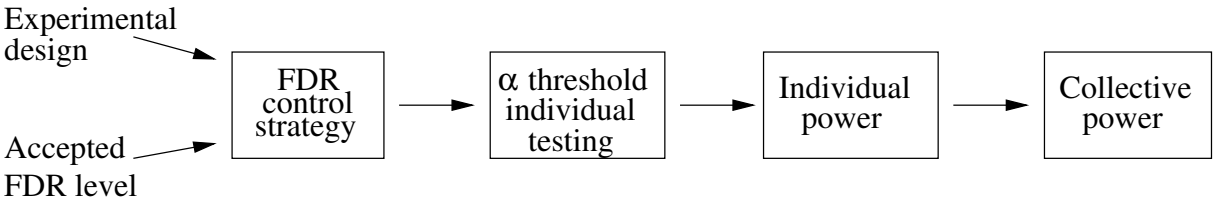

Figure 2: Causal chain between accepted FDR level and average power.

$$FDR_{garbage} = \frac{m_0 \cdot \alpha + G \cdot \alpha}{m_0 \cdot \alpha + m_1 \cdot (1-\beta) + G \cdot \alpha} \tag{5}$$

$$P = 1 - \beta \tag{6}$$

FDR estimations are directly affected by garbage peaks, whose numbers appear in this new FDR expression. $FDR_{garbage}$ equals the FDR from equation (1) when no garbage peaks are found, and tends to 1 as G tends to infinity. This means that the more garbage peaks are kept in the analysis, the higher $FDR_{garbage}$ will be.

On the other hand, statistical power is not affected directly by ghost peaks, but can be lowered when trying to control $FDR_{garbage}$ instead of the true FDR (when no ghost peak is present), leading to a too stringent control of the number of false positives. Indeed, $(1-\beta) = p(H1|DEP)$ depends on the individual type 1 error as in equation (7).

$$1 - \beta = \phi(\alpha) = 1 - F_1\left(F_0^{-1}(1-\alpha)\right) \tag{7}$$

In this equation, F_0 represents the cumulative density function (cdf) of the test statistic under the null hypothesis H_0, while F_1 represents the test statistic cdf under the alternative hypothesis H_1. Type 1 error α, in turn, depends on the accepted FDR, with link l defined by $\alpha = l(FDR)$. This dependence chain is represented in figure 2. In the simple Bonferonni setting, function l can be written as $\alpha = l(\alpha_B) = \frac{\alpha_B}{m}$, where m is the total number of tests performed, α_B is the global type 1 error, accounting for all m tests. For a fixed α_B, if the number of tests increases, α_0 decreases, leading to a loss of power, as seen in equation (7).

When using the FDR as the controled type 1 error, the link l between α and the FDR is not straightforward. Indeed, the FDR computation requires the estimation of quantities m_0 and m_1 from table 2, and of the test distributions. These quantities must therefore also be estimated in order to find the link l. A mixture model was proposed by Pawitan et al. [11], allowing to derive the desired link as in equation (8), where F_0 is the cdf of the test statistic under H_0, F is a mixture

of F_0 and F_1 (cdf under the alternative hypothesis), c is a critical value corresponding to a type 1 error at level α (e.g. $c = F_0^{-1}(1-\alpha)$ for a one-sided test).

$$FDR_{pawitan} = \frac{\pi_0 \cdot (1 - F_0(c))}{1 - F(c)} \tag{8}$$

Several articles [12, 13, 14, 15] based on equation (8) developed an expression of the link between α and FDR. In all cases, the key requirement is an expression of F_0 and F. Both functions require hypotheses made on the test distributions under H_0 and H_1. Recently, in the context of proteomics, Cairns et al. [16] proposed to use the simple definition from equation (9), where π_0 is the proportion of non-differentially expressed proteins.

$$FDR_{cairns} = \frac{\alpha \cdot \pi_0}{\alpha \cdot \pi_0 + (1-\beta) \cdot (1 - \pi_0)} \tag{9}$$

This equation is not based on test distributions. It can provide a useful approximation when both α and $1-\beta$ are set, but it fails to render the dependency between the two risks (expressed by equation (7)).

Therefore, a modified version of this last FDR definition was used, taking into account $1-\beta$ through $\phi(\alpha)$ (see equation (7)), resulting in equation (10) when garbage peaks are included.

$$FDR = \frac{\alpha \cdot \pi_0 + G \cdot \alpha}{\alpha \cdot \pi_0 + \phi(\alpha) \cdot (1 - \pi_0) + G \cdot \alpha} \tag{10}$$

Functions F_0 and F_1 must be defined to compute the associated FDR. They depend on the test performed. For a simple t-test, the cdf of F_0 is $T(n-2, 0)$ where $T(n-2, 0)$ represents the student distribution with $n-2$ degrees of freedom and a null non-centrality parameter (NCP), and the cdf of F_1 is $T(n-2, \lambda)$ where λ is some chosen NCP. In this context, the link between λ and the differential effect $\frac{\Delta}{\sigma}$, where $\Delta = \mu_0 - \mu_1$ is the difference between the group means and σ is the standard deviation, is given by equation 11, following classical derivations of the t-test statistic. We assume identical standard deviations within groups and two groups of identical sizes $n/2$.

$$\frac{\Delta}{\sigma} \cdot \sqrt{\frac{n}{4}} = \lambda \tag{11}$$

2.4 Inclusion of ghost and garbage peaks

In a second step, ghost peaks are introduced. This is the actual situation faced when analyzing mass-spectrometry data. As previously, it is possible to re-write FDR and power using fewer pa-

rameters, by introducing individual type 1 error and individual power, but also a peak detection probability. Ghost peaks must be considered when the peak detection ability of the chosen preprocessing algorithms, on a given data set, is not perfect. The differences between false positive found and not found (ghost), $V = V^0 - V'$, and true positive found and not found (ghost), $S = S^0 - S'$, can therefore be re-expressed using p_d, the probability of peak detection, regardless of its category. The introduction of p_d allows the definitions of equations (12) and (13).

$$E(V) = p_d \cdot m_0 \cdot \alpha \tag{12}$$

$$E(S) = p_d \cdot m_1 \cdot (1 - \beta) \tag{13}$$

A non-trivial link between p_d and G_1 exists. It depends both on characteristics of the mass spectrometer itself and of the preprocessing algorithms used. This link imposes constraints on the choice of the latter when the former is set, and reciprocally.

Using p_d and G leads to an expression of FDR_{gg} (standing for both Garbage and Ghost peaks consideration) and P_{gg} that only depend on two parameters linked to the preprocessing strategy, as described in equation (14) and (15).

$$FDR_{gg} = \frac{p_d \cdot m_0 \cdot \alpha + G \cdot \alpha}{p_d \cdot (m_0 \cdot \alpha + m_1 \cdot \phi(\alpha)) + G \cdot \alpha} \tag{14}$$

$$P_{gg} = p_d \cdot (1 - \beta) = p_d \cdot \phi(\alpha) \tag{15}$$

$$\frac{dFDR_{gg}}{dp_d} = \frac{-G \cdot \alpha(m_1 \cdot \phi(\alpha))}{(p_d \cdot (m_0 \cdot \alpha + m_1 \cdot \phi(\alpha)) + G \cdot \alpha)^2} \tag{16}$$

Equation (14) and its derivative presented in equation (16) show that:

- the general behavior of FDR_{gg} with G remains the same as with $FDR_{garbage}$,
- FDR_{gg} does not depend on p_d when no garbage peak is present, since the detection ability will equally apply to true and false discoveries,
- FDR_{gg} decreases monotonically in p_d, as shown by its derivative with respect to p_d. Indeed, all quantities of this derivative's numerator are positive by definition. Interestingly, the slope of FDR_{gg} along p_d is steeper with bigger values of G_1 or S.

Equation (15) directly shows that average power is a linear function of p_d, increasing with this probability as one would expect: the more true peaks we find in spectra, the higher power gets. However, garbage peaks also impact statistical power. As in the *garbage only* setting, this dependency can be seen through the $\phi(\alpha)$ function, from equation (7).

In order to investigate numerically the impact of garbage and ghost peaks on statistical power, the following strategy was used :

1. Set accepted FDR level, F_0 and F_1 (requiring setting the number of samples n and differential effect)
2. Set values for G and p_d
3. Solve equation (14) for α, using a bisection method for root search
4. Compute average power P_{gg} using equations (15) and (7)

The classical t-test setting was used, with test distributions $T(n-2,0)$ under H_0 and $T(n-2,\lambda)$ under H_1, where λ is the chosen non-centrality parameter of the T distribution. Iterating the process described here for various values of G and p_d allowed to explore the effect of both parameters. Obtained results may be compared to the result for the ideal situation (where $G = 0$ and $p_d = 1$).

The choice of the non-centrality parameter is arbitrary and should depend on the profile of the differentially expressed proteins. The actual quantity of interest is the differential effect $\frac{\Delta}{\sigma}$, depending on the biomarker's properties and on the measurement variability (adding to the biological variability). If we hypothesize a differential effect $\frac{\Delta}{\sigma} = 0.3$, the value of λ must be set to $\lambda = 0.3 \cdot \sqrt{\frac{200}{4}} = 2.12$ for $n = 200$ subjects (100 in each group). This value of λ was used in this study.

3 Results

3.1 Evidences for garbage and ghost peaks

Figure 3 presents the intensity histograms of detected peaks for the calibration and the Hodgkin data sets. These intensities do not follow the same distribution for the two data sets. The calibration data set shows two apparently independent distributions, corresponding to true and garbage peaks (this was verified using the true knowledge available from this data). The Hodgkin data set, on the contrary, does not show the same bimodal profile. It is not possible to isolate two different

groups of intensities. While it was easy for the calibration data set to choose an intensity threshold separating true and garbage peaks, it is not possible to find a simple threshold for the Hodgkin data set. True biological samples will most likely exhibit the same profile as the Hodgkin data set, making the separation between true and garbage peaks a more difficult task.

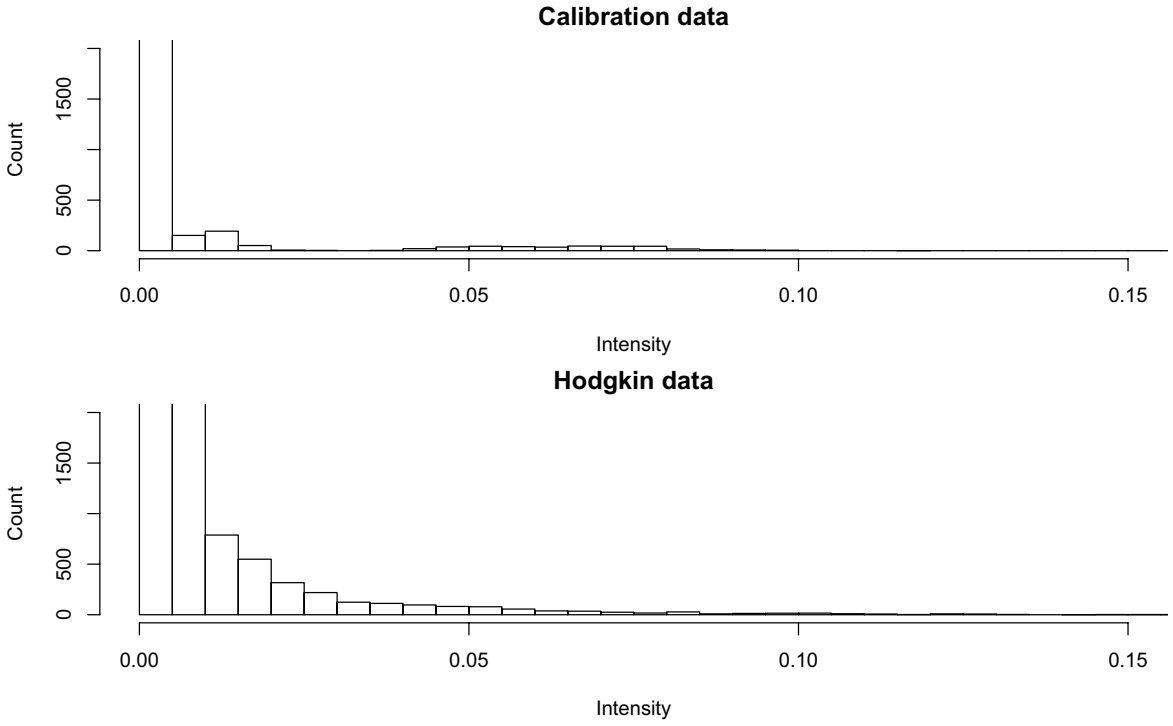

Figure 3: Histogram of peak intensities for calibration and hodgkin data sets. Intensities for the calibration data set present a gap around intensities ~ 0.03, explaining the separability of garbage and ghost peaks. This gap is not seen for the Hodgkin data set.

Figure 4 presents the proportion of garbage peaks (solid line) among all detected peaks with respect to the Signal to Noise Ratio (SNR) threshold chosen for peak filtering, and the true peak detection rate (dashed line). This figure shows that a SNR threshold filtering 90% garbage peaks (SNR=350) implies a peak detection lowered to 80%. This result is all the more significant that it is based on calibration data, where peaks distinctly emerge from the rest of the signal. In other words, even clearly defined peaks are filtered out by garbage peaks filtering methods, although the situation seemed simple from figure 3.

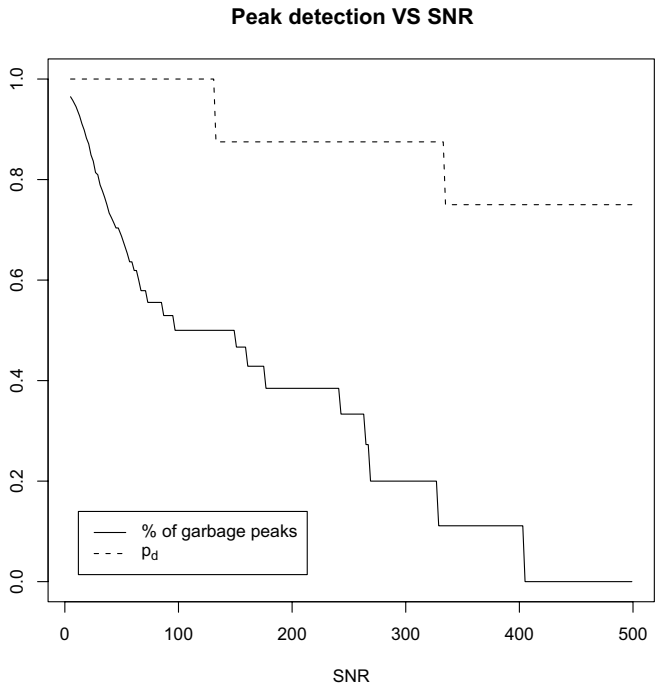

Figure 4: Evolution of the proportion of garbage peaks among all peaks and of the true peak detection rate p_d with the SNR threshold. Calibration data.

3.2 Analytical expressions investigation

The behavior of FDR_{gg}, defined in equation (14) can be investigated graphically in a three-dimensional representation of its variation with p_d and G. Figure 5 shows these variations, using arbitrary values for $m_0 = 200$, $m_1 = 10$, $\alpha = 0.001$ (corresponding to the individual type 1 error risk) and power $1 - \beta$ computed as in equation 7 with non-centrality parameter $\lambda = 2.12$ for F_1, and $n = 200$ subjects (following current studies). In this situation, the FDR reaches 8.34% for $G_1 = 0$ and $p_d = 1$. However, a notable deviation from this value is seen when either G or p_d move away from their ideal values. As recently shown [7], $p_d \simeq 0.55$ at best on SELDI/MALDI mass-spectrometry data. Using this value for p_d and $G = 70$ garbage peaks (i.e. about one garbage peak detected for three true peaks included), FDR_{gg} rises to 12.96%, although the available information from the data is the same. Correcting this deviation will require using a lower individual type 1 error (α), thereby lowering individual power (as shown in equation 7).

This effect of FDR control on power can be seen on figure 6. Here again, the quantity of interest

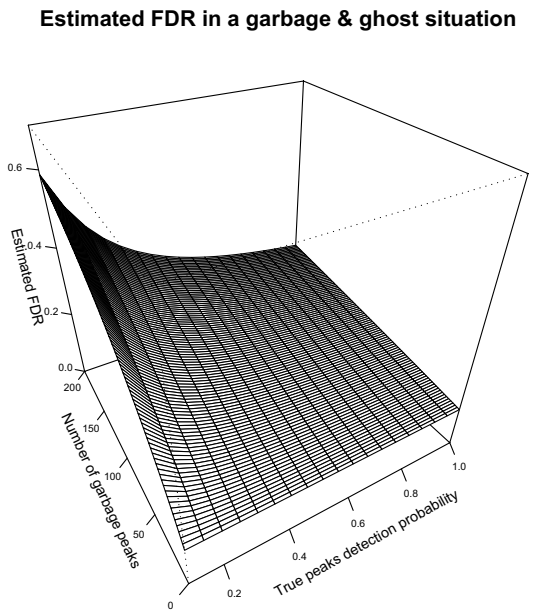

Figure 5: Evolution of $FDR_{perceived}$ with $m_0 = 200$, $m_1 = 10$, $\alpha = 0.001$ and a non-centrality parameter $\lambda = 2.12$, for $n = 200$ subjects. Actual FDR without garbage or ghost peaks is 8.34%.

is plotted against p_d and G. As expected from equation (15), power decreases linearly with p_d. This figure also shows the decrease of power with the number of garbage peaks, linked to the FDR control. Coming back to the example developed above with $p_d = 0.55$ and $G = 70$, power equals 10% although it reaches 23% in the perfect situation with $p_d = 1$ and $G = 0$.

4 Discussion

The classical paradigm for FDR definition and estimation, based on a 2 situations / 2 conclusions table, does not allow to interpret MS results correctly. Indeed, it introduces a bias in estimations of both FDR and statistical power. This bias results from the inclusion of garbage peaks and the absence of ghost peaks. The consequence of this bias is an under-estimation of power.

Investigating the distribution of intensities for found peaks on true spectra showed that garbage peaks are unavoidable on current MS medical experiments if we want to limit the number of ghost peaks. Even calibration data do not allow a perfect peak detection, without garbage or ghost peaks. This necessary trade-off between over-(garbage peaks) and under-detection (ghost peaks) requires

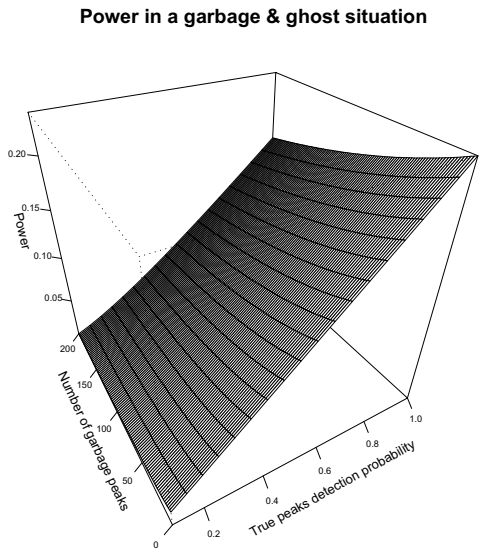

Figure 6: Evolution of P with $m_0 = 200$, $m_1 = 10$, $FDR = 5\%$ and a non-centrality parameter $\lambda = 2.12$, for $n = 200$ subjects.

the paradigm redefinition discussed in this article.

Our results show that garbage peaks and ghost peaks impact FDR. Rewriting the definition of FDR allowed us to investigate the theoretical impact of garbage and ghost peaks on this quantity. It was shown that FDR can be increased by 50% of its ideal value (in a situation with no garbage or ghost peak). Correcting the individual type 1 error to control FDR below a chosen threshold will in turn lead to a power drop. Rewriting the definition of the average statistical power, the effects of garbage and ghost peaks on this quantity were presented. A net decrease of power is seen with garbage peaks, through a suboptimal FDR control and a too stringent individual type 1 error choice, and with ghost peaks, because of their absence.

Lee et al. [17] discussed the effect of the number of non-differentially expressed genes (NDEG) on statistical power, in the field of transcriptomics. NDEG define the size of the gene space but their number can not be limited if one wants to explore a large space. They showed a decrease of power when this number of NDEG increases, while controlling the Family Wise Error Rate (FWER). This conclusion supports the results presented here, as H_0 rejections for NDEG or garbage peaks have the same effect on type 1 error rates (FWER or FDR) computations. However, the very reason why NDEG or garbage peaks occur is different. Garbage peaks appear due to instrumental and

preprocessing imperfections. Their number should be as limited as possible.

The new writings proposed for FDR and average power are useful to investigate the power loss underwent with peak detection. Therefore, it is possible to estimate the expected gain of better MS instruments and peak detection strategies. All pretreatment steps, from data acquisition to peak intensity readings, can lead to garbage and ghost peaks. A better understanding is needed as to why garbage peaks appear and ghost peaks are missed. This better understanding involves a better characterization of intensity distributions for all kind of peaks, whether true, garbage or ghost. Some peak detection algorithms make hypothesis on this topic [18]. However, the lack of a gold standard experiment for the MALDI technology provides neither an actual knowledge of these distributions nor the ability to correct our hypotheses on them. We therefore suggest to perform such a gold standard experiment, of great value for all past and future researches on MALDI-TOF MS experiments. Its data could be used to evaluate parameters p_d and G for all preprocessing algorithms, providing a common scale for comparisons of these algorithms.

Acknowledgement

The author wish to thank Hadrien Charvat for valuable discussions.

Funding

This work was supported by the french National Cancer Institute (INCa), for the project "Experimental settings analysis and experiments calibration for proteomic biomarkers identification".

References

[1] S. Dudoit, J.P. Shaffer, and J.C. Boldrick. Multiple hypothesis testing in microarray experiments. *Statistical Science*, 18(1):71–103, 2003.

[2] Y. Benjamini and Y. Hochberg. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, 57(1):289–300, 1995.

[3] Anat Reiner, Daniel Yekutieli, and Yoav Benjamini. Identifying differentially expressed genes using false discovery rate controlling procedures. *Bioinformatics*, 19(3):368–375, Feb 2003.

[4] JD Storey. A direct approach to false discovery rates JR Stat. *Journal of the Royal Statistical Society. Series B (Methodological)*, 64:479, 2002.

[5] John D Storey and Robert Tibshirani. Statistical significance for genomewide studies. *Proc Natl Acad Sci U S A*, 100(16):9440–9445, Aug 2003.

[6] Melanie Hilario, Alexandros Kalousis, Christian Pellegrini, and Markus Mueller. Processing and classification of protein mass spectra. *Mass Spectrom Rev*, 25(3):409–449, 2006.

[7] Vincent A Emanuele and Brian M Gurbaxani. Benchmarking currently available seldi-tof ms preprocessing techniques. *Proteomics*, 9(7):1754–1762, Apr 2009.

[8] Xiaochun Li. *PROcess: Ciphergen SELDI-TOF Processing*, 2005. R package version 1.12.0.

[9] Kevin R Coombes, Spiridon Tsavachidis, Jeffrey S Morris, Keith A Baggerly, Mien-Chie Hung, and Henry M Kuerer. Improved peak detection and quantification of mass spectrometry data acquired from surface-enhanced laser desorption and ionization by denoising spectra with the undecimated discrete wavelet transform. *Proteomics*, 5(16):4107–4117, Nov 2005.

[10] Jeffrey S Morris, Kevin R Coombes, John Koomen, Keith A Baggerly, and Ryuji Kobayashi. Feature extraction and quantification for mass spectrometry in biomedical applications using the mean spectrum. *Bioinformatics*, 21(9):1764–1775, May 2005.

[11] Yudi Pawitan, Stefan Michiels, Serge Koscielny, Arief Gusnanto, and Alexander Ploner. False discovery rate, sensitivity and sample size for microarray studies. *Bioinformatics*, 21(13):3017–3024, Jul 2005.

[12] Stan Pounds and Cheng Cheng. Sample size determination for the false discovery rate. *Bioinformatics*, 21(23):4263–4271, Dec 2005.

[13] Sin-Ho Jung. Sample size for fdr-control in microarray data analysis. *Bioinformatics*, 21(14):3097–3104, Jul 2005.

[14] Peng Liu and J. T Gene Hwang. Quick calculation for sample size while controlling false discovery rate with application to microarray analysis. *Bioinformatics*, 23(6):739–746, Mar 2007.

[15] Tommy S Jrstad, Herman Midelfart, and Atle M Bones. A mixture model approach to sample size estimation in two-sample comparative microarray experiments. *BMC Bioinformatics*, 9:117, 2008.

[16] David A Cairns, Jennifer H Barrett, Lucinda J Billingham, Anthea J Stanley, George Xinarianos, John K Field, Phillip J Johnson, Peter J Selby, and Rosamonde E Banks. Sample size determination in clinical proteomic profiling experiments using mass spectrometry for class comparison. *Proteomics*, 9(1):74–86, Jan 2009.

[17] Mei-Ling Ting Lee and G. A. Whitmore. Power and sample size for dna microarray studies. *Stat Med*, 21(23):3543–3570, Dec 2002.

[18] Martijn Dijkstra, Han Roelofsen, Roel J Vonk, and Ritsert C Jansen. Peak quantification in surface-enhanced laser desorption/ionization by using mixture models. *Proteomics*, 6(19):5106–5116, Oct 2006.

Chapitre 5

Discussion et perspectives

La puissance statistique, notion souvent accessoire dans les études biomédicales, est un élément de décision majeur lorsqu'il s'agit d'expériences de détection de nouveaux biomarqueurs. Elle traduit rien de moins que la capacité de détection de ces biomarqueurs, mais reste pourtant sous-estimée dans la littérature concernant la spectrométrie de masse voire plus généralement la protéomique. Ce travail de thèse avait pour objet l'étude de cette puissance, dont l'étude des déterminants est une approche directement utile. En effet, les conclusions portées permettent de mieux appréhender les critères permettant une bonne étude, au sens de la probabilité de détection de biomarqueurs. On peut ainsi mettre en évidence les défauts d'études passées et fixer des recommandations pour les études à venir.

Un des messages principaux de ce travail est la nécessité d'augmenter les effectifs des études protéomiques de détection de biomarqueurs. Les schémas expérimentaux de la majorité des études actuelles ne permettent tout simplement pas la détection de biomarqueurs, quelles que soient les qualités de la technique : même la donnée des concentrations parfaites ne permet alors pas la mise en évidence de nouveaux biomarqueurs avec une puissance suffisante. Ce message ne présente pas de nouveauté conceptuelle, mais reste peu pris en compte dans la planification et la réalisation des expériences actuelles. La donnée quantitative de la puissance en fonction du nombre de sujets inclus, spécifiquement pour la spectrométrie de masse, constitue alors un message pédagogique d'importance.

Il faut préciser que l'exigence de taille d'étude soulevée ici est à mettre en rapport avec une hypothèse majeure de ce travail : l'effet différentiel attendu pour les nouveaux biomarqueurs est faible, inférieur à l'unité. Cairns et al. [59] ont pourtant recommandé des études comportant moins d'une dizaine de spectres par groupe pour identifier des marqueurs. Ce travail, conceptuellement intéressant et ayant le mérite de fournir des formules permettant le calcul du nombre d'individus à inclure, repose toutefois implicitement sur des effets différentiels improbables, de l'ordre de la dizaine. Il est ainsi important de tempérer toute vision optimiste des tailles d'expérience en prenant en compte les caractéristiques attendues des biomarqueurs.

Au delà de la mauvaise calibration de la majorité des schémas expérimentaux actuels, l'étude réalisée fait ressortir différents déterminants de la puissance. L'erreur de mesure, ou variabilité instrumentale, en est un majeur. Cela permet d'espérer améliorer la puissance des études en limitant cette erreur. Toutefois, celle-ci n'est pas seulement due à la qualité de la mesure, mais aussi aux techniques de traitement du signal employées. On a ainsi vu l'impact de ces techniques

au travers de deux exemples particuliers : les imperfections dans la détection des pics eux-mêmes, et les différentes estimations de l'intensité associée à un pic. L'amélioration de ces deux aspects permet aussi d'espérer améliorer la puissance. En outre, des interactions entre ces déterminants existent. La synergie péjorative mise en évidence dans le chapitre 3 en est un exemple. La variabilité expérimentale et le contrôle du FDR concourent alors à une puissance très affaiblie.

L'importance de la variabilité instrumentale suggère son analyse, avec l'utilisation de réplicats techniques pour chaque sujet de l'étude. L'idée n'est pas nouvelle mais n'a pas été développée dans le cadre de ce travail. La simulation de réplicats est aisément réalisable, mais l'analyse des données obtenues pose de nouvelles questions. Elle suppose ainsi l'utilisation de modèles mixtes ou d'erreur de mesure, selon le point de vue adopté sur les données. Ces modèles devront être utilisés pour étudier l'effet sur la puissance de ces répétitions. Ce travail doit aussi permettre d'améliorer l'estimation de la variabilité instrumentale.

Bien que les résultats présentés ici doivent servir à calibrer des expériences réelles, ils restent basés sur des simulations. Leur validation sur un jeu de données réel est donc un objectif important, dont la réalisation est prévue à court terme, sur des échantillons artificiellement modifiés pour y faire apparaître des biomarqueurs désignés. Les méthodes d'analyse développées ici pourront être reprises, valorisant le travail réalisé.

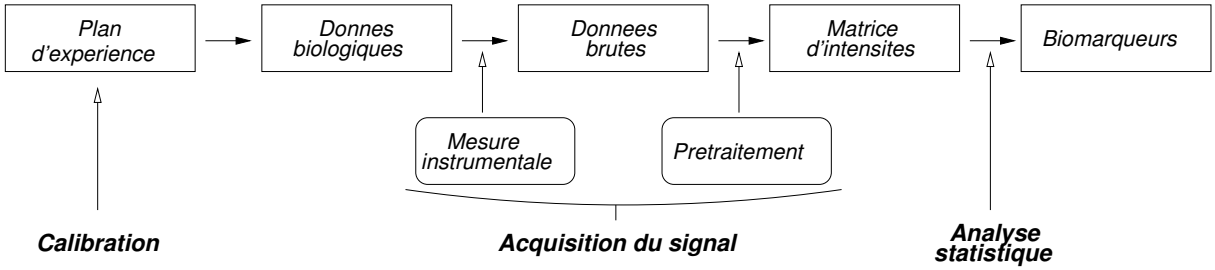

FIG. 5.1 – Parallèle entre chaîne de prétraitement et chaîne de perte de puissance : de la conception de l'étude aux biomarqueurs, les maillons d'intervention possible.

Au final, les déterminants de la puissance en spectrométrie de masse s'intègrent dans une chaîne d'analyse, parallèle à une chaîne de pertes de puissance, présentées en figure 5.1. Trois niveaux d'intervention possible pour améliorer la puissance ont été investigués dans cette thèse : i) une meilleure calibration expérimentale, ii) la minimisation de la variabilité instrumentale, iii) l'optimisation des algorithmes de prétraitement. Ces trois points ne représentent pas une nouveauté mais rappellent trois pistes d'amélioration de la spectrométrie de masse, afin d'en exploiter tout le potentiel. Le choix de la stratégie d'analyse statistique impose lui aussi une réflexion, reflètant une quatrième piste d'amélioration, dont les résultats s'appliqueront plus largement aux technologies *omiques*.

Index

Bibliographie

[1] T. R. Golub, D. K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J. P. Mesirov, H. Coller, M. L. Loh, J. R. Downing, M. A. Caligiuri, C. D. Bloomfield, and E. S. Lander. Molecular classification of cancer : class discovery and class prediction by gene expression monitoring. *Science*, 286(5439) :531–537, Oct 1999.

[2] Emanuel F Petricoin, Ali M Ardekani, Ben A Hitt, Peter J Levine, Vincent A Fusaro, Seth M Steinberg, Gordon B Mills, Charles Simone, David A Fishman, Elise C Kohn, and Lance A Liotta. Use of proteomic patterns in serum to identify ovarian cancer. *Lancet*, 359(9306) :572–577, Feb 2002.

[3] Jinong Li, Zhen Zhang, Jason Rosenzweig, Young Y Wang, and Daniel W Chan. Proteomics and bioinformatics approaches for identification of serum biomarkers to detect breast cancer. *Clin Chem*, 48(8) :1296–1304, Aug 2002.

[4] Claudio Belluco, Emanuel F Petricoin, Enzo Mammano, Francesco Facchiano, Sally Ross-Rucker, Donato Nitti, Cosimo Di Maggio, Chenwei Liu, Mario Lise, Lance A Liotta, and Gordon Whiteley. Serum proteomic analysis identifies a highly sensitive and specific discriminatory pattern in stage 1 breast cancer. *Ann Surg Oncol*, 14(9) :2470–2476, Sep 2007.

[5] Marie-Christine Gast, Jan Schellens, and Jos Beijnen. Clinical proteomics in breast cancer : a review. *Breast Cancer Res Treat*, 116 :17–29, Dec 2008.

[6] Kevin R Coombes, Herbert A Fritsche, Charlotte Clarke, Jeng-Neng Chen, Keith A Baggerly, Jeffrey S Morris, Lian-Chun Xiao, Mien-Chie Hung, and Henry M Kuerer. Quality control and peak finding for proteomics data collected from nipple aspirate fluid by surface-enhanced laser desorption and ionization. *Clin Chem*, 49(10) :1615–1623, Oct 2003.

[7] John M Koomen, Lichen Nancy Shih, Kevin R Coombes, Donghui Li, Lian chun Xiao, Isaiah J Fidler, James L Abbruzzese, and Ryuji Kobayashi. Plasma protein profiling for diagnosis of pancreatic cancer reveals the presence of host response proteins. *Clin Cancer Res*, 11(3) :1110–1118, Feb 2005.

[8] Judith Y M N Engwegen, Helgi H Helgason, Annemieke Cats, Nathan Harris, Johannes M G Bonfrer, Jan H M Schellens, and Jos H Beijnen. Identification of serum proteins discriminating colorectal cancer patients and healthy controls using surface-enhanced laser desorption ionisation-time of flight mass spectrometry. *World J Gastroenterol*, 12(10) :1536–1544, Mar 2006.

[9] Mirre E de Noo, Bart J A Mertens, Aliye Ozalp, Marco R Bladergroen, Martijn P J van der Werff, Cornelis J H van de Velde, Andre M Deelder, and Rob A E M Tollenaar. Detection of colorectal cancer using maldi-tof serum protein profiling. *Eur J Cancer*, 42(8) :1068–1076, May 2006.

[10] Bao-Ling Adam, Yinsheng Qu, John W Davis, Michael D Ward, Mary Ann Clements, Lisa H Cazares, O. John Semmes, Paul F Schellhammer, Yutaka Yasui, Ziding Feng, and George L Wright. Serum protein fingerprinting coupled with a pattern-matching algorithm distinguishes prostate cancer from benign prostate hyperplasia and healthy men. *Cancer Res*, 62(13) :3609–3614, Jul 2002.

[11] Yutaka Yasui, Margaret Pepe, Mary Lou Thompson, Bao-Ling Adam, George L Wright, Yinsheng Qu, John D Potter, Marcy Winget, Mark Thornquist, and Ziding Feng. A data-analytic strategy for protein biomarker discovery : profiling of high-dimensional proteomic data for cancer detection. *Biostatistics*, 4(3) :449–463, Jul 2003.

[12] Yutaka Yasui, Dale McLerran, Bao-Ling Adam, Marcy Winget, Mark Thornquist, and Ziding Feng. An automated peak identification/calibration procedure for high-dimensional protein measures from mass spectrometers. *J Biomed Biotechnol*, 2003(4) :242–248, 2003.

[13] Weida Tong, Qian Xie, Huixiao Hong, Leming Shi, Hong Fang, Roger Perkins, and Emanuel F Petricoin. Using decision forest to classify prostate cancer samples on the basis of seldi-tof ms data : assessing chance correlation and prediction confidence. *Environ Health Perspect*, 112(16) :1622–1627, Nov 2004.

[14] G. Ball, S. Mian, F. Holding, R. O. Allibone, J. Lowe, S. Ali, G. Li, S. McCardle, I. O. Ellis, C. Creaser, and R. C. Rees. An integrated approach utilizing artificial neural networks and seldi mass spectrometry for the classification of human tumours and rapid identification of potential biomarkers. *Bioinformatics*, 18(3) :395–404, Mar 2002.

[15] Fatima W Khwaja, John David Larkin Nolen, Savaas E Mendrinos, Melinda M Lewis, Jeffrey J Olson, Jan Pohl, Erwin G Van Meir, James C Ritchie, and Daniel J Brat. Proteomic analysis of cerebrospinal fluid discriminates malignant and nonmalignant disease of the central nervous system and identifies specific protein markers. *Proteomics*, 6(23) :6277–6287, Dec 2006.

[16] L. C. Whelan, K. A R Power, D. T. McDowell, J. Kennedy, and W. M. Gallagher. Applications of seldi-ms technology in oncology. *J Cell Mol Med*, 12(5A) :1535–1547, 2008.

[17] Hasan H Otu, Handan Can, Dimitrios Spentzos, Robert G Nelson, Robert L Hanson, Helen C Looker, William C Knowler, Manuel Monroy, Towia A Libermann, S. Ananth Karumanchi, and Ravi Thadhani. Prediction of diabetic nephropathy using urine proteomic profiling 10 years prior to development of nephropathy. *Diabetes Care*, 30(3) :638–643, Mar 2007.

[18] Michael L Merchant and Jon B Klein. Proteomics and diabetic nephropathy. *Semin Nephrol*, 27(6) :627–636, Nov 2007.

[19] Annunziata Lapolla, Roberta Seraglia, Laura Molin, Katherine Williams, Chiara Cosma, Rachele Reitano, Annalisa Sechi, Eugenio Ragazzi, and Pietro Traldi. Low molecular weight proteins in urines from healthy subjects as well as diabetic, nephropathic and diabetic-nephropathic patients : a maldi study. *J Mass Spectrom*, 44(3) :419–425, Mar 2009.

[20] Marie-Alice Meuwis, Marianne Fillet, Laurence Lutteri, Raphael Maree, Pierre Geurts, Dominique de Seny, Michel Malaise, Jean-Paul Chapelle, Louis Wehenkel, Jacques Belaiche, Marie-Paule Merville, and Edouard Louis. Proteomics for prediction and characterization of response to infliximab in crohn's disease : a pilot study. *Clin Biochem*, 41(12) :960–967, Aug 2008.

[21] Eleftherios P Diamandis. Mass spectrometry as a diagnostic and a cancer biomarker discovery tool : opportunities and potential limitations. *Mol Cell Proteomics*, 3(4) :367–378, Apr 2004.

[22] Eleftherios P Diamandis and Da-Elene van der Merwe. Plasma protein profiling by mass spectrometry for cancer diagnosis : opportunities and limitations. *Clin Cancer Res*, 11(3) :963–965, Feb 2005.

[23] Glen L Hortin. The maldi-tof mass spectrometric view of the plasma proteome and peptidome. *Clin Chem*, 52(7) :1223–1237, Jul 2006.

[24] Jeffrey S Morris, Kevin R Coombes, John Koomen, Keith A Baggerly, and Ryuji Kobayashi. Feature extraction and quantification for mass spectrometry in biomedical applications using the mean spectrum. *Bioinformatics*, 21(9) :1764–1775, May 2005.

[25] Anestis Antoniadis, Sophie Lambert-Lacroix, Frederique Letue, and Jeremie Bigot. Nonparametric preprocessing methods and inference tools for analyzing time-of-flight mass spectrometry data. *Current Analytical Chemistry*, 3(2) :127–147, April 2007.

[26] Theodore Alexandrov, Jens Decker, Bart Mertens, Andre M Deelder, Rob A E M Tollenaar, Peter Maass, and Herbert Thiele. Biomarker discovery in maldi-tof serum protein profiles using discrete wavelet transformation. *Bioinformatics*, 25(5) :643–649, Mar 2009.

[27] Joshua W K Ho, Maurizio Stefani, Cristobal G dos Remedios, and Michael A Charleston. Differential variability analysis of gene expression and its application to human diseases. *Bioinformatics*, 24(13) :i390–i398, Jul 2008.

[28] Baolin Wu, Tom Abbott, David Fishman, Walter McMurray, Gil Mor, Kathryn Stone, David Ward, Kenneth Williams, and Hongyu Zhao. Comparison of statistical methods for classification of ovarian cancer using mass spectrometry data. *Bioinformatics*, 19(13) :1636–1643, Sep 2003.

[29] R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58(1) :267–288, 1996.

[30] B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani. Least angle regression. *Annals of Statistics*, 32(2) :407–499, 2004.

[31] Y. Benjamini and Y. Hochberg. Controlling the false discovery rate : A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, 57(1) :289–300, 1995.

[32] Y. Benjamini, Y. Hochberg, and Y. Kling. False discovery rate control in multiple hypotheses testing using dependent test statistics. 1997.

[33] Y. Benjamini and D. Yekutieli. The control of the false discovery rate in multiple testing under dependency. *Annals of statistics*, 4 :1165–1188, 2001.

[34] S. Dudoit, J.P. Shaffer, and J.C. Boldrick. Multiple hypothesis testing in microarray experiments. *Statistical Science*, 18(1) :71–103, 2003.

[35] JD Storey. A direct approach to false discovery rates JR Stat. *Journal of the Royal Statistical Society. Series B (Methodological)*, 64 :479, 2002.

[36] John D Storey and Robert Tibshirani. Statistical significance for genomewide studies. *Proc Natl Acad Sci U S A*, 100(16) :9440–9445, Aug 2003.

[37] Hui-Rong Qian and Shuguang Huang. Comparison of false discovery rate methods in identifying genes with differential expression. *Genomics*, 86(4) :495–503, Oct 2005.

[38] Sin-Ho Jung and Woncheol Jang. How accurately can we control the fdr in analyzing microarray data ? *Bioinformatics*, 22(14) :1730–1736, Jul 2006.

[39] Xin Lu and David L Perkins. Re-sampling strategy to improve the estimation of number of null hypotheses in fdr control under strong correlation structures. *BMC Bioinformatics*, 8 :157, 2007.

[40] Kyung In Kim and Mark A van de Wiel. Effects of dependence in high-dimensional multiple testing problems. *BMC Bioinformatics*, 9 :114, 2008.

[41] Yudi Pawitan, Stefan Michiels, Serge Koscielny, Arief Gusnanto, and Alexander Ploner. False discovery rate, sensitivity and sample size for microarray studies. *Bioinformatics*, 21(13) :3017–3024, Jul 2005.

[42] Stan Pounds and Cheng Cheng. Sample size determination for the false discovery rate. *Bioinformatics*, 21(23) :4263–4271, Dec 2005.

[43] Sin-Ho Jung. Sample size for fdr-control in microarray data analysis. *Bioinformatics*, 21(14) :3097–3104, Jul 2005.

[44] Peng Liu and J. T Gene Hwang. Quick calculation for sample size while controlling false discovery rate with application to microarray analysis. *Bioinformatics*, 23(6) :739–746, Mar 2007.

[45] Jonathan Taylor, Robert Tibshirani, and Bradley Efron. The 'miss rate' for the analysis of gene expression data. *Biostatistics*, 6(1) :111–117, Jan 2005.

[46] Andrew W Norris and C. Ronald Kahn. Analysis of gene expression in pathophysiological states : balancing false discovery and false negative rates. *Proc Natl Acad Sci U S A*, 103(3) :649–653, Jan 2006.

[47] Mei-Ling Ting Lee and G. A. Whitmore. Power and sample size for dna microarray studies. *Stat Med*, 21(23) :3543–3570, Dec 2002.

[48] Chen-An Tsai, Sue-Jane Wang, Dung-Tsa Chen, and James J Chen. Sample size for gene expression microarray experiments. *Bioinformatics*, 21(8) :1502–1508, Apr 2005.

[49] Liat Ein-Dor, Itai Kela, Gad Getz, David Givol, and Eytan Domany. Outcome signature genes in breast cancer : is there a unique set ? *Bioinformatics*, 21(2) :171–178, Jan 2005.

[50] Liat Ein-Dor, Or Zuk, and Eytan Domany. Thousands of samples are needed to generate a robust gene list for predicting outcome in cancer. *Proc Natl Acad Sci U S A*, 103(15) :5923–5928, Apr 2006.

[51] Jiang Gui and Hongzhe Li. Penalized cox regression analysis in the high-dimensional and low-sample size settings, with applications to microarray gene expression data. *Bioinformatics*, 21(13) :3001–3008, Jul 2005.

[52] Sin-Ho Jung, Heejung Bang, and Stanley Young. Sample size calculation for multiple testing in microarray data analysis. *Biostatistics*, 6(1) :157–169, Jan 2005.

[53] Tommy S Jrstad, Herman Midelfart, and Atle M Bones. A mixture model approach to sample size estimation in two-sample comparative microarray experiments. *BMC Bioinformatics*, 9 :117, 2008.

[54] Richard D Pearson. A comprehensive re-analysis of the golden spike data : towards a benchmark for differential expression methods. *BMC Bioinformatics*, 9 :164, 2008.

[55] Caroline Truntzer, Delphine Maucort-Boulch, and Pascal Roy. Comparative optimism in models involving both classical clinical and gene expression information. *BMC Bioinformatics*, 9 :434, 2008.

[56] K R Coombes, J M Koomen, K A Baggerly, J S Morris, and R Kobayashi. Understanding the characteristics of mass spectrometry data through the use of simulation. *Cancer Inform*, 1 :41–52, 2005.

[57] Alejandro Cruz-Marcelo, Rudy Guerra, Marina Vannucci, Yiting Li, Ching C Lau, and Tsz-Kwong Man. Comparison of algorithms for pre-processing of seldi-tof mass spectrometry data. *Bioinformatics*, 24(19) :2129–2136, Oct 2008.

[58] Chao Yang, Zengyou He, and Weichuan Yu. Comparison of public peak detection algorithms for maldi mass spectrometry data analysis. *BMC Bioinformatics*, 10 :4, 2009.

[59] David A Cairns, Jennifer H Barrett, Lucinda J Billingham, Anthea J Stanley, George Xinarianos, John K Field, Phillip J Johnson, Peter J Selby, and Rosamonde E Banks. Sample size determination in clinical proteomic profiling experiments using mass spectrometry for class comparison. *Proteomics*, 9(1) :74–86, Jan 2009.