N° d'ordre : 490 IVS

THÈSE

présentée par

Fabien BERNARD

Pour obtenir le grade de Docteur
de l'École Nationale Supérieure des Mines de Saint-Étienne

Spécialité : image, vision, signal

IDADIGE : PROCÉDÉ DE TRAITEMENT DES IMAGES DE GELS D'ÉLECTROPHORÈSE BIDIMENSIONNELLE DIFFÉRENTIELLE DANS LE CONTEXTE DE LA RECHERCHE DE MARQUEURS PROTEIQUES

Soutenue à Saint-Étienne le 26 septembre 2008

Membres du jury

Président :	
ANTONIADIS Anestis	Professeur, Université Joseph Fourier
Rapporteurs :	
JOUBERT-CARON Raymonde	Ingénieur de Recherche, HDR, Université Paris13
PRÉTEUX Françoise	Professeur, TELECOM & Management SudParis
Examinateurs :	
LACROIX Bruno	PhD, Entreprise bioMérieux
BAY Xavier	Maître Assistant, École des Mines de Saint-Étienne
CHOQUET-KASTYLEVSKY Geneviève	MD/PhD, Entreprise bioMérieux
GRUY Frédéric	Maître de recherche, École des Mines de Saint-Étienne
Directeur(s) de thèse :	
PINOLI Jean-Charles	Professeur, École des Mines de Saint-Étienne
Invité :	
LAMBERT Claude	Médecin, Centre Hospitalier Universitaire de Saint-Étienne

◘ Spécialités doctorales : **Responsables :**

SCIENCES ET GENIE DES MATERIAUX	**J. DRIVER** Directeur de recherche – Centre SMS
MECANIQUE ET INGENIERIE	**A. VAUTRIN** Professeur – Centre SMS
GENIE DES PROCEDES	**G. THOMAS** Professeur – Centre SPIN
SCIENCES DE LA TERRE	**B. GUY** Maitre de recherche – Centre SPIN
SCIENCES ET GENIE DE L'ENVIRONNEMENT	**J. BOURGOIS** Professeur – Centre SITE
MATHEMATIQUES APPLIQUEES	**E. TOUBOUL** Ingénieur – Centre G2I
INFORMATIQUE	**O. BOISSIER** Professeur – Centre G2I
IMAGE, VISION, SIGNAL	**JC. PINOLI** Professeur – Centre CIS
GENIE INDUSTRIEL	**P. BURLAT** Professeur – Centre G2I
MICROELECTRONIQUE	**Ph. COLLOT** Professeur – Centre CMP

◘ Enseignants-chercheurs et chercheurs autorisés à diriger des thèses de doctorat (titulaires d'un doctorat d'Etat ou d'une HDR)

AVRIL	Stéphane	MA	Mécanique & Ingénierie	CIS
BATTON-HUBERT	Mireille	MA	Sciences & Génie de l'Environnement	SITE
BENABEN	Patrick	PR 2	Sciences & Génie des Matériaux	SMS
BERNACHE-ASSOLANT	Didier	PR 1	Génie des Procédés	CIS
BIGOT	Jean-Pierre	MR	Génie des Procédés	SPIN
BILAL	Essaïd	DR	Sciences de la Terre	SPIN
BOISSIER	Olivier	PR 2	Informatique	G2I
BOUCHER	Xavier	MA	Génie Industriel	G2I
BOUDAREL	Marie-Reine	MA	Sciences de l'inform. & com.	DF
BOURGOIS	Jacques	PR 1	Sciences & Génie de l'Environnement	SITE
BRODHAG	Christian	MR	Sciences & Génie de l'Environnement	SITE
BURLAT	Patrick	PR 2	Génie industriel	G2I
CARRARO	Laurent	PR 1	Mathématiques Appliquées	G2I
COLLOT	Philippe	PR 1	Microélectronique	CMP
COURNIL	Michel	PR 1	Génie des Procédés	SPIN
DAUZERE-PERES	Stéphane	PR 1	Génie industriel	CMP
DARRIEULAT	Michel	ICM	Sciences & Génie des Matériaux	SMS
DECHOMETS	Roland	PR 1	Sciences & Génie de l'Environnement	SITE
DESRAYAUD	Christophe	MA	Mécanique & Ingénierie	SMS
DELAFOSSE	David	PR 2	Sciences & Génie des Matériaux	SMS
DOLGUI	Alexandre	PR 1	Génie Industriel	G2I
DRAPIER	Sylvain	PR 2	Mécanique & Ingénierie	CIS
DRIVER	Julian	DR	Sciences & Génie des Matériaux	SMS
FOREST	Bernard	PR 1	Sciences & Génie des Matériaux	CIS
FORMISYN	Pascal	PR 1	Sciences & Génie de l'Environnement	SITE
FORTUNIER	Roland	PR 1	Sciences & Génie des Matériaux	CMP
FRACZKIEWICZ	Anna	MR	Sciences & Génie des Matériaux	SMS
GARCIA	Daniel	CR	Génie des Procédés	SPIN
GIRARDOT	Jean-Jacques	MR	Informatique	G2I
GOEURIOT	Dominique	MR	Sciences & Génie des Matériaux	SMS
GOEURIOT	Patrice	MR	Sciences & Génie des Matériaux	SMS
GRAILLOT	Didier	DR	Sciences & Génie de l'Environnement	SITE
GROSSEAU	Philippe	MR	Génie des Procédés	SPIN
GRUY	Frédéric	MR	Génie des Procédés	SPIN
GUILHOT	Bernard	DR	Génie des Procédés	CIS
GUY	Bernard	MR	Sciences de la Terre	SPIN
GUYONNET	René	DR	Génie des Procédés	SPIN
HERRI	Jean-Michel	PR 2	Génie des Procédés	SPIN
KLÖCKER	Helmut	MR	Sciences & Génie des Matériaux	SMS
LAFOREST	Valérie	CR	Sciences & Génie de l'Environnement	SITE
LI	Jean-Michel	EC (CCI MP)	Microélectronique	CMP
LONDICHE	Henry	MR	Sciences & Génie de l'Environnement	SITE
MOLIMARD	Jérôme	MA	Sciences & Génie des Matériaux	SMS
MONTHEILLET	Frank	DR 1 CNRS	Sciences & Génie des Matériaux	SMS
PERIER-CAMBY	Laurent	PR1	Génie des Procédés	SPIN
PIJOLAT	Christophe	PR 1	Génie des Procédés	SPIN
PIJOLAT	Michèle	PR 1	Génie des Procédés	SPIN
PINOLI	Jean-Charles	PR 1	Image, Vision, Signal	CIS
STOLARZ	Jacques	CR	Sciences & Génie des Matériaux	SMS
SZAFNICKI	Konrad	CR	Sciences de la Terre	SITE
THOMAS	Gérard	PR 1	Génie des Procédés	SPIN
VALDIVIESO	François	MA	Sciences & Génie des Matériaux	SMS
VAUTRIN	Alain	PR 1	Mécanique & Ingénierie	SMS
VIRICELLE	Jean-Paul	MR	Génie des procédés	SPIN
WOLSKI	Krzysztof	CR	Sciences & Génie des Matériaux	SMS
XIE	Xiaolan	PR 1	Génie industriel	CIS

Glossaire :

PR 1	Professeur 1ère catégorie
PR 2	Professeur 2ème catégorie
MA(MDC)	Maître assistant
DR (DR1)	Directeur de recherche
Ing.	Ingénieur
MR(DR2)	Maître de recherche
CR	Chargé de recherche
EC	Enseignant-chercheur
ICM	Ingénieur en chef des mines

Centres :

SMS	Sciences des Matériaux et des Structures
SPIN	Sciences des Processus Industriels et Naturels
SITE	Sciences Information et Technologies pour l'Environnement
G2I	Génie Industriel et Informatique
CMP	Centre de Microélectronique de Provence
CIS	Centre Ingénierie et Santé

Remerciements

Je voudrais avant tout remercier les différentes personnes qui par leur soutien, leur expérience, leurs connaissances ainsi que par leurs qualités humaines et leur patience ont rendu ce projet possible. Ce travail est le résultat d'une collaboration avec des personnes issues de structures différentes, mais il est surtout le fruit de beaucoup d'amitiés.

Pour l'École des Mines de Saint-Étienne, je tiens à remercier mon directeur de thèse, Jean-Charles PINOLI moteur de la collaboration avec bioMérieux, ainsi que Frédéric GRUY et Xavier BAY qui par leurs analyses ont régulièrement su faire progresser la réflexion.

Je remercie de la même manière Claude LAMBERT, de la faculté de médecine de Saint-Étienne, qui, avec grande gentillesse, s'est toujours montré disponible et efficace.

Je remercie les différents partenaires du projet de recherche NODDICCAP et en particulier Raymonde et Michel CARON du Laboratoire de Biochimie des Protéines et Protéomique de l'université Paris 13.

Ayant principalement travaillé au sein de l'entreprise bioMérieux, j'y ai côtoyé de nombreuses personnes qui m'ont accompagné et beaucoup apporté, durant toute la durée de mon travail. Je remercie en particulier mon co-directeur de thèse, Bruno LACROIX qui a réussi à recadrer les travaux quand cela a été nécessaire et qui a été présent tout au long du projet, me fournissant notamment des pistes de travail et du matériel bibliographique.

Bien sûr, je veux remercier toute l'équipe de protéomique dont la responsable, Geneviève CHOQUET-KASTYLEVSKY, à l'initiative de la collaboration et du projet de recherche, m'a communiqué sa motivation à travers son énergie, sa curiosité et sa spontanéité. Je voudrais dire un grand merci à Jean-Philippe CHARRIER et à Florence POIRIER avec qui j'ai pu avoir de longues réflexions et qui m'ont apporté leur expérience, leurs connaissances et leur attention. Florence est également beaucoup intervenu, avec beaucoup de minutie et de rigueur, à certaines étapes de mes travaux et a contribué à la qualité des données exploitées. Je remercie Sylvie PONS, Corinne BEAULIEU, Aymeric MORLA qui ont été bien plus que des amis. Outre leurs connaissances et leurs réflexions, ils m'ont apporté leurs encouragements et ont participé à l'excellente ambiance qui régnait dans l'équipe. Sylvie, avec qui j'ai passé beaucoup de temps en laboratoire m'a également permis de toucher du doigt l'aspect purement expérimental de l'électrophorèse bidimensionnelle. Son travail a permis de constituer des bases de données de qualité. Je remercie aussi les différents stagiaires qui se sont succédés, et en particulier Hader Haidous pour sa gentillesse et son implication à certaines étapes de mon travail.

Je tiens ensuite à remercier les nombreuses personnes du département IT (Technologie de l'Information) de bioMérieux, département dirigé par Jean-François GORSE, toujours accessible et attentif.

Dans ce département, je remercie tout spécialement l'ensemble du service de biomathématiques qui m'a offert un cadre de travail idéal. Françoise GUERILLOT-MAIRE, chef de ce service, s'est largement impliquée dans le projet et m'a ouvert sur les méthodes de travail en entreprise. Je remercie également, et tout particulièrement, Audrey LARUE, Maud ARSAC, Christian RAGEADE et Malick PAYE avec qui j'ai beaucoup échangé et qui m'ont apporté leur expertise et leurs points de vue toujours pertinents de biomathématiciens et de biostatisticiens.

Je remercie également les personnes de l'ingénierie de la connaissance, autre service du département IT de bioMérieux, qui, de près ou de loin, ont joué un rôle dans ce projet. Parmi eux, j'adresse un remerciement particulier à Frank FOS-CHIATTI qui m'a accompagné durant une partie de mon travail et avec qui j'ai pu avoir de nombreuses interactions.

Je remercie enfin toutes les personnes qui ont compté, pour leur soutien, leur amitié et leur humour. Pêle-mêle, Amandine BAKRI, Nadège DEBILLY, Nadine FALCHERO, Catherine ARTHAUD, Édith MICLET, Alexandra DUMITRU, Lise BARBIER, Imen CANOVA, Stéphanie LABICH, Pierre-Jean COTE-PATA, René VACHON, Didier POIRAULT, Bertrand BONNAUD, Julien HENRY,…

Table des matières

Table des illustrations

I. Lexique

- **Électrophorèse**. Déplacement, sous l'effet d'un champ électrique, de granules, de particules chargées, en solution ou en émulsion. (Cette technique a de nombreuses applications en chimie, biologie, médecine et dans l'industrie.)
- **Gel d'électrophorèse.** Le gel d'électrophorèse est un des supports de l'électrophorèse, sur lequel les particules chargées se déplacent. Dans le cadre de l'étude des protéines les gels utilisés sont des gels de polyacrylamide qui peuvent être de deux types : SDS-PAGE (Sodium Dodecyl Sulfate PolyAcrylamide Gel Electrophoresis) ou Tris-Tricine. Les gels SDS-PAGE sont utilisés pour faire migrer des protéines tandis que les gels Tris-Tricine permettent de visualiser des protéines de petite taille dont le nombre d'acides aminés est inférieur à 150 : les peptides.
- **Électrophorèse bidimensionnelle**. Technique qui consiste à séparer les protéines : dans une première dimension par focalisation isoélectrique (IEF) selon leur point isoélectrique (pI) et dans une deuxième direction, perpendiculaire à la première, selon leur poids moléculaire.
- **Électrophorèse bidimensionnelle différentielle (DIGE)**. Séparation bidimensionnelle de populations protéiques marquées par différents fluorophores, par co-migration au sein d'un même gel.
- **Lignée Cellulaire** : Une lignée cellulaire est une population homogène de cellules, stables après des mitoses successives, et ayant en théorie une capacité illimitée de division. Il s'agit en général de cellules cancéreuses prélevées chez un patient, transformées artificiellement par un oncogène ou encore mutées pour des gènes impliqués dans la régulation du cycle cellulaire. Elles sont d'un usage quotidien dans les laboratoires de recherches de biologie.
- **Lignée Caco-2**. Lignée cellulaire issue du cancer colorectal.
- **Protéine**. Molécule constituée d'une longue chaîne d'acides aminés arrangés suivant une séquence donnée. Les protéines sont essentielles à la structure et au fonctionnement des tissus et des organes. Chaque protéine a une ou plusieurs fonctions qui lui sont propres.
- **Protéome.** Le protéome est l'ensemble des protéines produites par un génome dans des conditions données, à un moment donné. Ce terme d'origine anglo-saxonne a été inventé en 1994 par un chercheur de l'Université Macquarie de Sidney. Le protéome provient de la traduction du génome en protéines. Cette traduction est dépendante des conditions de vie des cellules. La taille et la complexité du protéome sont plus importantes que celle du génome car un gène peut coder pour plusieurs protéines. Ceci est dû à des modifications de la maturation des ARNm (molécules intermédiaires de la traduction, entre le génome et le protéome), mais aussi à des modifications post-traductionnelles des protéines comme les phosphorylations et les glycosylations. Le protéome est de nature dynamique : À la différence du génome qui reste constant (si on ne tient pas compte des mutations) dans les cellules d'un organisme, le protéome varie suivant le type cellulaire, l'activité cellulaire ou le micro-environnement entourant ces cellules.
- **Protéomique.** analyse systématique des profils de protéines dans des extraits protéiques donnés. La génomique permet d'établir les relations entre l'activité d'un gène et une maladie. Mais beaucoup de processus de maladies se manifestent au niveau des protéines. D'où l'intérêt de la protéomique.

- **Fluorophore.** Molécules pouvant être excitées à une certaine longueur d'onde lumineuse en réémettant de la lumière à une autre longueur d'onde également spécifique. Les principales caractéristiques d'un fluorophore sont les suivantes :
- *Longueurs d'onde :* celles qui correspondent aux pics des spectres d'excitation et d'émission,
- *Coefficient d'extinction (ou absorption molaire)* : il relie la quantité de lumière absorbée, pour une longueur d'onde donnée, à la concentration du fluorophore en solution (M-1 cm-1)
- *Rendement quantique :* efficacité relative de la fluorescence comparée aux autres voies de désexcitation (c'est le nombre de photons émis divisé par le nombre de photons absorbés)
- *Durée de vie à l'état excité :* c'est la durée moyenne pendant laquelle la molécule reste à l'état excité avant de retourner à son état basal (psec).
- **Photo blanchiment (photobleaching) :** lorsque la molécule est à l'état excité, il existe une certaine probabilité pour qu'elle participe à des réactions chimiques (on parle alors de réactions photochimiques), en particulier avec l'oxygène sous forme de radicaux libres. Le fluorophore perd alors ses propriétés de fluorescence. Autrement dit, quand on excite une solution de molécules fluorescentes, une certaine proportion d'entre elles est détruite à chaque instant et par conséquent l'intensité de fluorescence décroît au cours du temps. Ce phénomène peut être gênant, notamment en microscopie à fluorescence, mais il peut également être mis à profit pour mesurer la mobilité moléculaire par la méthode de redistribution de fluorescence après photo blanchiment.
- **IPG strip.** De l'anglais « Immobilized pH gradient strip », gel utilisé pour la première dimension de migration lors de l'électrophorèse bidimensionnelle.
- **Hétéroscédasticité :** En statistique, lorsque l'on constate que la variance des termes d'erreur, conditionnellement à l'une des variables explicatives, a tendance à adopter un comportement systématique, on parle d'hétéroscédastivité.
- **Test ELISA :** acronyme d'un examen de laboratoire appelé en anglais : Enzyme-linked immunosorbent assay. C'est une technique immuno-enzymatique utilisée en immunologie pour détecter la présence d'une molécule dans un échantillon à l'aide d'anti-corps spécifiques.

II. Introduction et Contexte

En 2002, l'entreprise bioMérieux et l'École Nationale Supérieure des Mines de Saint Etienne se sont rencontrées autour d'une problématique d'intérêt commun : le traitement des images de gels d'électrophorèse bidimensionnelle. Ceci s'est traduit par la réalisation d'un stage de DEA de 6 mois au sein de bioMérieux sur cette thématique. Ce stage a permis de mettre en évidence la perfectibilité du traitement des images réalisé par les logiciels alors disponibles. Un sujet de thèse a ainsi pu être défini en élargissant le sujet de DEA à la maîtrise du processus de traitement de l'ensemble des données issues de la protéomique, depuis leur acquisition jusqu'à leur interprétation. Les choix stratégiques du laboratoire de protéomique de bioMérieux ont conduit par la suite à focaliser la thèse sur la mise en place et l'optimisation d'un processus d'exploitation d'une technique nouvelle d'exploration du protéome : l'électrophorèse bidimensionnelle différentielle.

Le sujet de cette thèse s'inscrit dans le cadre du projet NODDICCAP initié par l'entreprise bioMérieux et visant le développement de Nouveaux Outils pour le Dépistage, le DIagnostic, l'évaluation du pronostic et le suivi du Cancer Colorectal par une Approche Protéomique. Ce projet est d'un grand intérêt car les cancers colorectaux représentent un problème majeur en cancérologie. En effet, seul un diagnostic de ces tumeurs à des stages précoces offre l'espoir d'un traitement curatif efficace.

Pour répondre à cet enjeu majeur de santé publique, il a été nécessaire de mettre en œuvre une approche faisant appel aux dernières avancées technologiques en protéomique, afin d'identifier de nouveaux marqueurs tumoraux discriminants et spécifiques du cancer colorectal. Une fois identifiés, il sera possible de développer, pour chacun d'eux, des tests de dosage qui pourront être utilisés pour le dépistage et le diagnostic précoce, ainsi que pour établir le pronostic et définir une stratégie de surveillance du cancer.

Le premier objectif du projet NODDICCAP consistait donc en l'identification de marqueurs tumoraux discriminants et spécifiques du cancer colorectal. Un second objectif est maintenant l'évaluation de différents outils pour la détection de ces biomarqueurs au niveau sérique grâce à des immuno-essais de type ELISA (Enzyme-Linked Immunosorbent Assay). Le troisième objectif sera de déterminer la sensibilité, la spécificité et la valeur prédictive des marqueurs identifiés pour chacune des applications cliniques.

Le projet NODDICCAP a été organisé autour d'une équipe pluridisciplinaire associant des cliniciens, des épidémiologistes, des biologistes ainsi que des bioinformaticiens pour la création de bases de données, la gestion du système informatique, pour l'analyse des images électrophorétiques et des signaux de spectrométrie de masse, ainsi que pour le traitement des données. L'apport de ces compétences diverses a permis de couvrir l'ensemble de la problématique liée à la recherche de marqueurs tumoraux.

Pour mener à bien ce projet, bioMérieux s'est doté d'une plate-forme technologique performante avec, notamment, un spectromètre de masse permettant de réaliser des analyses de masse en tandem, avec fragmentation peptidique et a utilisé, par

des voies innovantes, des techniques récentes telles que l'électrophorèse bidimensionnelle différentielle (DIGE).

III. Problématique

L'électrophorèse bidimensionnelle consiste à séparer les protéines :

- tout d'abord par focalisation isoélectrique selon leur point Isoélectrique, ce qui constitue la première dimension,
- puis dans une deuxième dimension, perpendiculaire à la première, selon leur poids moléculaire.

La préparation des gels d'électrophorèse bidimensionnelle suit un protocole très précis. La qualité des images obtenues relève non seulement du respect de ce protocole, des matières premières utilisées, mais également de l'étape importante qu'est la coloration (marquage en fluorescence pour la DIGE). Cette étape de coloration permet de rendre visibles les taches protéiques soit directement pour l'œil humain soit par l'intermédiaire d'un appareil tel qu'un scanner à fluorescence. La qualité de l'image en terme de définition spatiale et de niveaux de gris varie en fonction des technologies employées, du matériel utilisé et des conditions d'acquisition.

La caractérisation des taches (on parle également de « spots ») protéiques est une étape essentielle de l'analyse protéomique. Elle permet de passer de la simple description des variations observées sur les cartes protéiques (des taches apparaissent, disparaissent ou varient d'intensité) à une véritable interprétation : rattacher une tache protéique à une protéine déjà répertoriée dans les banques de données c'est lui attribuer une (ou des) fonction(s), l'associer à une famille de protéines, à un (ou des) lieu(x) d'expression dans la cellule ou le tissu, l'inclure dans une chaîne réactionnelle, etc. Ce travail d'identification est long et fastidieux, car il nécessite, au préalable, d'avoir repéré correctement toutes les taches protéiques sur les images de gel. Les interprétations nécessitent également de comparer des gels entre eux afin de mener des études différentielles par classe. C'est à ce niveau de l'analyse que se situe l'objet de l'étude présentée ici. Les difficultés proviennent de la qualité très variable des gels en terme de bruit, de dérive d'éclairement ou encore de déformation.

L'exploitation de la technologie DIGE, qui constitue le thème principal de cette thèse, couvre le domaine du traitement d'images de part la nature des données étudiées ainsi que le domaine plus général du traitement de l'information (plans d'expérience, normalisation des données, statistiques, etc.…) au cours de la recherche en biologie.

IV. État des lieux de la technologie DIGE

IV.1 Électrophorèse bidimensionnelle

IV.1.1 Introduction

La technologie DIGE est une évolution de la technique bien connue de l'électrophorèse bidimensionnelle. Cette évolution, qui permet le multiplexage de plusieurs populations protéiques au sein d'un même gel, est simplement basée sur l'utilisation d'un marquage en fluorescence. À l'exception de ce nouveau type de marquage, une expérience DIGE repose sur les mêmes principes qu'une expérience d'électrophorèse bidimensionnelle classique. Afin de mieux cerner les caractéristiques et les problématiques dont la DIGE a hérité, il est donc nécessaire, dans un premier temps, de s'intéresser aux principes de l'électrophorèse bidimensionnelle classique.

IV.1.2 Électrophorèse bidimensionnelle et protéomique

L'électrophorèse bidimensionnelle est l'un des outils fondamentaux de la protéomique, un domaine scientifique récent qui complète et enrichit la génomique. En effet, les récents développements dans le domaine de la génomique ont fourni une grande quantité d'informations faisant le lien entre les activités des gènes et certaines pathologies et en permettant d'obtenir les séquences théoriques des protéines, accessibles dans des bases de données publiques depuis le séquençage complet du génome humain. Cependant, comme le soulignent M.R. Wilkins et al, pionniers de la protéomique [59], l'information sur la séquence des gènes ne permet pas de connaître précisément l'abondance d'une protéine ni sa structure finale ou son activité. En effet, pour un génome de mammifères, un seul gène peut coder en moyenne jusqu'à 6 familles de protéines ce qui rend le protéome beaucoup plus complexe que le génome. Par ailleurs, les progrès de la technique de spectrométrie de masse permettent aujourd'hui l'analyse de grosses molécules comme les protéines.

Les protéines sont directement impliquées dans les processus biochimiques normaux et pathologiques. C'est pourquoi, l'étude des protéines présentes dans les tissus ou les cellules pathologiques, permet une meilleure compréhension de la maladie. Il s'agit là du principe de base de l'analyse protéomique dont les résultats, tant pour la biologie que la clinique, sont potentiellement immenses [60].

La première étape d'une étude protéomique est la collecte des échantillons d'origine clinique, végétale ou de culture. Après l'extraction des protéines à partir de ces échantillons, les protéines sont prétraitées : la solubilisation, la dénaturation et la réduction permettent de supprimer toute interaction potentielle entre les protéines et d'éliminer tout le contenu non protéique. L'étape suivante consiste à séparer le maximum de protéines présentes afin de les identifier et de les quantifier.

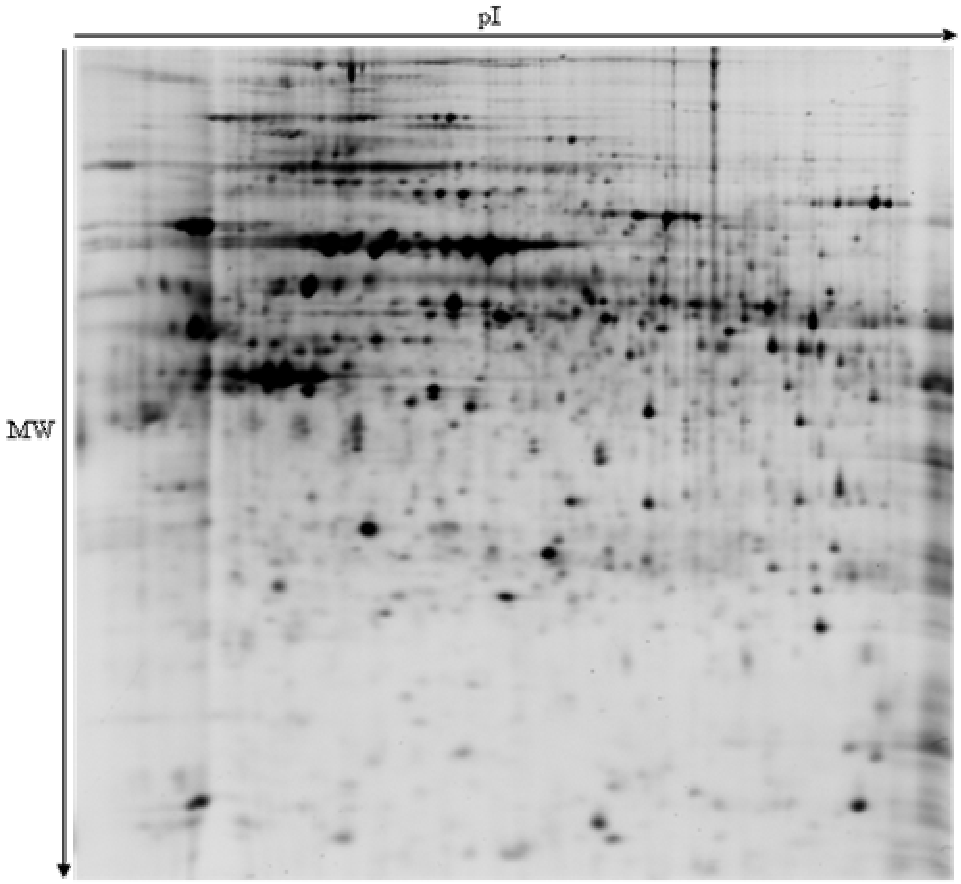


Figure 1:Image typique d'un gel obtenue par électrophorèse bidimensionnelle. Les protéines se répartissent à la surface du gel selon deux dimensions : Leur point isoélectrique (pI) et leur masse moléculaire (MW).

L'électrophorèse bidimensionnelle sur les gels de polyacrylamide (voir Figure 1) est, à l'heure actuelle, la seule méthode permettant la séparation et la quantification simultanée de plusieurs milliers de protéines (jusqu'à 10000 protéines [29]). Elle reste, depuis plus de 20 ans, la méthode la plus résolutive pour la séparation des protéines, malgré l'émergence de nombreuses techniques concurrentes [45]. Bien que de récentes avancées en matière de spectrométrie de masse, en particulier celles basées sur l'utilisation de marqueurs isotopiques stables (ICAT : Isotope Coded Affinity Tags), soient prometteuses, l'électrophorèse bidimensionnelle reste une méthode de choix, beaucoup moins onéreuse, pour l'étude de l'expression différentielle des protéines entre différents échantillons.

IV.1.3 Principes de l'électrophorèse bidimensionnelle

L'exploration du protéome nécessite de pouvoir extraire le maximum de polypeptides d'un échantillon donné. Comme l'illustre le schéma de la Figure 2, l'électrophorèse bidimensionnelle répond à cette nécessité grâce à une double séparation correspondant à deux caractéristiques physico-chimiques intrinsèques de chaque protéine : son point isoélectrique et sa masse moléculaire. Ces caractéristiques sont exploitées grâce à la migration des protéines, induite par un champ électrique, sur un gel de polyacrylamide.

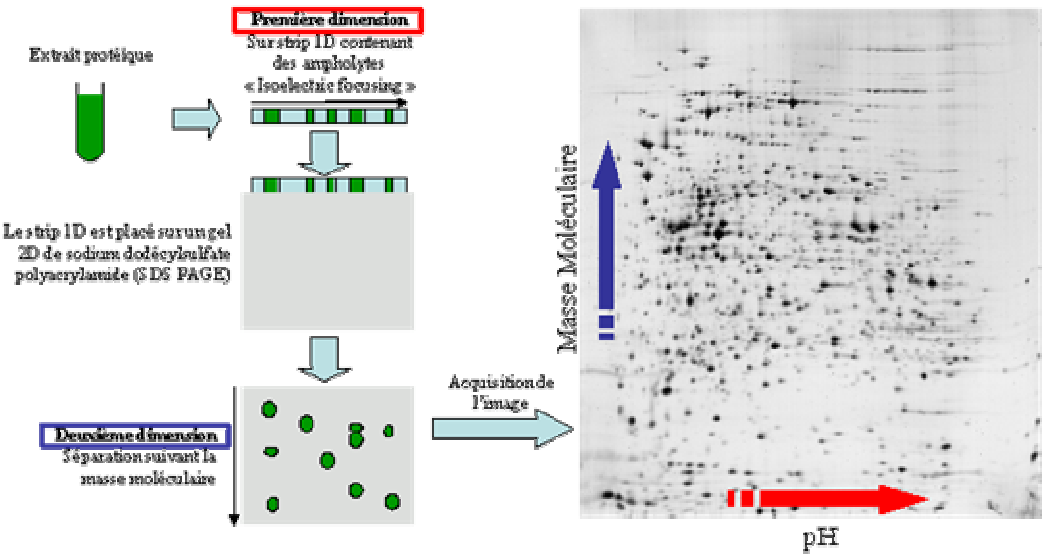


Figure 2: Schéma du principe de l'électrophorèse bidimensionnelle classique : la migration achevée dans la première dimension, on effectue une migration perpendiculaire à la première en utilisant une deuxième technique (deuxième dimension). Les molécules, séparées selon deux critères, se répartissent donc dans un système a deux coordonnées ce qui permet une grande résolution : si beaucoup de molécules ont un poids moléculaire ou un pH proche, les molécules ayant les deux paramètres en commun sont rares. Avec cette variante, il est possible de travailler sur des extraits cellulaires complets.

Les paragraphes suivants détaillent les trois principales étapes de l'électrophorèse bidimensionnelle :

- l'électrofocalisation,
- l'électrophorèse en gel de polyacrylamide en présence de sodium dodecylsulfate (SDS-PAGE)
- et la coloration des protéines pour leur visualisation.

IV.1.3.1 Première dimension : l'électrofocalisation (IEF)

La première dimension de l'électrophorèse bidimensionnelle permet de séparer les protéines suivant leur point isoélectrique (pI). C'est l'étape d'électrofocalisation, réalisée au laboratoire de protéomique de bioMérieux, grâce au « Protean IEF Cell », un instrument dédié, développé par le constructeur Bio-Rad (voir Figure 3).

Les protéines sont des molécules amphotères, elles peuvent être chargées positivement, négativement ou ne pas avoir de charge selon le pH de la solution dans laquelle elles se trouvent. La charge nette d'une protéine est la somme des charges négatives et positives de ses extrémités et des chaînes latérales des acides aminés qui la composent. Le point isoélectrique correspond au pH où la charge nette de la protéine est égale à 0. Les protéines sont chargées positivement à pH inférieur à leur pI et négativement à pH supérieur.

Figure 3: Photo du matériel utilisé au laboratoire de protéomique de bioMérieux pour l'électrofocalisation (Bio-Rad PROTEAN™ IEF Cell)

En solution, les protéines sont chargées et lorsqu'on les soumet à un gradient de pH, sous l'influence d'un champ électrique :

- les protéines chargées positivement migrent, à travers le gradient de pH, vers la cathode en perdant, au fur et à mesure, leurs charges positives pour arriver à une charge nette nulle lorsqu'elles atteignent leur pI.
- les protéines chargées négativement migrent, à travers le gradient de pH, vers l'anode en perdant, au fur et à mesure, leurs charges négatives pour arriver à une charge nette nulle lorsqu'elles atteignent leur pI.

Si elles diffusaient au-delà de ce point isoélectrique, elles seraient à nouveau chargées et elles reviendraient à cette valeur.

IV.1.3.2 Deuxième dimension : le « SDS-PAGE »

Dans la seconde dimension, orthogonale à la première, les protéines sont séparées selon leur masse relative (M_r). Le matériel utilisé au laboratoire de protéomique de bioMérieux est le « Protean Plus Dodeca » développé par le constructeur Bio-Rad (voir Figure 4).

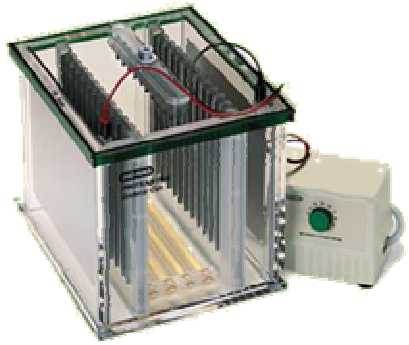


Figure 4: Photo du matériel utilisé au laboratoire de protéomique de bioMérieux pour la deuxième dimension de l'électrophorèse (Bio-Rad PROTEAN Plus™ Dodeca™)

Lors d'une électrophorèse en gel de polyacrylamide (PAGE), en présence de sodium dodecylsulfate (SDS), la migration n'est pas déterminée par les charges électriques des protéines mais par leur poids moléculaire. Le SDS est un agent anionique qui se fixe sur les protéines, masquant la charge propre des protéines, en formant un complexe anionique ayant une charge nette négative. Comme le montre la Figure 5, le gel de polyacrylamide sert de tamis moléculaire.

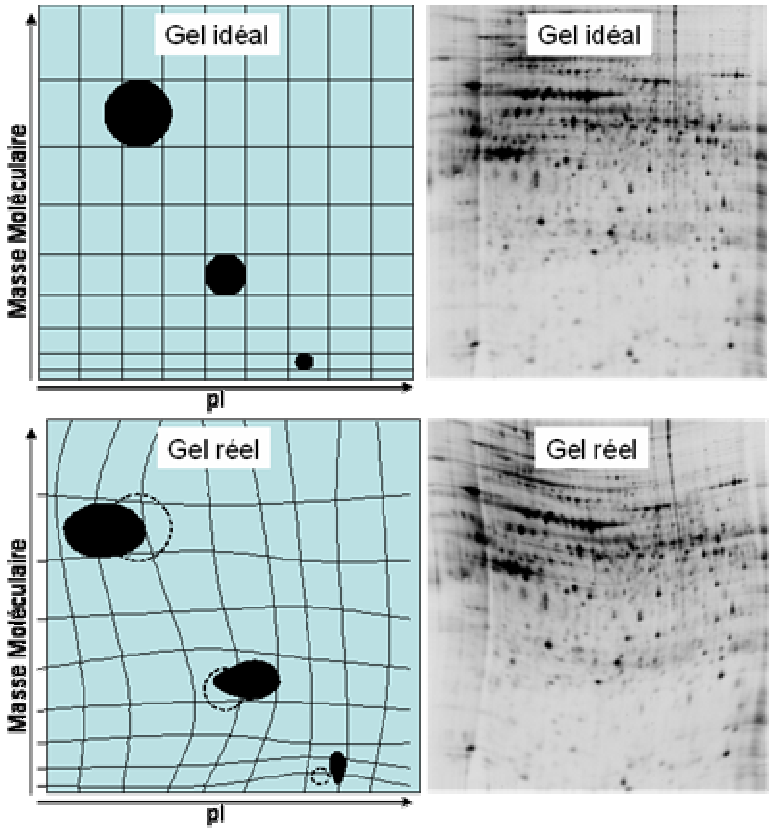


Figure 5 : Schémas du tamis moléculaire d'un gel d'électrophorèse 2D et visualisation d'un gel dans le cas idéal et dans le cas réel. Les protéines de forte masse moléculaire sont arrêtées par les premières mailles tandis que les plus petites protéines poursuivent leur migration vers le bas. Les pointillés sur le gel réel figurent les localisations des protéines dans le cas idéal et illustrent la problématique d'alignement entre différents gels.

Ainsi, la séparation des protéines se fera exclusivement en fonction de leur poids moléculaire. La distance de migration du polypeptide-SDS est proportionnelle au logarithme de son poids moléculaire et varie en fonction de la concentration en polyacrylamide du gel. Le gel est réalisé de manière à présenté un gradient de cette concentration suivant l'axe des ordonnées qui sera l'axe des valeurs de masse moléculaire et une concentration constante suivant l'axe des abscisses. Le gradient de concentration constitue un véritable tamis, retenant les molécules de forte masse moléculaire sur le haut du gel, en début de migration et les protéines de masse inférieure, dans la partie basse du gel, en fin de migration. C'est donc ce gradient qui, associé au champ électrique appliqué, va définir le chemin de migration des protéines. C'est pourquoi le gradient de concentration en polyacrylamide doit être le plus régulier et le plus répétable possible et le champ électrique le plus homogène malgré les fuites de courant déjà évoquées. Dans la réalité des manipulations en laboratoire la perfection n'est jamais atteinte et les chemins de migration subissent une déformation qui peut être amplifiée par la déformation physique du gel lors de l'acquisition des images. Le gel n'est en effet pas une matière rigide et l'assèchement inhomogène de sa surface peut entraîner des contractions locales.

IV.1.3.3 Marquage des protéines

À l'issue des deux étapes de migration, les protéines se sont réparties en fonction de leurs caractéristiques (pI et masse moléculaire) sur la surface du gel. Cependant, ce gel n'est pas directement exploitable car les protéines restent invisibles. Différentes techniques de colorations existent. La coloration à l'argent est considérée comme la référence en ce qui concerne la sensibilité pour visualiser les protéines minoritaires, mais d'autres colorants sont utilisés comme le bleu de Coomassie, le SYPRO ruby (BioRad, Hercules, CA, USA) [40]. Nous reviendrons plus en détail sur la technique de marquage en fluorescence qui a permis le développement de la technologie DIGE. Le résultat d'une expérience d'électrophorèse bidimensionnelle est l'image acquise au moyen de système d'imagerie de technologies diverses et dont la dynamique des niveaux de gris est généralement de 12 ou 16 bits.

IV.1.4 Applications de l'électrophorèse bidimensionnelle

L'un des objectifs de ce type d'approche peut être d'identifier l'expression différentielle des protéines entre deux types d'échantillons ayant migrés sur des séries de gels d'électrophorèse bidimensionnelle. L'analyse de cette expression différentielle revêt une importance primordiale dans le cadre du projet NODDICCAP. Il s'agit, en effet, d'identifier sur les gels, les taches protéiques qui, d'un type d'échantillon à l'autre, sont inhibées (disparitions), induites (apparitions) ou bien encore changent de niveau d'expression entre les deux échantillons, de façon significative et pour des raisons biologiques. Une fois ces taches d'intérêt sélectionnées, il est nécessaire d'identifier les protéines qu'elles contiennent par spectrométrie de masse.

IV.1.5 Considérations générales sur les images de gels d'électrophorèse bidimensionnelle

Afin d'envisager une exploitation efficace des images il est nécessaire de prendre en considération les caractéristiques particulières de la technique d'électrophorèse bidimensionnelle, ainsi que les nombreux biais liés à son caractère expérimental. Ce paragraphe fait le point sur l'ensemble des particularités et des biais qui sont à l'origine d'autant de difficultés lors de l'analyse et l'interprétation des images.

Le travail d'identification des taches protéiques est un point crucial de l'interprétation des gels. Ce travail est long et fastidieux, car il nécessite, au préalable, d'avoir repéré correctement toutes les taches protéiques sur les images de gel. Les interprétations nécessitent également souvent la possibilité de comparer des gels entre eux afin de mener des études différentielles par classe.

Bien qu'à première vue la résolution de l'électrophorèse bidimensionnelle soit impressionnante, elle reste largement insuffisante comparée au nombre et à la diversité des protéines cellulaires. Cette résolution pâtit également du phénomène très commun de co-migration de protéines au sein d'une même tache protéique [42]. Par ailleurs dans certaines régions des gels, le chevauchement ajouté à la saturation des taches protéiques rend leur localisation très difficile voire impossible ce qui signifie

qu'il ne sera pas possible de leur associer un groupement protéique et encore moins de les quantifier. L'emploi de gels à l'échelle de pH étroite permet de séparer des protéines dont les points isoélectriques sont proches, et apporte une solution à ce problème. Cependant cette technique a ses limites car il est très difficile de combiner plusieurs de ces gels afin d'obtenir l'information sur une large échelle de pH. Par ailleurs, une autre difficulté vient de la diffusion des taches protéiques et de la présence de traînées, suivant les deux dimensions de la migration.

Anderson et al., dans leur article sur les perspectives offertes par l'étude du protéome humain [4], constatent que la dynamique d'expression des différentes protéines des cellules et des tissus est très grande. Il a ainsi été estimé qu'un tiers des taches protéiques pouvait constituer plus de 75% de la quantité totale des protéines d'un échantillon, le tiers le plus faible en représentant alors moins de 6%. La gamme dynamique entre les protéines les moins exprimées et les plus exprimées peut atteindre un facteur de 10^6 pour les cellules et les tissus et jusqu'à 10^{12} pour les fluides corporels comme le plasma. Étant donné que la gamme dynamique de l'électrophorèse bidimensionnelle est de l'ordre de 10^4, de nombreuses taches correspondant à des protéines minoritaires ne sont pas détectables ou alors sont noyées dans le bruit et nécessitent un œil expert afin les localiser. Par ailleurs, l'intensité du fond d'une image de l'électrophorèse bidimensionnelle peut varier suivant la zone considérée. Les plus grosses variations d'intensité ont lieu suivant la deuxième dimension de migration (la masse moléculaire). Les bords du gel présentent également un biais puisque, généralement, l'intensité y est particulièrement élevée.

L'analyse des gels nécessite non seulement de s'intéresser à l'intensité des taches protéiques mais également à leur disposition sur la surface du gel. Par la suite, nous parlerons de « patron » pour désigner cette répartition géographique et la géométrie des taches protéiques sur un gel. Les patrons subissent des déformations qui sont principalement liées au processus de fabrication des gels, au phénomène de polymérisation et à l'étape de migration des protéines. Il existe deux principaux modèles de migration des protéines dans un gel : le modèle standard de Ogston-Morris-Rodbard-Chrambach [47] [36] et le modèle de reptation [24]. Le principal facteur de déformation du patron est dû aux fuites de courant provoquant une inhomogénéité du champ électrique nécessaire à la migration des protéines. Il faut y ajouter la combinaison de facteurs mineurs entraînant des distorsions locales.

Ce sont quelques unes des difficultés auxquelles doit faire face l'automatisation de la détection des taches protéiques dans le processus d'exploitation informatique des images de gels. L'application de méthodes statistiques et automatisées pour la protéomique est devenue indispensable. En effet, pour le biochimiste, l'analyse différentielle de deux gels d'électrophorèse bidimensionnelle pouvant contenir des milliers de protéines candidates nécessite de nombreuses heures. Par ailleurs, la difficulté d'alignement inhérente à la déformation des gels vient s'ajouter à la lourdeur de la tâche et rend ce travail pratiquement impossible. Comme nous le verrons dans le paragraphe suivant, la technologie DIGE ne résout que partiellement ce problème d'alignement car le problème resurgit quand il s'agit de comparaisons inter-gels. Par ailleurs, il peut également se produire que des taches protéi-

ques de faible intensité disparaissent d'un gel à l'autre à cause de la limite de sensibilité du marquage ou de l'acquisition.

IV.2 La DIGE : électrophorèse bidimensionnelle différentielle

IV.2.1 Principe de la technologie DIGE

La technique de l'électrophorèse bidimensionnelle différentielle (disponible commercialement sous le nom bioMérieux DIGE; Amersham biosciences, Uppsala, Suède) [56] consiste à employer différents esters de succinimidyle de colorants cyanines afin de marquer en fluorescence plusieurs populations protéiques différentes avant de mélanger ces populations et de les faire migrer au sein du même gel. Il s'agit en fait du multiplexage des populations protéiques au sein d'un même gel (voir Figure 6).

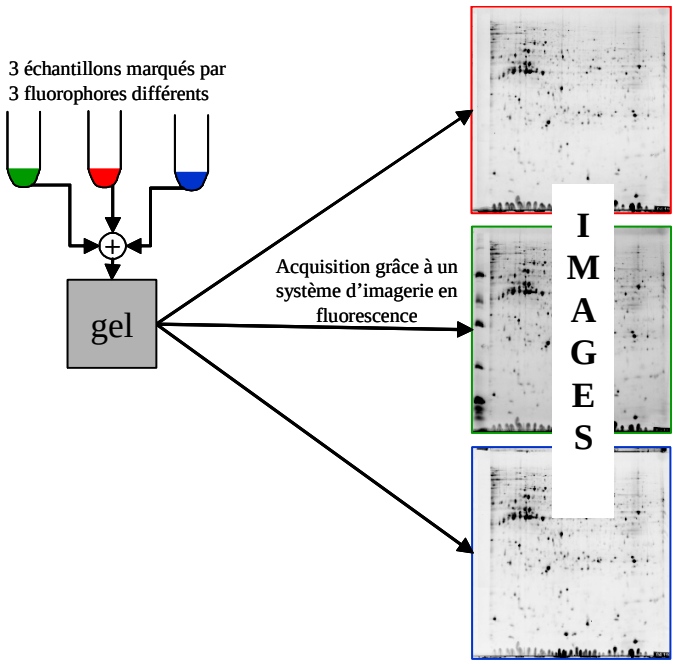


Figure 6 : Principe du multiplexage DIGE. Un seul et même gel est utilisé comme le support du procédé d'électrophorèse bidimensionnelle de 3 populations protéiques distinctes.

Classiquement, les trois fluorophores utilisés sont trois cyanines différentes usuellement notées : Cy2, Cy3 et Cy5. Ces 3 fluorophores ont leur masse et leur charge équivalente et présentent des spectres d'émission suffisamment différents pour être séparés à l'aide d'un appareillage optique adapté.

IV.2.2 Acquisition des images issues de la technique DIGE à l'aide d'un système d'imagerie à fluorescence

Afin de mieux appréhender la nature des images que nous aurons à exploiter par la suite, il est nécessaire de s'intéresser à la méthode et à la technologie employées pour leur acquisition. La plateforme technologique du projet NODDICCAP comprend le système d'imagerie par fluorescence nommé « ProXPRESS™ Proteomic

Imaging System » dont une photographie est présentée en Figure 7. Ce système d'imagerie, développé par PerkinElmer®, permet l'acquisition des images issues des gels DIGE.

Figure 7: Photographie du système d'imagerie par fluorescence utilisé au laboratoire de protéomique de bioMérieux (PerkinElmer ProXpress)

Le ProEXPRESS est un scanner à fluorescence à longueurs d'onde multiples capable de réaliser des mesures de densité optique et de fluorescence sur divers media tels que les gels, les lames de microscopie ou bien encore sur des milieux culture. La surface d'acquisition maximale est de 230 x 280 mm avec une résolution pouvant aller jusqu'à 50 microns. Les mesures de fluorescence peuvent être faites en utilisant n'importe quelle combinaison de six longueurs d'ondes d'excitation et d'émission dans un intervalle allant de 400 à 750 nm. Les longueurs d'ondes d'excitation correspondent aux différentes longueurs d'ondes obtenues après l'application de filtres à la source d'excitation. Les longueurs d'ondes d'émissions sont les longueurs d'ondes émises par l'échantillon et observables à l'aides de filtres d'absorption. Bien que ce ne soit pas utile dans le cadre de l'acquisition d'images de gels DIGE, il faut noter qu'il est également possible de réaliser des acquisitions sans illumination pour les applications en luminescence ainsi que des acquisitions dites en "transillumination" pour les applications nécessitant une excitation dans l'ultraviolet ou en lumière blanche.

Le gel, maintenu par deux plaques de verres (à faible fluorescence), est placé horizontalement sur un tiroir automatisé de l'appareil. Il est important d'apporter un soin particulier à la position du gel sur ce tiroir afin qu'il se situe dans le plan focal de la caméra et que l'illumination et l'acquisition se fassent dans les meilleures conditions. La caméra CCD (« Charge-Coupled Device ») acquiert l'image en capturant le nombre nécessaire d'imagettes dont la taille fixe est d'environ 50 x 70 mm. Ces imagettes sont ensuite recombinées afin de former l'image complète de la surface précisée préalablement par l'opérateur.

IV.2.3 Avantages, applications et limitations de la DIGE

Le principal avantage de la technique DIGE est qu'elle supprime le problème lié à la mise en correspondance des taches protéiques entre différents échantillons. Cet avantage n'est cependant effectif qu'à condition de travailler sur les images issues d'un même gel. Afin d'exploiter au mieux le potentiel lié à cette technique, il est devenu courant d'utiliser, au cours d'une étude différentielle entre deux classes d'échantillons (une classe d'échantillon désigne un sous-groupe d'échantillons réunissant une caractéristique commune : il s'agit typiquement d'un sous-groupe d'échantillons témoins ou d'un autre d'échantillons pathologiques), un standard commun à tous les gels. La Figure 8 propose une représentation schématique des images produites à l'issue d'une expérience DIGE menée selon ce schéma expérimental. L'échantillon standard est constitué du mélange stœchiométrique de tous les échantillons étudiés. Il est marqué en fluorescence par une troisième cyanine. Le standard permet la normalisation inter-gels des mesures de l'abondance en protéine [61] et [1]. Des plates-formes logicielles commerciales dédiées à la technologie DIGE sont disponibles (Decyder, Amersham Biosciences, ImageMaster 2D Platinium). Néanmoins, l'approche de normalisation est toujours dépendante de la qualité des données générées lors de l'étape de mise en correspondance inter-gels des taches protéiques. Finalement, l'analyse des images issues d'une expérience se heurte aux mêmes problématiques que pour l'électrophorèse classique du fait de la nécessité du raisonnement « inter-gels ». D'un point de vue qualitatif, les variations du fond, observées sur les images issues d'expériences de DIGE, s'expliquent par l'imperfection de la technologie et surtout par la présence de fluorescences parasites inhérentes à l'échantillon lui-même ou à la matrice (gel). Les imperfections technologiques proviennent notamment de la difficulté de focalisation lors de l'acquisition (le signal et le bruit voisins du point considéré peuvent être intégrés au signal acquis) et de fuites du laser lors de l'acquisition [56]. Elles sont cependant négligeables devant la fluorescence de fond liée à l'échantillon même.

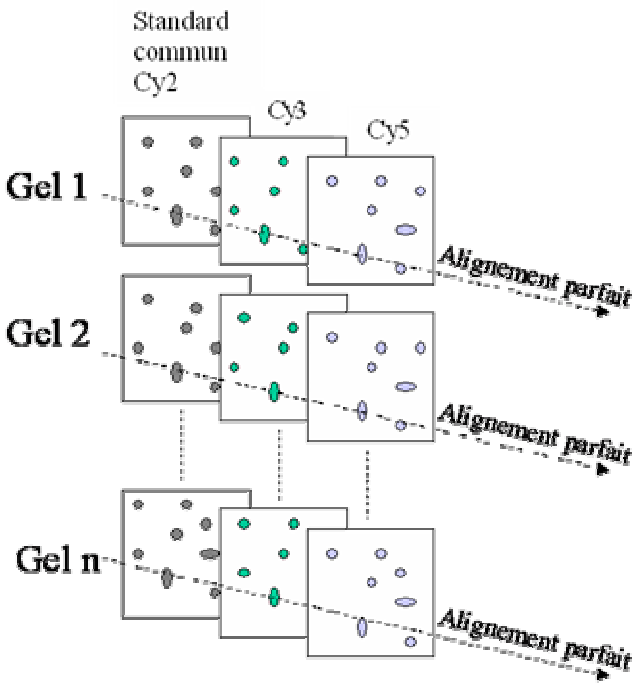


Figure 8 : Représentation schématique des images issues des n gels d'une expérience DIGE. Les 3 fluorophores permettent de distinguer 3 populations. L'échantillon témoin marqué par le Cy2, est appelé standard commun et est constitué du mélange de tous les échantillons utilisés.

La DIGE résout de nombreux problèmes associés à l'électrophorèse bidimensionnelle classique. Les variations gel à gel observées classiquement sont mieux maîtrisées d'une part grâce à l'utilisation d'un seul gel pour plusieurs populations et d'autre part grâce à la présence d'un standard interne permettant la normalisation inter-gels. Par ailleurs, la gamme dynamique est beaucoup plus étendue qu'en électrophorèse bidimensionnelle classique et permet des études quantitatives plus sensibles et plus précises.

IV.3 Méthodes de la littérature pour l'exploitation de la technologie DIGE

IV.3.1 Introduction

Comme le montre la Figure 9, il n'existe pas un unique processus pour l'exploitation des données de l'électrophorèse bidimensionnelle. Néanmoins, les différents processus possibles contiennent nécessairement une étape de prétraitement d'image et généralement des étapes de détection, de quantification et de mise en correspondance des taches protéiques.

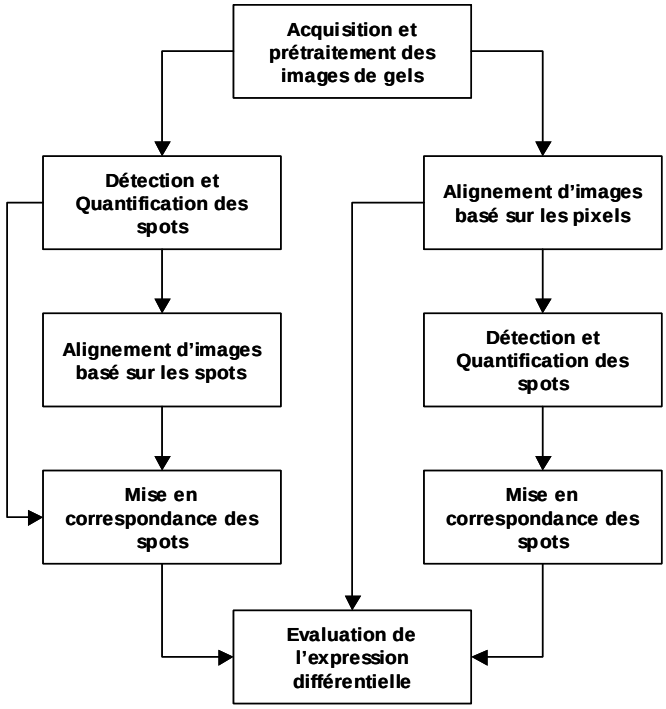


Figure 9: Principales approches pour le traitement informatique des images de gels d'électrophorèse bidimensionnelle dans le contexte d'une étude différentielle. L'alignement des images peut être basé sur les taches protéiques ou bien directement sur les pixels.

Par la suite nous allons présenter différentes méthodes de traitement d'images et de données reconnues et publiées, pouvant entrer dans un processus d'exploitation des données issues de l'électrophorèse bidimensionnelle DIGE. Si la liste des méthodes présentées ne peut être exhaustive, elle se veut pertinente, englobant à la fois les méthodes les plus courantes dans le domaine ou dans des domaines voisins ainsi que d'autres, pressenties comme judicieuses. L'intérêt de chacune des méthodes sera discuté, qu'elles soient ou non déjà mises en œuvre dans notre domaine d'application. Cette partie constituera un état de l'art à partir duquel, dans une seconde partie, nous construirons un processus adapté aux besoins et aux contraintes du projet NODDICCAP.

IV.3.2 Prétraitement des images

Les images de gel d'électrophorèse bidimensionnelle sont des images numériques, donc échantillonnées. Ces images sont vues comme des fonctions de $\mathbb{Z} \times \mathbb{Z}$ dans $\mathbb{Z}$. La représentation la plus utilisée est un tableau à 2 dimensions représentant les dimensions spatiales de l'image, dans lequel chaque valeur correspond à la mesure du signal d'intensité lumineuse en un point (en réalité une surface, la plus petite possible) du gel.

Les « signaux-images » sont sujets à de nombreux biais expérimentaux. Ces biais se traduisent par :

- une déformation géométrique du patron, c'est-à-dire de la répartition géographique et la géométrie des taches protéiques sur un gel,
- la présence d'un bruit de fond souvent non homogène sur la surface de l'image
- ainsi que par la présence de bruits de haute fréquence spatiale d'origines diverses telles que des particules de poussière, des traînées et parfois des bulles d'air. Les traînées, souvent horizontales, sont le résultat de la contamination par des acides nucléiques ou de la présence de sels lors de la première migration.

Le paragraphe présente les méthodes de traitement d'image (et donc du traitement du signal) les plus adaptées pour la suppression de ces biais qu'elles appartiennent ou non à la bibliographie existante dans le domaine de l'électrophorèse bidimensionnelle.

IV.3.2.1 Suppression du bruit

Quelle que soit la technique employée pour leur obtention, les images de gels d'électrophorèse bidimensionnelle sont affectées de bruits de nature diverse. Ces bruits en deux familles selon leur fréquence spatiale : les bruits de haute fréquence spatiale qui sont dû aux poussières, aux bruits électroniques de l'appareil d'acquisition, aux bulles d'air, aux taches de coloration, aux cassures éventuelles du gel, aux traînées des taches protéiques et les bruits basse fréquence qui se traduisent par une inhomogénéité du fond de l'image et qui peuvent être dus à l'imperfection de l'appareil d'acquisition, à un défaut de coloration ou du marquage, ou bien encore à des propriétés intrinsèques de la matrice comme la fluorescence naturelle du gel dans le cas de l'utilisation de fluorophores pour le marquage.

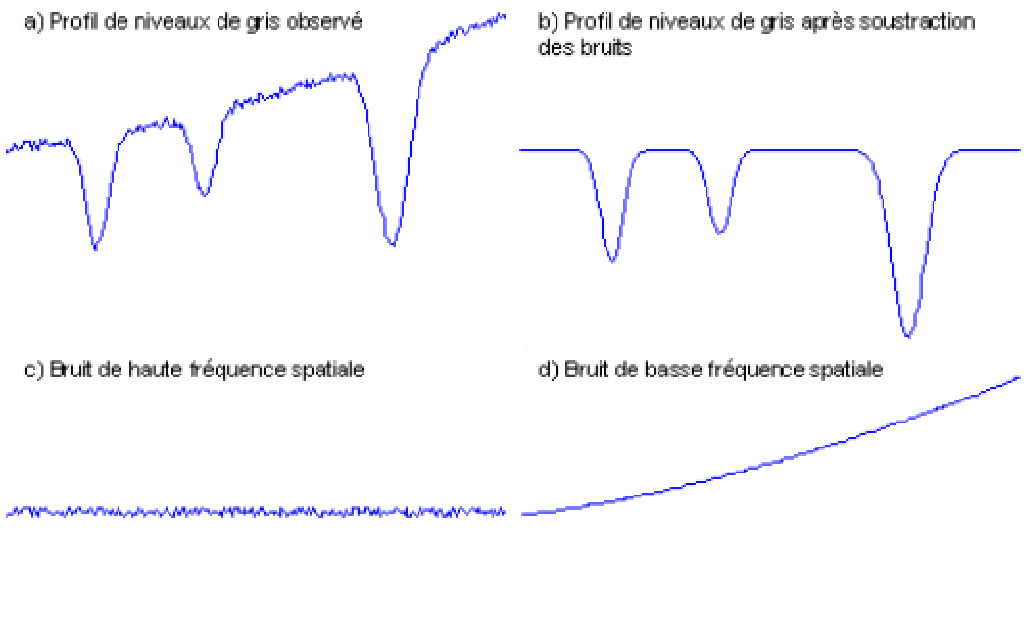


Figure 10: Profils des niveaux de gris. Le signal observé (a), le signal idéal recherché (b), le bruit haute fréquence spatiale (c) et la ligne de base (d).

Les bruits hautes fréquences entravent la détection des taches protéiques notamment lorsque celle-ci est basée sur la dérivée du signal (image) qui est très sensible à ce type de bruit. Ils provoquent de fausses identifications et de mauvaises estimations des frontières des taches protéiques. L'inhomogénéité du fond, quand elle n'est pas due à un phénomène biologique, doit être prise en compte car elle affecte non seulement la quantification des taches protéiques mais aussi la qualité des résultats de méthodes basées sur la segmentation de l'histogramme. Ces inhomogénéités entraînent également des difficultés d'alignement d'images lorsque la méthode employée consiste à maximiser la similarité.

3.2.1.1 Suppression du bruit de haute fréquence spatiale

Un simple seuillage peut parfois supprimer le bruit et améliorer l'impression visuelle de contraste. Dans le cas des images de gels d'électrophorèse, l'intensité des différents bruits couvre toute la plage d'intensité occupée par les taches protéiques. Comme nous allons le voir par la suite, un filtrage local est plus approprié. Le filtrage local peut être linéaire, non linéaire, adaptatif ou non adaptatif.

a) Filtrage linéaire

Le filtrage linéaire regroupe l'ensemble des méthodes calculant la nouvelle valeur $I'(i,j)$ du niveau de gris du pixel courant (i,j) à partir d'une combinaison linéaire des niveaux de gris des pixels du voisinage. Cette nouvelle valeur du niveau de gris peut ainsi s'écrire :

$$I'(i,j) = \sum_{(m,n)\in V(i,j)} h(I(m,n)),$$

avec $I(m,n)$ la valeur du niveau de gris du pixel (m,n), $V(i,j)$ le voisinage du pixel courant (i,j) et h une fonction de pondération. Cette équation représente la technique de filtrage la plus communément employée comme par exemple par le lo-

giciel Melanie™ [6]. Elle est également appelée convolution discrète de l'image par le noyau h. Par exemple, lorsque les coefficients de pondération décrits par h sont égaux, le filtre est appelé « filtre moyen », la nouvelle valeur attribuée au pixel courant étant la moyenne des intensités des pixels de son voisinage.

De la même manière, si le noyau h est gaussien ou bien calculé suivant les lois de la diffusion le filtrage devient gaussien ou par diffusion.

Ces formes de filtrage suppriment efficacement le bruit mais présentent aussi le désavantage d'altérer les contours des taches protéiques. C'est pourquoi il est opportun d'utiliser des filtrages linéaires adaptatifs, c'est à dire d'adapter la fonction de pondération h à la topologie locale de l'image. C'est le cas du filtrage de Wiener [35] qui dans les régions de grande variance atténue la force du filtrage et respecte ainsi le contour des taches protéiques. Ce filtrage correspond à la recherche d'un optimum parmi les filtres linéaires, par rapport à un critère de minimalisation d'erreur quadratique moyenne instantanée (méthode bayésienne). Une approximation simple du filtre de Wiener peut être réalisée à partir de l'estimation de la moyenne et de la variance locale du bruit :

$$\mu = \frac{1}{MN} \sum_{(m,n)\in V(i,j)} I(m,n)$$

$$\sigma^2 = \frac{1}{MN} \sum_{(m,n)\in V(i,j)} I^2(m,n) - \mu^2$$

avec M, N les dimensions de la fenêtre considérée pour le voisinage $V(i,j)$. La valeur de sortie du filtre de Wiener est alors :

$$I'(i,j) = \mu + \frac{\sigma^2 - v^2}{\sigma^2}\left(I(i,j) - \mu\right)$$

avec v^2 la variance globale du bruit qui peut être estimée comme la moyenne des variances locales.

L'approche proposée par Seillier-Moiseiwitsch [51], qui fait appel à la transformée en ondelette, appartient également à la catégorie des filtres linéaires. Les images de gels d'électrophorèse bidimensionnelle contiennent des taches protéiques de tailles et d'intensités très variables et correspondent donc au champ des applications de la transformée en ondelette. En effet, cette transformation fournit l'outil « espace-fréquence » et « temps-échelle » et permet la distinction entre le bruit et le signal (c'est-à-dire l'intensité mesurée liée à la quantité protéique) quelque soit l'échelle considérée. La méthode se décompose en trois étapes :

- il s'agit d'abord de décomposer l'image sur une base orthonormée d'ondelettes
- puis de seuiller judicieusement les coefficients des ondelettes
- et enfin d'appliquer la transformation inverse à partir des coefficients seuillés.

b) Filtrage non linéaire

Il s'agit notamment des filtres par diffusion inhomogène, des filtres d'ordre et des filtres morphologiques.

Le filtre médian [22], [52] possède des propriétés intéressantes pour l'application aux images de gels d'électrophorèse bidimensionnelle. Il est notamment indépendant d'éventuels étirements de contraste, c'est à dire qu'il agira de la même

manière pour deux images dont la dynamique des niveaux de gris est différente. Tout comme les filtres linéaires adaptatifs, le filtre médian permet également une bonne conservation des contours. Le filtre médian transforme l'image I en une image I', telle que pour tout pixel (i, j), le niveau de gris $I'(i, j)$ est la valeur médiane des niveaux de gris des pixels I appartenant au voisinage $V(i, j)$:

$$I'(i, j) = med\left(I(m,n) \middle| (m,n) \in V(i, j)\right)$$

Étant donné que la médiane représente une alternative à la moyenne, le filtre médian constitue une forme de lissage.

Un autre type de filtrage, mise en œuvre par David Fournier [20] dans une problématique voisine (analyse de bactériophages), permet également le respect des contours. Il s'agit des filtres morphologiques basés sur les opérateurs de reconstruction géodésique tels que les décrit Jean Serra dans son ouvrage de référence "Image Analysis and mathematical Morphology" [52]. Le choix le plus approprié aux images de gel d'électrophorèse est le filtrage morphologique alterné séquentiel. Il correspond à la composition de filtres alternés d'activité croissante :

$$\varphi_n\gamma_n \circ \ldots \circ \varphi_3\gamma_3 \circ \varphi_2\gamma_2 \circ \varphi_1\gamma_1$$

où γ et φ sont l'ouverture et la fermeture morphologiques par un élément structurant. Cet élément structurant augmente avec le niveau de la composition allant de 1 à n. L'ouverture agit sur les structures claires, la fermeture sur les structures sombres. Le nombre d'itérations du filtrage ainsi que l'ordre des combinaisons des séquences d'ouvertures et de fermetures sont à définir de manière empirique. L'ordre n'est cependant pas un facteur clef car il n'induit pas une différence notable des résultats. Pour des images médicales (analyse de bactériophages) d'aspects proches des images de gels d'électrophorèse bidimensionnelle, deux itérations ont été nécessaires au filtrage.

Les filtrages que nous venons de présenter sont valables pour les bruits de hautes fréquences spatiales « classiques » tels que les poussières, les bulles ou le bruit électronique du système d'acquisition, mais ils sont souvent inefficaces en ce qui concerne les bruits liés aux traînées et aux cassures éventuelles au sein du gel. Ces bruits se caractérisent par leur forme allongée, souvent verticale ou horizontale.

L'usage de filtres morphologiques apparaît alors tout à fait approprié et peut venir compléter un premier filtrage classique du bruit de hautes fréquences spatiales. Ces filtres consistent en l'usage des opérateurs d'ouverture et de fermeture définis en morphologie mathématique et généralisés aux images en niveaux de gris. Ainsi, le choix d'une barre verticale, horizontale ou bien encore d'une direction particulière comme élément structurant de ces opérateurs permet d'estimer l'image où seules restent les traînées et les cassures [53]. Il suffit alors de soustraire cette image contenant uniquement les motifs artefactuels indésirables de l'image originale. Il est cependant important d'employer ces filtres avec prudence tant la notion de traînée sur un gel d'électrophorèse reste délicate : on peut être amené à dénaturer le signal porteur d'un sens biologique, notamment lorsque le chevauchement de plusieurs taches protéiques constituent des formes allongées non discernables de réelles traînées par cette méthode. Par ailleurs, sous certaines conditions, les traînées et cassures peuvent être

considérées comme un bruit de basse fréquence spatiale et ainsi être supprimées lors de l'étape de soustraction de « la ligne de base ».

3.2.1.2 Suppression du bruit de basse fréquence spatiale

L'opération de suppression du fond consiste à retrancher des niveaux de gris de l'image d'un gel, la fonction de niveau de gris traduisant les variations de grandes amplitudes spatiales du niveau de gris. Comme décrit plus haut, cette étape est la plupart du temps indispensable pour une meilleure détection, quantification et mise en correspondance des taches protéiques ainsi que pour l'alignement des images.

La première idée consiste en l'évaluation d'un bruit de fond moyen, homogène sur toute la surface de l'image. Pour cela, il est par exemple possible, à partir d'une zone de l'image ne contenant que ce que l'expert biologiste aura jugé comme du bruit, de calculer un niveau moyen que l'on retranchera à l'image entière. Cette technique est très perfectible car elle ne tient pas compte des variations spatiales. Dès 1986, Tyson et Haralick [55] ont proposé une estimation du fond par interpolation des niveaux de gris entre les minima locaux préalablement identifiés. Le logiciel Melanie™ [5] identifie le niveau de gris minimum de l'image afin de lui retrancher puis ajuste un polynôme du troisième degré à l'image dont les taches protéiques ont été retirées. Une autre solution est une nouvelle fois apportée par la morphologie mathématique : l'utilisation d'une sphère comme élément structurant pour l'opération d'ouverture morphologique (les niveaux de gris étant vus comme une hauteur) permet en effet d'estimer le fond de l'image. Le diamètre de la sphère doit être plus grand que la plus grande des taches protéiques de l'image mais de manière à ce que la courbure de la sphère reste supérieure à celle du fond.

IV.3.2.2 Amélioration du contraste et manipulation de l'histogramme

L'histogramme d'une image de gel d'électrophorèse bidimensionnelle prend généralement l'allure de celui de la Figure 4.

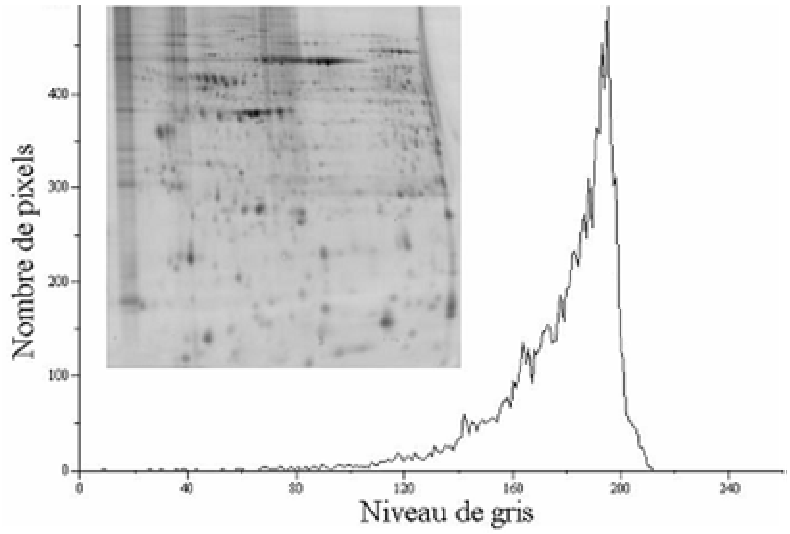


Figure 11: Histogramme d'une image de gel d'électrophorèse bidimensionnelle

Les valeurs dans la région des pixels clairs sont dues à l'arrière plan de l'image. Celles dans la région des pixels foncés viennent des taches protéiques. Pour améliorer l'affichage des taches protéiques à l'écran, on peut ré-étaler l'histogramme. Cela permet simplement d'améliorer la qualité visuelle d'une image et donc le confort pour le biologiste dans ses interprétations.

Le principe de l'algorithme consiste à identifier le niveau de gris minimal, $\min$, et le niveau de gris maximal, $\max$ de l'image afin de normaliser entre 0 et 1 le niveau de gris de chaque pixel. Il s'agit ensuite d'appliquer une transformation F judicieusement choisie (il s'agira souvent d'une sigmoïde), sachant que l'œil humain a une meilleure sensibilité dans le sombre que dans le clair et que l'on s'intéresse souvent à une gamme de niveaux de gris limitée, correspondant à celle occupée par les taches protéiques. La nouvelle intensité $I'(i,j)$ d'un pixel de l'image sera comprise dans l'intervalle $[MIN, MAX]$ choisi par l'opérateur :

$$I'(i,j) = F\left(\frac{(I(i,j) - \min}{\max - \min}\right) * (MAX - MIN) + MIN$$

L'étalement de l'histogramme est une opération linéaire sur les niveaux de gris qui respecte une quantification relative : les rapports volumiques établis entre les spots d'une même image sont inchangés par la transformation. Par ailleurs, dans les procédés classiques de traitement et d'analyse des quantifications des taches protéiques, une étape de normalisation, consistant souvent en l'égalisation des distributions des quantifications « inter-images », peut rendre l'opération d'étalement transparente.

Le logiciel Mélanie™ et toutes ses évolutions (ImageMaster™), proposent une égalisation de l'histogramme de type exponentiel, « Rayleigh », racine cubique ou logarithmique. Ils proposent également le ré-étalement de l'histogramme ainsi que l'amélioration du contraste.

IV.3.2.3 Correction de la distorsion de l'image

La principale cause de déformation des images de gels d'électrophorèse bi-dimensionnelle est liée au phénomène physique de fuite de courant lors de la phase de deuxième dimension. En effet, ces fuites de courant sont à l'origine de l'inhomogénéité de la migration sur la surface du gel, entraînant une courbure, surtout visible sur les bords du gel, des lignes horizontales de masse égale. John S. Gustafsson [25] tient compte de ce phénomène afin de modéliser la déformation subie. Il est crucial de corriger cette déformation lorsqu'il s'agit d'atteindre les coordonnées pI et M_r des taches protéiques, mais ce traitement est superflu en ce qui concerne les expériences d'analyse différentielle pour lesquelles l'alignement inter-gels est suffisant.

IV.3.3 Analyse des images

Quel que soit le domaine d'application, l'analyse d'image a pour but l'extraction de l'information utile. Cette extraction est souvent basée sur des connaissances a priori quant à la nature de l'information attendue. Par exemple, l'analyse d'une image de code barre est très simple du fait de la connaissance précise des motifs recherchés : des traits verticaux avec un fort contraste. L'analyse peut alors consister en un simple seuillage du profil de niveau de gris observé sur une ligne horizontale. Le résultat de ce seuillage porte toute l'information recherchée.

Dans notre contexte, pour chaque polypeptide, l'information recherchée est le changement de son abondance d'une image à l'autre. Pour parvenir à extraire cette information, il faut dans un premier temps pouvoir détecter chaque polypeptide (c'est-à-dire chaque tache protéique), puis, dans un second temps, être capable de mettre en correspondance, d'une image à l'autre, les polypeptides détectés identiques. La première étape de l'analyse est donc la détection des taches protéiques pour lesquelles nous disposons de connaissances a priori basées sur les observations de taches typiques. Nous détaillerons ces observations dans le paragraphe IV.3.3.1. La seconde étape est la mise en correspondance de ces taches.

IV.3.3.1 Observation et caractérisation des taches protéiques

3.3.1.1 Considérations générales sur la forme des taches protéiques

Visuellement, l'identification d'une tache protéique requiert une certaine pratique ainsi que des connaissances en protéomique et sur le procédé de l'électrophorèse. La Figure 12 présente un exemple de la vue dans le plan ainsi que la vue pseudo tridimensionnelle (la troisième dimension étant le niveau de gris des pixels) d'un regroupement de taches protéiques tels qu'ils sont classiquement observés sur les images de gel d'électrophorèse.

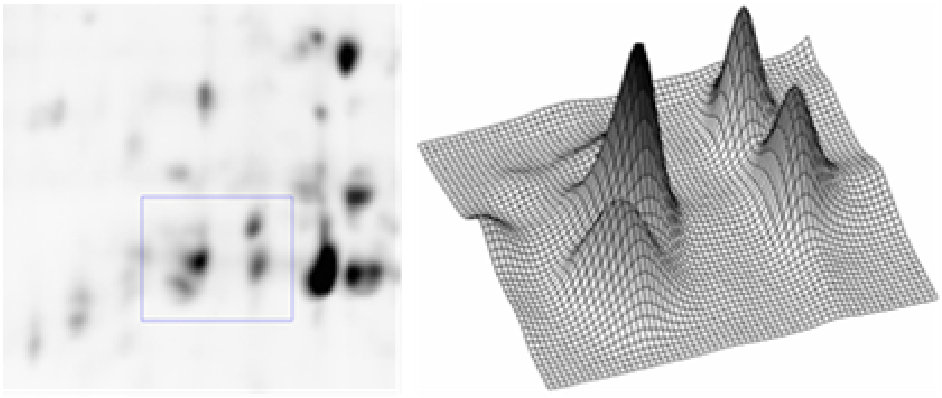


Figure 12: visualisation 3D du groupe de taches protéiques isolé dans le rectangle bleu. Cette visualisation 3D met en évidence la présence d'épaulements entre les taches proches.

D'une manière générale et comme l'illustre la Figure 13, toutes les petites surfaces foncées, distinctes du fond et grossièrement circulaires sont comptées comme des taches protéiques. Du fait de la migration, leur forme peut se rapprocher de celle d'une goutte.

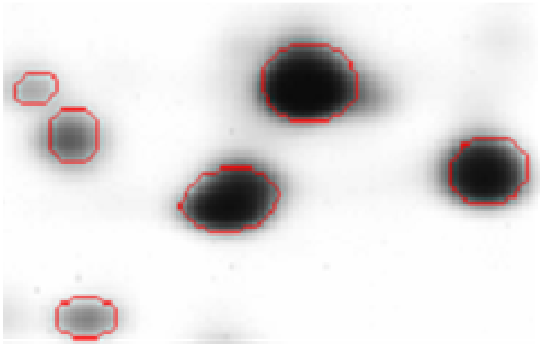


Figure 13 : Détail d'une image de gel d'électrophorèse. Les taches protéiques détourées en rouge présentent ici une allure grossièrement circulaire. C'est la forme la plus couramment observée.

La migration des protéines dans le gel se faisant verticalement (par rapport au poids moléculaire) et horizontalement (par rapport au pH), les taches protéiques possèdent généralement une symétrie suivant l'axe horizontal et l'axe vertical. En présence d'une symétrie autre, on peut soupçonner la présence de plusieurs taches protéiques se chevauchant.

3.3.1.2 Situations particulières

Sur une image de gel d'électrophorèse bidimensionnelle, certaines régions doivent faire l'objet d'une attention particulière.

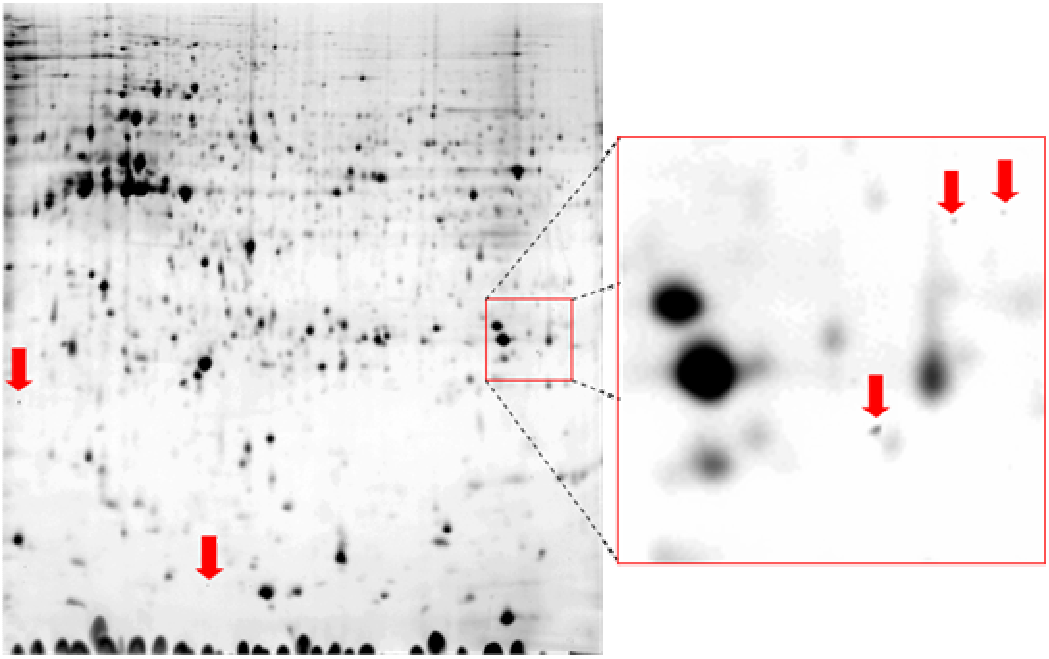


Figure 14:Les flèches rouges indiquent la présence d'un bruit que l'on peut qualifier de poivre et sel et qui est dû à des poussières ou à de petites taches de coloration. Il se caractérise par une aire faible et par un fort gradient.

Les petites taches (

Figure 14) n'occupant que quelques pixels ne doivent pas être prises en compte car elles sont dues aux poussières ou bien aux défauts du gel. Les cassures du gel doivent également être exclues de l'analyse. C'est le rôle du prétraitement décrit précédemment.

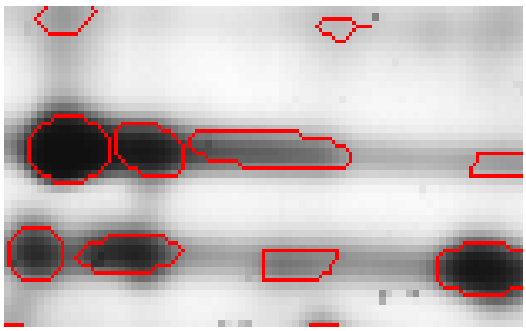


Figure 15: Identification (en rouge) de taches protéiques dans les traînées engendrées lors de la migration de première dimension (pH).

Dans les traînées (Figure 15), l'étude du « relief des niveaux de gris » doit permettre, par identification des épaulements, de déterminer la présence de taches protéiques sur une traînée.

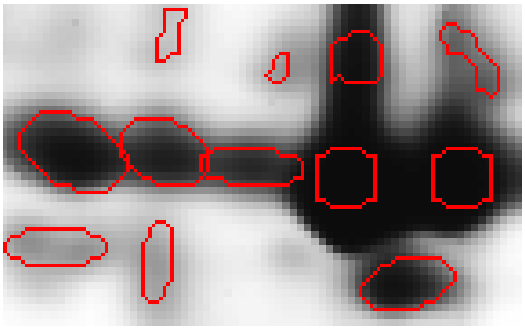


Figure 16: Identification (en rouge) de taches protéiques dans un amas

Certains motifs, comme celui présenté sur la Figure 16, sont de forme complexe. Ils peuvent contenir une ou plusieurs taches protéiques. Ils doivent faire l'objet d'un traitement particulier afin d'identifier la présence éventuelle de chevauchements. Ce traitement peut tenir compte de la forme volumique dans l'espace des niveaux de gris, avec une recherche d'épaulements. Elle peut tenir compte, de l'aspect de la forme planaire du regroupement (seule source d'information en cas de saturation). Dans ce cas là, il faut rechercher sur le contour de l'amas, les zones de resserrement (zones concaves), qui permettent de séparer deux taches protéiques. Les principales situations conflictuelles sont représentées sur le schéma de la Figure 17 :

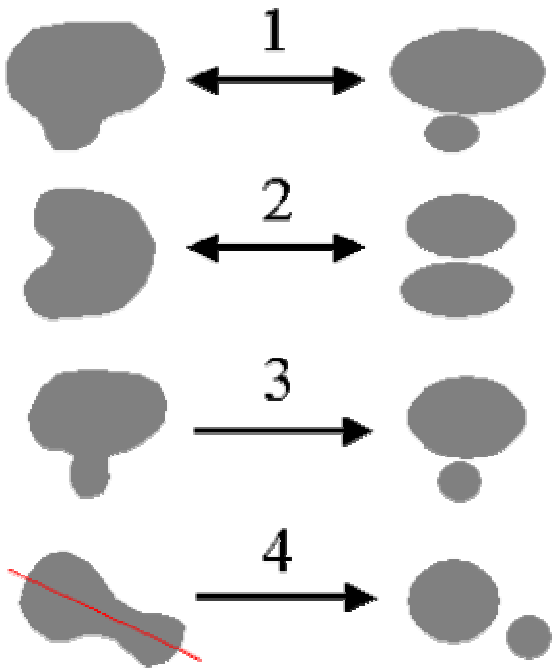


Figure 17 : Schémas des principales situations conflictuelles des taches protéiques

- Dans le cas n°1, le spot présente une protubérance sans resserrement, il est difficile de savoir s'il ne s'agit que d'une ou de deux taches protéiques. Une étude au niveau des épaulements de la forme volumique peut être décisive.
- Dans le cas n°2, il y a présence d'une seule zone concave sur la forme, il faut là encore se baser sur la forme volumique.
- Dans le cas n°3, l'excroissance présente un resserrement, il s'agit donc bien de deux taches protéiques distinctes.

- Enfin dans le cas n°4, la symétrie, qui n'est ni verticale, ni horizontale, doit permettre de soupçonner la présence de 2 taches protéiques.

Pour résumer, une tache protéique est une région du plan image qui peut être vue comme une montagne dans le relief en niveau de gris en considérant le noir comme le haut de l'échelle. Ces taches sont de taille variable, leur base est de forme grossièrement circulaire. Leur profil volumique est également très variable surtout au niveau des chevauchements. Ce profil volumique peut par ailleurs être affectée par le phénomène de saturation qui crée un plateau au centre des taches.

- Toute segmentation de l'image ayant pour but de repérer les taches protéiques doit donc s'efforcer de tenir compte des paramètres suivants :
- Son aire afin d'éliminer les cas de bruits ou de taches de coloration qui n'auraient pas été supprimés par un prétraitement.
- du dénivelé de niveau de gris entre son sommet et le fond de l'image,
- de son contour qui doit être grossièrement circulaire (et grossièrement convexe),
- de sa symétrie prépondérante suivant les axes horizontaux et verticaux,
- et de sa forme volumique qui, lorsqu'il n'y a pas saturation, peut se visualiser comme un sommet dans une chaîne montagneuse avec la présence de cols (épaulements) en cas de chevauchement.

3.3.1.3 Modélisation des taches protéiques

Dans la littérature, il n'existe que peu d'études ayant cherché à caractériser la forme des taches protéiques. La modélisation la plus répandue est basée sur une fonction gaussienne qui permet de restituer l'impression visuelle d'une tâche et qui permet une première approximation pour la quantification. Une autre approche consiste à tenir compte de la physique de formation des taches. Ces dernières sont formées par diffusion des protéines au sein du gel. Une modélisation basée sur le principe de diffusion, comme celle proposée par Bettens et al. [9], permet donc une représentation plus adéquate. En effet, le modèle basée sur la fonction gaussienne se rapproche du modèle de diffusion seulement lorsque la distribution initiale de concentration occupe une aire suffisamment petite et cette condition n'est que rarement observée dans la pratique.

Cependant, les modélisations gaussiennes et par diffusion supposent que la diffusion dans le gel soit parfaite. S'agissant des contours des taches, ceux générés par le processus physique théorique de diffusion sont réguliers et symétriques. Dans la pratique, le processus de diffusion est imparfait car il est sujet à de nombreux biais dont les paramètres ne sont pas maîtrisables. C'est pourquoi, les taches protéiques présentent des contours inattendus et irréguliers, s'éloignant parfois considérablement du modèle théorique de la diffusion parfaite.

Partant de ce constat Rogers et al. [48], proposent une approche plus pragmatique basée sur la construction automatique d'un modèle statistique des taches protéomiques. Le modèle est construit à partir d'un jeu de données réelles et permet d'identifier les principaux modes de déformation des spots.

L'approche consiste dans un premier temps à adopter un modèle de distribution de points permettant une représentation du contour des taches d'un jeu de données réelles et d'obtenir ainsi un jeu de données d'apprentissage. Le modèle complet

de la tache est obtenu par convolution de la surface (extraite par segmentation) par une fonction gaussienne bivariée (Figure 18). En considérant que la région extraite par segmentation correspond à la distribution protéique initiale, cette convolution permet en effet de simuler le phénomène de diffusion. Les différents paramètres du modèle, dont l'intensité de la tache et la dispersion de la gaussienne bivariée, sont ajustés au mieux à la forme réelle de la tache. La méthode d'optimisation utilisée l'algorithme de descente de gradient de Levenberg Marquardt [38]. Au final, le modèle de distribution de points se base sur 25 points régulièrement espacés sur le contour de chaque tache.

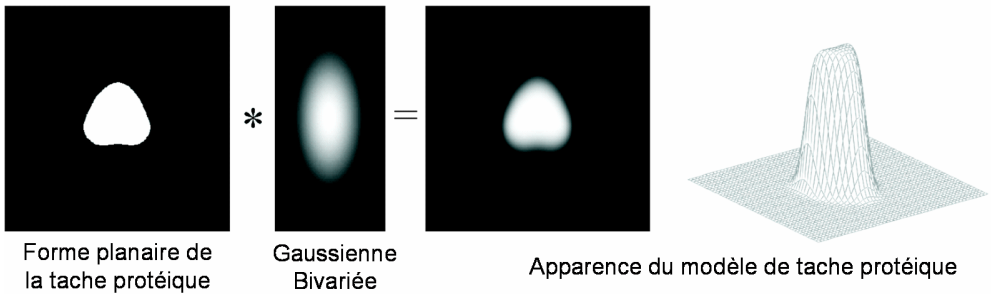


Figure 18: Formation du modèle complet d'une tache protéique par convolution de la forme planaire de la tache protéique avec une fonction gaussienne bivariée. Ce procédé correspond au processus de diffusion.

Dans un deuxième temps, Rogers et al [48] appliquent une analyse en composantes principales robuste sur le jeu de données d'apprentissage afin de déterminer les différents modes de variation du contour, tout en restant relativement insensible aux taches aberrantes inévitables. Les différents modes de déformation obtenus par Rogers et al à partir de leurs jeux de données sont représentés sur la Figure 19 et sont représentatifs du comportement des taches protéiques généralement observées.

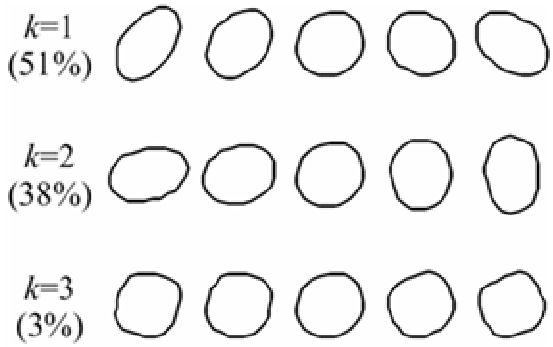


Figure 19 : Les trois premiers modes du modèle de distribution de points construit à l'aide d'une analyse en composantes principales robuste, permettant d'exclure de l'analyse les contours aberrants (correspondants généralement à des taches se chevauchant)

– Pour valider leur modèle de tache protéique, Rogers et al [48] le comparent aux modèles gaussien et par diffusion. Pour cela, chacun des modèles est ajusté par descente de gradient de Levenberg-Marquardt [38] à un même ensemble de taches protéiques préalablement défini, la valeur minimisée ayant été judicieusement choisie, les résidus de l'ajustement permettent une comparaison directe des modèles.

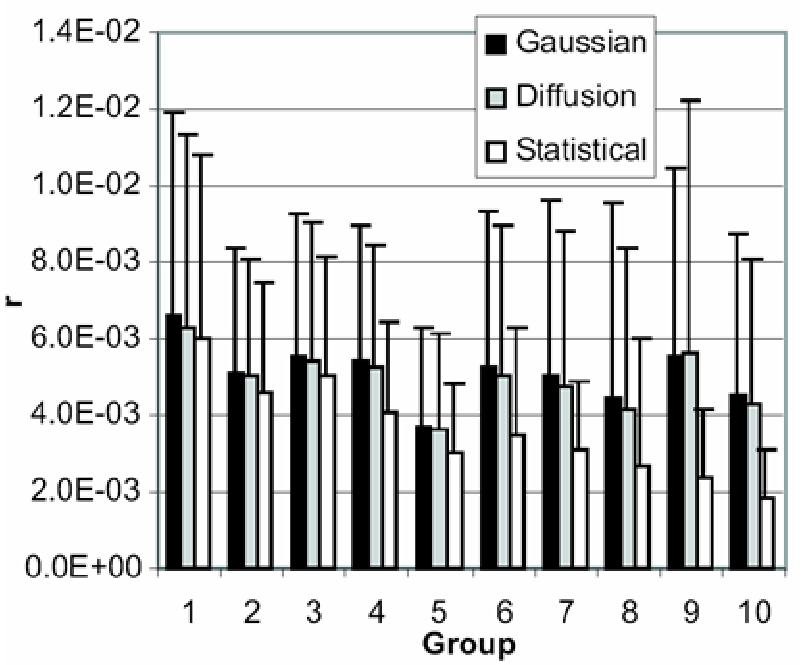


Figure 20 : Résidus de l'ajustement de chacun des modèles pour les taches protéiques de tailles croissantes (chacun des 10 groupes contient 10% des taches, et la taille de ces taches croît avec le numéro du groupe)

La Figure 20 résume le résultat de la comparaison des modèles et met en évidence la supériorité du modèle de distribution de points.

IV.3.3.2 Détections des taches protéiques

Tous les logiciels dédiés à l'analyse des images de gels d'électrophorèse bidimensionnelle proposent une solution pour la détection des taches protéiques. Cette problématique de segmentation est bien connue dans le domaine du traitement d'image et elle en est une étape primordiale. À ce jour, il existe de nombreuses méthodes de segmentation, que l'on peut regrouper en quatre principales classes :

1. Segmentation basée sur les régions (en anglais : region-based segmentation). On y trouve par exemple : la croissance de région (en anglais : region-growing) et la décomposition/fusion (en anglais : split and merge)
2. Segmentation basée sur les contours (en anglais : edge-based segmentation)
3. Segmentation basée sur une approche globale de l'image (par exemple : seuillage (en anglais : thresholding), histogramme, approches basées sur le nuage couleur...)
4. Segmentation basée sur la coopération entre les trois premières segmentations

Toutes ne sont pas forcément adaptées aux images de gels d'électrophorèse bidimensionnelle. Les méthodes les plus couramment utilisées pour les gels d'électrophorèse sont basées sur la technique du laplacien du gaussien (« LoG » pour Laplacian of Gaussian) [22], qui consiste à retenir les intersections de la dérivée se-

conde de l'image comme les centres des taches protéiques et d'établir une première segmentation en considérant les zones où elle est négative. La sensibilité au bruit de la dérivée seconde rend nécessaire le préfiltrage de l'image par un noyau gaussien. La première segmentation est ensuite étendue, par croissance de région, aux pixels voisins répondant à des règles empiriques basées sur la valeur de l'intensité et de la dérivée seconde. La détection proposée par le logiciel ImageMaster™ [5], fait partie de cette catégorie de méthode basée sur le laplacien et les dérivées secondes des intensités des pixels. Le principe de cette détection peut être résumé en considérant un point $\vec{P}$ du plan image, dont l'intensité est $I(\vec{P})$. Il appartient à une tache protéique si :

$$\min\left(\frac{\partial^2}{\partial x^2} I(\vec{P}) - R, \frac{\partial^2}{\partial y^2} I(\vec{P}) - C\right) > 0$$

et

$$-\Delta I(\vec{P}) - L \geq 0$$

Avec $\Delta I(\vec{P})$ le laplacien et $\mathrm{L, R\ et\ C}$ des constantes positives. Pour les pixels saturés, c'est-à-dire lorsque l'intensité dépasse un seuil T donné, la condition devient :

$$\min\left(\frac{\partial^2}{\partial x^2} I(\vec{P}), \frac{\partial^2}{\partial y^2} I(\vec{P})\right) > 0$$

L'inconvénient majeur de ce type de méthodes est qu'elles ne tiennent pas compte de la forme du contour des taches protéiques et peuvent conduire à des identifications erronées. Pour pallier à ce problème, de nouvelles méthodes ont été développées comme celle proposée par Takahashi et al. [54] qui utilisent un opérateur annulaire (« ring operator »). Il s'agit d'un opérateur permettant de n'identifier que les maximum locaux caractéristiques des sommets des taches protéiques en imposant un critère de circularité.

Dans le cadre d'une étude sur l'analyse automatique des gels d'électrophorèse bidimensionnelle, Conradsen et Pedersen [14] présentent une méthode de détection d'un motif dans une image largement inspirée de la morphologie mathématique. La méthode consiste en l'application successive de filtres médians, de filtres « maximums locaux », de détection de contour et d'érosions binaires. Le filtrage médian permet la suppression du bruit tandis que le filtrage du « maximum local » correspond à une érosion morphologique des niveaux de gris. La détection des contours est basée sur la dérivée seconde de l'image. Le résultat de la détection des contours est une image binaire sur laquelle on applique une érosion afin de supprimer les bruits et de distinguer certaines taches protéiques se chevauchant. Comme le montre le schéma de la Figure 21, toutes ces opérations sont répétées à plusieurs niveaux de résolution afin de détecter les taches protéiques de toutes tailles : les voisinages 3x3, 5x5, 7x7 et 9x9 sont ainsi successivement utilisés.

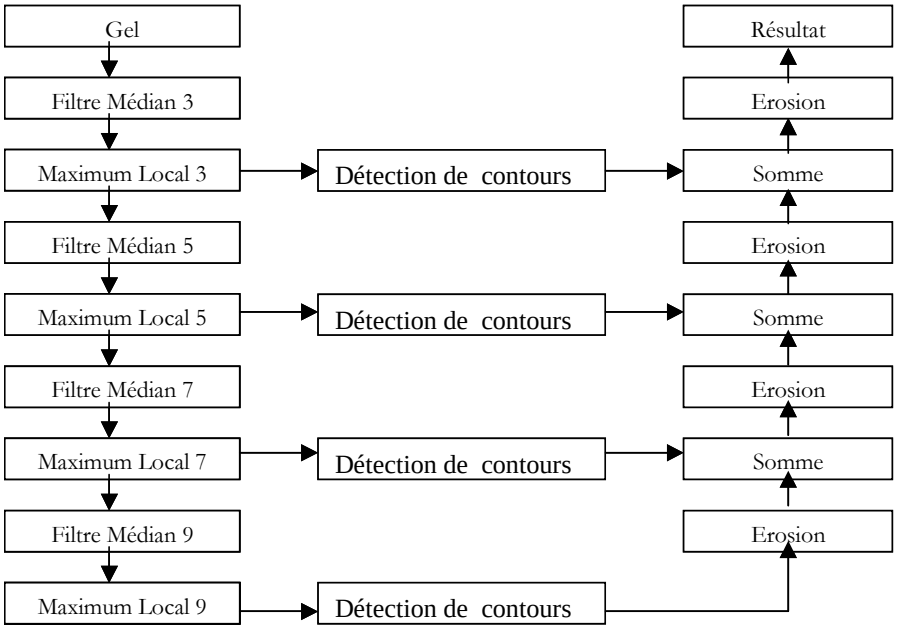


Figure 21: Principe de la méthode de détection mise en place par K. Conradsen et J. Pedersen [14].

Les résultats obtenus sur des gels d'électrophorèse et présentés par Conradsen et Pedersen [14] dans leur article ne sont pas entièrement satisfaisants car certaines petites taches protéiques ne sont pas pris en compte à cause de l'érosion binaire de leur noyau et la méthode est mise en défaut par certains chevauchements.

Un des algorithmes les plus simples pour l'extraction des taches protéiques se base sur la transformée « H-domes » [57]. Cette transformation, illustrée par la Figure 22, permet d'extraire les parties claires (ou foncées en travaillant sur l'image complémentaire) de manière indépendante des critères de forme et d'échelle. Les régions extraites répondent aux deux conditions suivantes :

- chaque pixel du dôme a son niveau de gris supérieur aux pixels immédiatement voisins du dôme,
- la différence maximale de niveau de gris entre deux pixels quelconque du dôme doit être inférieure ou égale à une constante h (définie par la suite).

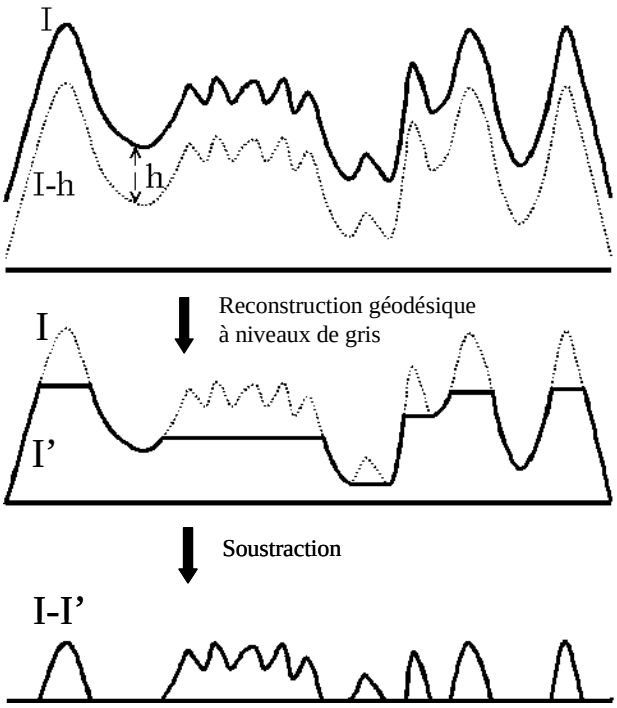


Figure 22: Illustration de la méthode "h-domes" pour l'extraction des taches protéiques visualisée sur le profil de niveau de gris

Le résultat $M_h(I)$ de la transformation "h-dome" peut être définie de la manière suivante : $M_h(I) = I - \rho_1(I-h)$, où I est l'image originale et où $(I-h)$ représente le résultat de la soustraction de l'image originale par une constante h choisie de manière adaptée. Le choix de h est lié au contraste de l'image de gel et à la volonté de l'utilisateur d'extraire plus ou moins les taches protéiques peu visibles. $\rho_1(I-h)$ est la reconstruction morphologique de l'image originale avec $(I-h)$ pris comme marqueur et I comme masque de reconstruction. Pour davantage de détails et une approche plus théorique, il sera utile de se reporter aux travaux de Vincent et Soille [57]. Une fois la transformation « h-domes » effectuée, les taches protéiques sont identifiées par seuillage et le résultat doit être encore affiné en jouant sur les paramètres de forme des objets binaires ou en appliquant un traitement complémentaire pouvant être une modélisation. En effet, cette méthode ne tient pas compte de la géométrie des taches protéiques (le contour d'une tache protéique est grossièrement circulaire) et elle conduit à l'obtention d'agrégats de taches protéiques mal séparées.

La méthode de la ligne de partage des eaux fournit une solution très populaire. Cette méthode est un outil de la morphologie mathématique introduit par S. Beucher et C. Lantuejoul [10] à la fin des années 70. Elle présente l'avantage de s'affranchir des variations basses fréquences du fond lors de la détection de contours. Comme l'illustre la Figure 23, elle consiste en la recherche des lignes de partage des eaux sur le gradient de l'image originale. Le gradient est un opérateur mathématique permettant de quantifier la variation de l'intensité dans le plan de l'image. L'opérateur gradient permet ainsi de mettre en évidence les contours des taches qui correspondent à des zones de forte variation d'intensité. Cependant, cette méthode conduit souvent à une sur-segmentation (Figure 24) si elle n'est pas contrôlée par des

marqueurs (Figure 25). Ces marqueurs sont des régions de l'images considérées comme faisant partie des taches protéiques à identifier. Le problème de la détection des taches et de leurs contours se voit alors transformé en la détection de ces marqueurs. En couplant cette technique à une modélisation des taches protéiques par diffusion, E. Bettens et al. [9] obtiennent des résultats satisfaisants sur des images de gel d'électrophorèse bidimensionnelle. Il faut noter que Vincent et Soille [57] ont développé des algorithmes rapides de ligne de partage des eaux.

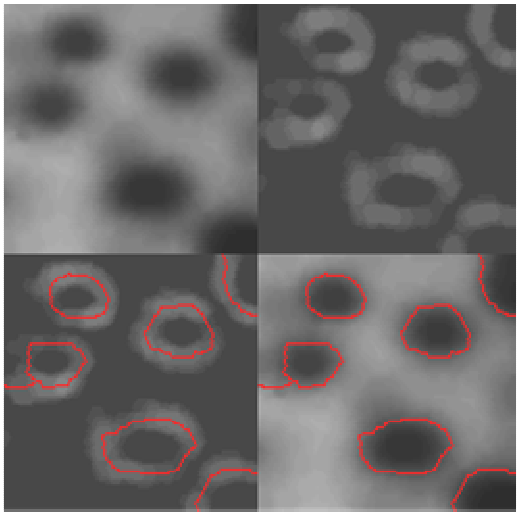


Figure 23: Détection des taches protéiques par la technique de ligne de partage des eaux. De gauche à droite et de bas en haut : l'image originale, le gradient de l'image, la ligne de partage des eaux sur le gradient et enfin, le résultat sur l'image originale.

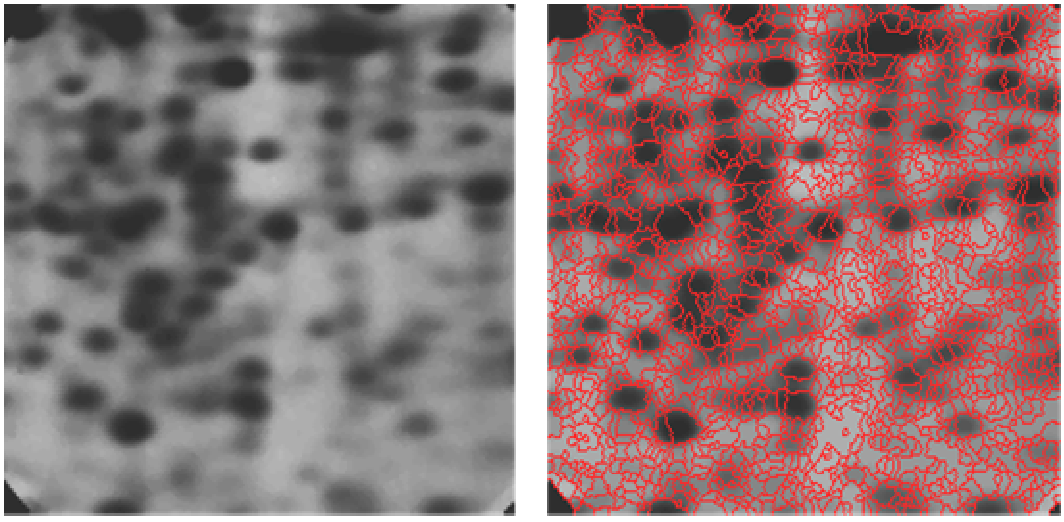


Figure 24: Problème de sur-segmentation dû aux bruits et aux irrégularités locales du gradient de l'image.

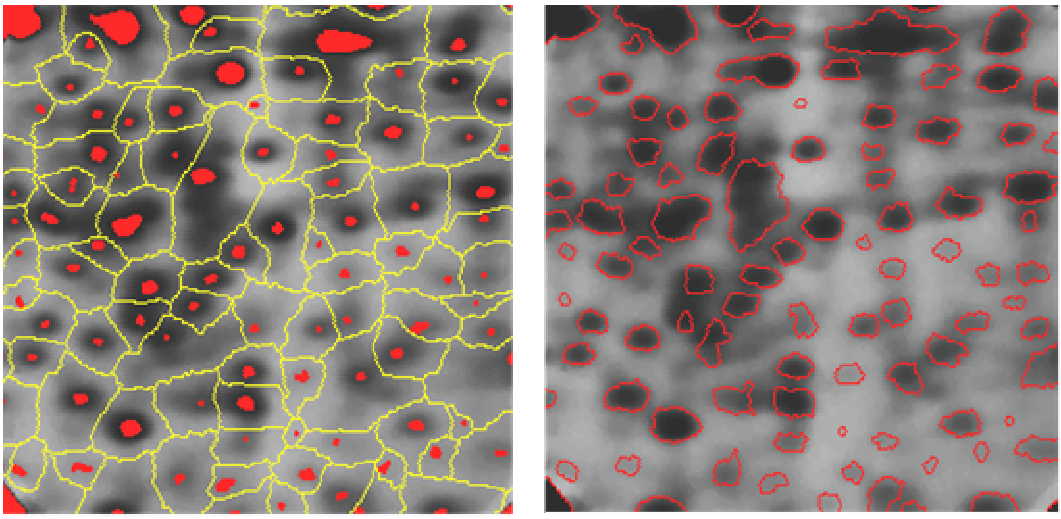


Figure 25: Technique de segmentation par ligne de partage des eaux controlée par des marqueurs (en rouge sur l'image de gauche). Comme le montre l'image de droite, cette technique permet de s'affranchir du problème de sur-segmentation.

Par ailleurs, une façon de séparer les traînées des pics pourrait être d'utiliser des bases de fonctions telles que les « ridgelets » ([12], Candes et Donoho, 1999) qui sont dédiées à la modélisation des objets linéaires sur une image.

En 1999, Klaus-Peter Pleißner [43] propose de nouveaux algorithmes pour la détection et la mise en correspondance des taches protéiques de gel d'électrophorèse . Il a développé une technique de ligne de partage des eaux hiérarchisée avec extraction de motifs qui permet des fusions partielles de taches protéiques, évitant ainsi la sur-segmentation. L'approche est résumée dans la Figure 26.

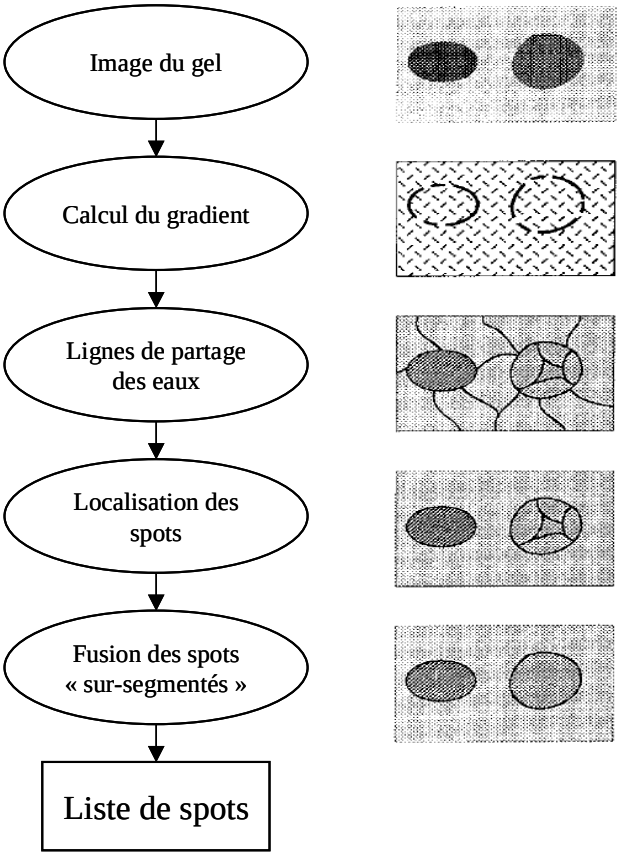


Figure 26: Les étapes de la détection des taches protéiques par ligne de partage des eaux hiérarchisée avec marqueurs.

Il faut cependant souligner que cette méthode est surtout efficace pour la phase segmentation, c'est à dire qu'elle suppose d'avoir effectué, au préalable, l'identifications de marqueurs des taches protéiques, tels que définis plus haut.

Certaines méthodes supposent que la forme des taches protéiques est à peu près constante : contour elliptique [27], densité gaussienne [5]. De telles suppositions ne sont pas systématiquement vérifiées dans la pratique. Les irrégularités proviennent notamment de la fusion de taches protéiques, de la superposition du bruit et de la saturation.

Les méthodes fondées sur la modélisation des taches protéiques font souvent suite aux segmentations présentées précédemment de manière à les affiner. En effet, les méthodes de modélisation permettent la reconstruction des formes altérées par le traitement, la détection des taches protéiques au sein des clusters mais aussi la quantification de ces taches. Parmi ces méthodes, on peut notamment citer celles qui correspondent au modèle gaussien et au modèle par diffusion qui consistent à identifier les formes en tant que volumes. Idéalement, chaque tache est une bi-gaussienne. En réalité, les taches ont une forme de goutte qui peut nécessiter une modélisation par 4 gaussiennes. On cherche la fonction gaussienne (ou bi-gaussienne ou tetra-gaussienne) qui modélise au mieux la tache. La méthode est similaire pour le modèle par diffusion.

Au sein des groupements de taches protéiques, la géométrie peut être complexe et il peut y avoir saturation du niveau de gris, ces méthodes ne sont alors plus justifiées et deviennent inefficaces. Efrat et al [18] ont proposé une solution fondée sur la seule vision planaire de l'image, c'est-à-dire sans tenir compte des niveaux de gris et donc moins sensible au phénomène de saturation. Comme l'illustre la Figure 27, cette solution consiste à modéliser les regroupements à l'aide d'ellipses.

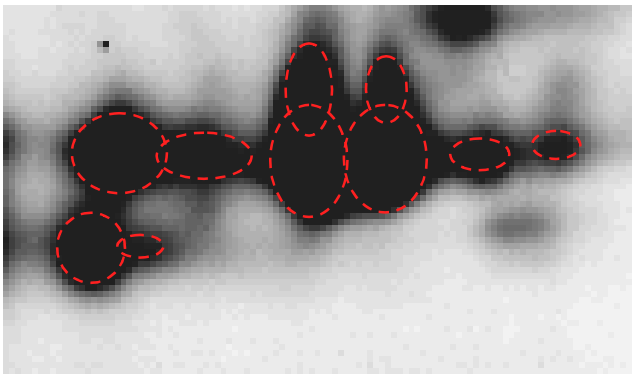


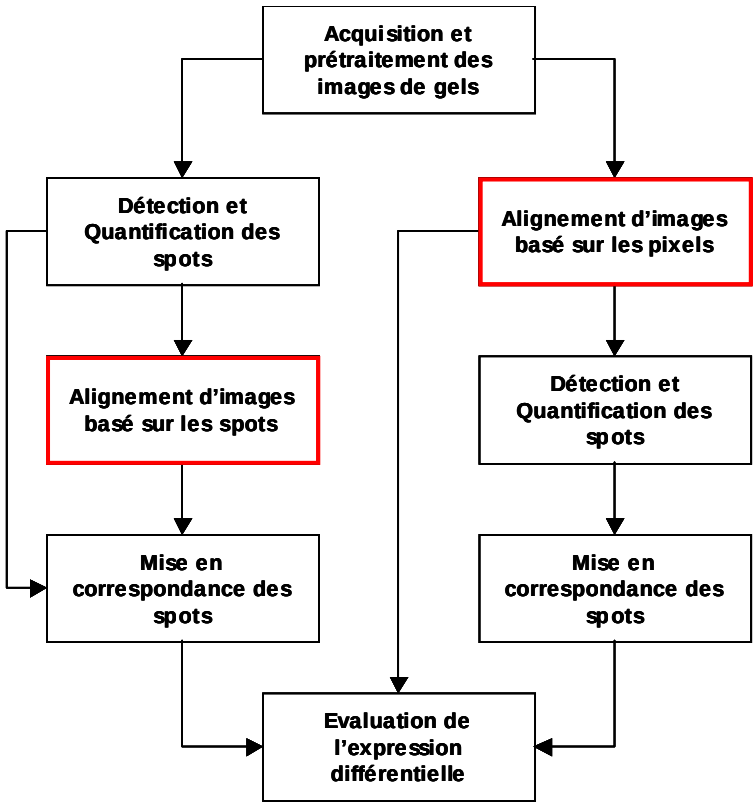
Figure 27: Modélisations d'un groupement de taches protéiques à l'aide d'ellipses (inter-gels et al., [31])

Une méthode plus souple, proposée par Mike Rogers et al. [48] consiste quant à elle, en l'utilisation d'un modèle statistique d'une tache protéique, basé sur un modèle de distribution de points (PDM). Il s'agit d'une représentation de la variation du profil d'un objet au sein d'une classe d'objets du même type. Elle est basée sur les statistiques issues d'un apprentissage sur une population représentative. Il est en effet possible d'obtenir la forme « moyenne » d'une tache protéique et ses principaux modes de variation grâce à une analyse en composantes principales à partir d'une population de taches protéiques représentatives. Ce modèle permet de décrire la forme d'une tache protéique de manière efficace avec un minimum de paramètres.

Figure 28: La modélisation PDM basée sur l'apprentissage offre de meilleurs résultats que les modélisations classiques

Comme le montre la Figure 28, cette dernière méthode, un peu plus lourde à mettre en place, permet cependant une meilleure modélisation que celles issues de la méthode gaussienne ou par diffusion. Il faut toutefois noter le risque de surmodélisation lié à ce type d'approche pouvant conduire, par exemple, à considérer une partie des traînées comme faisant partie d'une tache protéique.

IV.3.3.3 Mise en correspondance inter-images des taches protéiques



L'étape de l'alignement des gels se révèle indispensable pour la comparaison des gels et donc pour la mise en correspondance des spots. Elle consiste en une relocalisation de chaque pixel d'un gel et a pour but de rendre les gels géométriquement comparables deux à deux. Après l'alignement et dans l'idéal, le spot correspondant à une protéine donnée se situera aux mêmes coordonnées sur toutes les images de gel. Cette transformation géométrique, illustrée par la Figure 29, se base généralement sur un ensemble de spots connus et localisés manuellement sur les gels concernés. Ce sont les « landmarks ».

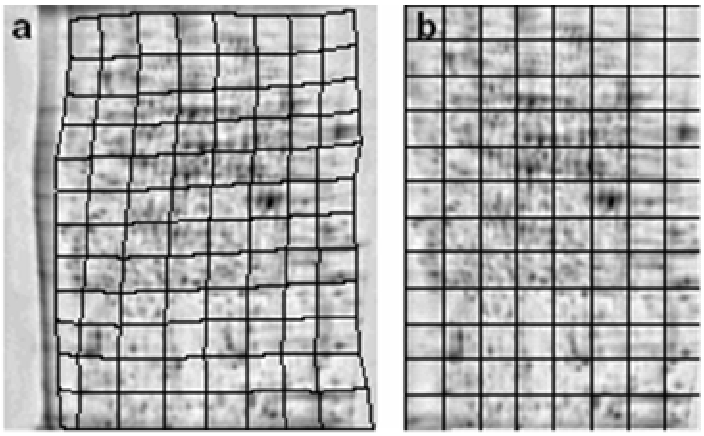


Figure 29: Alignement d'un gel (l'image a correspond au gel avant déformation, l'image b est le résultat de l'alignement avec un gel de référence)

Il est indispensable de pouvoir comparer des spots dans une série de gels. En effet, on s'intéresse souvent aux changements du profil d'une protéine sous diverses conditions expérimentales pour trouver les spots identiques et ceux qui diffèrent. L'étape de l'alignement des gels ne suffit pas forcément à superposer parfaitement les gels entre eux du fait de leur grande variabilité. C'est pour cela qu'il est parfois utile de distinguer l'alignement des gels et la mise en correspondance des spots. Les causes les plus courantes de la variabilité des gels sont :

- les différences dans la préparation des échantillons,
- les variations physiques et chimiques du gel,
- les variations des conditions expérimentales,
- la mobilité inégale des protéines suivant la région du gel.

De nombreux algorithmes sont disponibles pour le « gel matching », terme qui, ici, regroupe la notion d'alignement des gels et de mise en correspondance des spots. La plupart d'entre eux est basée sur les approches suivantes :

- géométrique,
- statistique,
- théorie des graphes et
- reconnaissance de motifs.

Dans ce paragraphe, après avoir défini précisément la notion de mise en correspondance des spots (« matching »), les techniques les plus courantes ayant été utilisées pour l'électrophorèse bidimensionnelle seront présentées. Pour chacune, il est conseillé au lecteur intéressé, de se référer à la littérature.

Par la suite, I_1 et I_2 désigneront les deux images de gels concernées par le matching, (x, y) seront les coordonnées d'un point de l'image I_1, (u, v) seront celles d'un point de I_2 et $\phi : I_1 \rightarrow I_2$ sera la transformation considérée.

3.3.3.1 Définition de la mise en correspondance des taches protéiques (matching)

Comme l'illustre la Figure 30, la mise en correspondance des taches protéiques consiste donc à définir des paires (entre 2 gels) ou des groupes (entre plus de 2 gels) de taches.

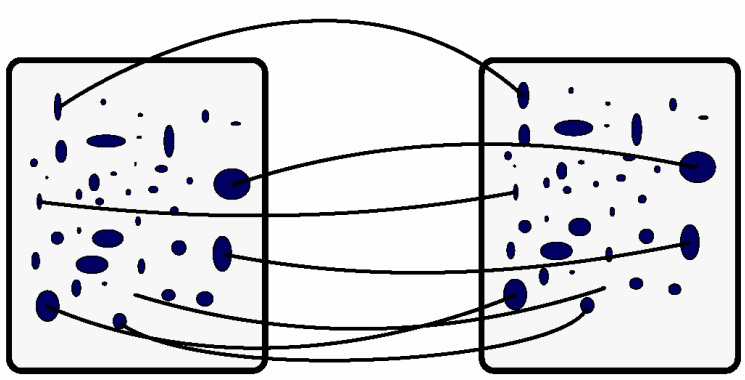


Figure 30: Illustration de la mise en correspondance des spots de deux gels différents (matching)

a) Entre deux gels

Soient deux gels g1 et g2, $S1 = (S1_1, S1_2, ..., S1_n)$, l'ensemble des taches de g1 et $S2 = (S2_1, S2_2, ..., S2_m)$, l'ensemble des taches de g2. Le matching revient à trouver l'ensemble des paires $P = (S1_i, S2_j)$ tel que $F(S1_i) = F(S1_j)$, F étant une fonction qui associe une tache au nom de la protéine correspondante.

Le matching peut se faire par nom (si les spots sont identifiés), par pattern (analyse des constellations de spots qui se retrouvent dans les deux gels), ou par d'autres méthodes que nous verrons plus loin.

b) Au sein d'un groupe de gels

Ce cas de figure correspond en fait à la mise en correspondance d'un gel de composition et un gel de référence. Un gel de composition est un gel théorique, créé artificiellement et sur lequel figurent toutes les taches protéiques apparaissant au moins une fois dans l'ensemble des gels qu'il représente.

Soit un gel de composition $G = (g_1, g_2, ...g_k)$ de k gels et un gel de référence g_r. Alors, un ensemble de spots $T = (S1, S2, .., Su)$ est un groupe :

- si chaque $Si_{1 \le i \le u}$ est un spot d'un des gels G,
- si Sr est un spot de g_r,
- et si $F(S1) = F(S2) ... = F(Sr)$.

3.3.3.2 La transformation affine

La transformation affine [21] est une des techniques d'alignement les plus simples. Elle prend la forme suivante :

$u = a_{10}x + a_{01}y + a_{00}$

$v = b_{10}x + b_{01}y + b_{00}$

L'idée est de choisir un ensemble de paires initiales sélectionnées soit de manière manuelle soit de manière statistique puis de calculer la transformation affine qui minimise la distance entre les points ayant été associés deux à deux. Ensuite cette procédure est répétée en partant à chaque fois des gels transformés et d'un ensemble de points choisis de manière statistique sur ces gels. L'algorithme prend fin lorsqu'un

certain critère de qualité est atteint. Cette méthode est peu adaptée à la complexité des gels d'électrophorèse dont les biais, précédemment présentés, sont nombreux.

3.3.3.3 La transformation polynomiale

La transformation polynomiale [21] prend la forme suivante :

$$u = \sum_{i=0}^{p}\sum_{j=0}^{q} a_{ij} x^i y^j$$

$$v = \sum_{i=0}^{p}\sum_{j=0}^{q} b_{ij} x^i y^j$$

avec p non nécessairement égal à q.

L'approche est ensuite identique à celle de la transformation affine. Traditionnellement [21], l'utilisation de fonctions polynomiales alliée à une minimisation des moindres carrés suffit à définir l'image résultant de l'alignement. Cependant, une telle modélisation ne permet pas de rendre compte des distorsions complexes inhérentes à la technique de l'électrophorèse.

3.3.3.4 Splines de type plaque mince (Thin-Plate Splines)

Il s'agit d'une technique de minimisation d'énergie [13] qui permet de faire un compromis entre la distance entre les points à mettre en correspondance et l'aspect harmonieux du résultat sachant que d'autres techniques peuvent conduire à des déformations induisant des discontinuités d'ordre 1 et 2. Soient $\{(x_i, y_i) | i = 1...k\}$ et $\{(u_i, v_i) | i = 1...k\}$ les ensembles de spots des images I_1 et I_2. Il s'agit alors de déterminer parmi une famille de fonction, la fonction f qui minimise l'énergie E_{TPS} définie par :

$$E_{TPS}(f) = \left(\sum_{i=1}^{k} (x_i - u_i)^2 + (y_i - v_i)^2 \right) + \lambda \iint \left(\left(\frac{\partial^2 f}{\partial x^2} \right)^2 + 2\left(\frac{\partial^2 f}{\partial x \partial y} \right)^2 + \left(\frac{\partial^2 f}{\partial y^2} \right)^2 \right) dxdy$$

Le paramètre λ permet de contrôler le degré de « douceur » du résultat : plus λ est grand, plus la transformation est respectueuse de la continuité de l'image résultat, mais moins le recalage point à point est efficace.

3.3.3.5 Modélisation physicochimique

La méthode développée par John S. Gustafsson [25] tient compte de l'origine physique de la déformation d'un gel. En effet, la principale cause de déformation est liée au phénomène de fuite de courant lors de la phase de deuxième dimension de l'électrophorèse et peut donc se modéliser. La première étape de la méthode, illustrée par la Figure 31, consiste donc à utiliser le modèle afin de corriger cette déformation.

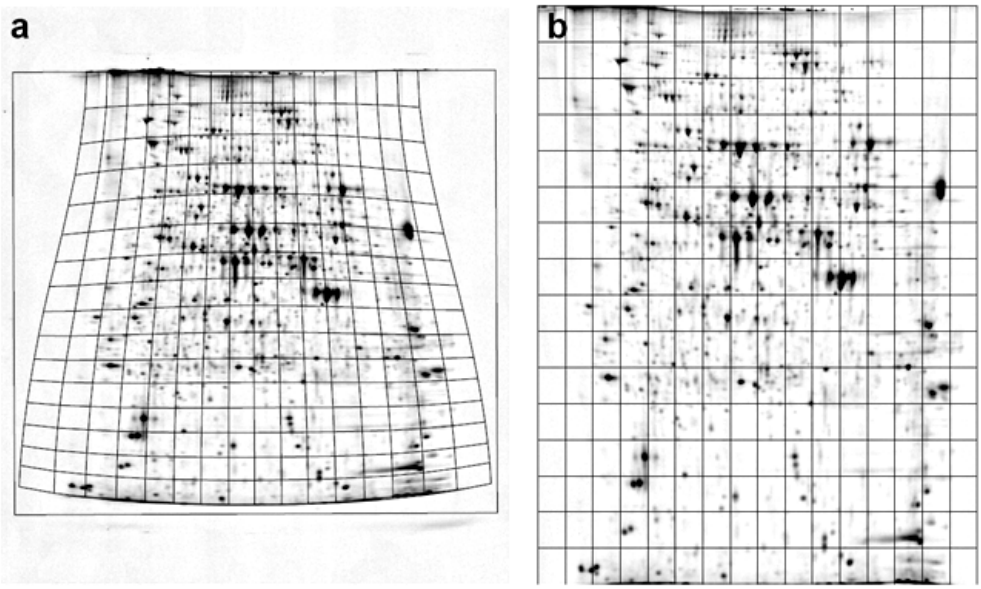


Figure 31: Étape 1. Correction de la déformation liée au phénomène de fuite de courant.

Lors de la deuxième étape, Les images des gels sont alignées à une référence grâce à une méthode classique de maximisation d'un critère de ressemblance (Figure 32).

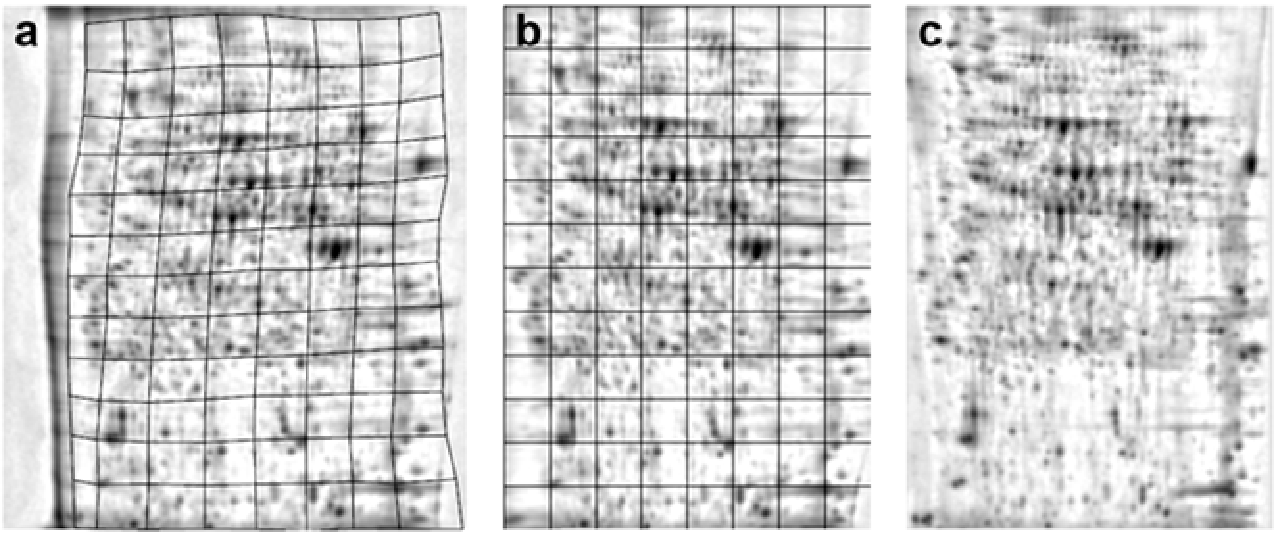


Figure 32: Étape 2. Alignement de l'image a) à l'image de référence c). L'image b) représente la version alignée de l'image a). La grille de déformation est représentée en surimpression sur l'image a).

Pour l'évaluation de sa méthode John S. Gustafsson [25] compare la similarité de la distribution des taches protéiques sur trois jeux d'images : les images originales, les images après l'étape 1 et les images après les étapes 1 et 2. Pour cela, il s'intéresse à la distribution de la longueur des vecteurs de déplacement des taches protéiques. Le résultat, présenté sur la Figure 33 montre que l'étape 1 n'améliore pas significativement la similarité entre les images.

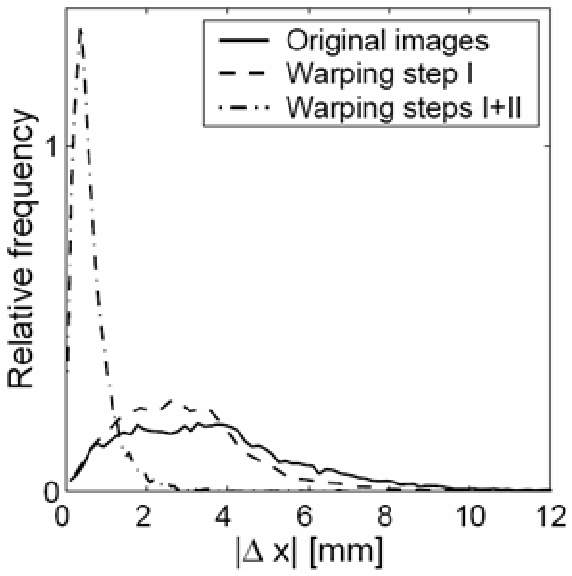


Figure 33: Comparaison de la similarité de la distribution des taches protéiques pour les images originales et sur les images après l'étape 1 et après les étapes 1 et 2. Le graphe représente la distribution de la longueur des vecteurs de déplacement des taches protéiques pour chacun des trois jeux d'images.

Cette approche apporte donc une légère amélioration dans l'optique de l'amélioration de la mise en correspondance des taches protéiques. Cependant, cette amélioration est minime et devient même dispensable lorsque l'on travaille sur des images issues de la technologie DIGE puisqu'elle permet de s'affranchir d'une grande partie des problèmes d'alignement.

3.3.3.6 Approche basée sur les graphes de voisinage

Les graphes de voisinages [18] [23] sont une classe de graphes planaires qui connecte des points à d'autres points voisins en respectant un certain nombre de règles. Leur utilité pour l'alignement d'image est basée sur l'idée que ces lois de connexion reflètent les propriétés locales de l'image et qu'elles peuvent être utilisées pour la mise en correspondance entre des points appariés de deux images (les « landmarks »). Ceci évite le calcul explicite d'une fonction de déformation.

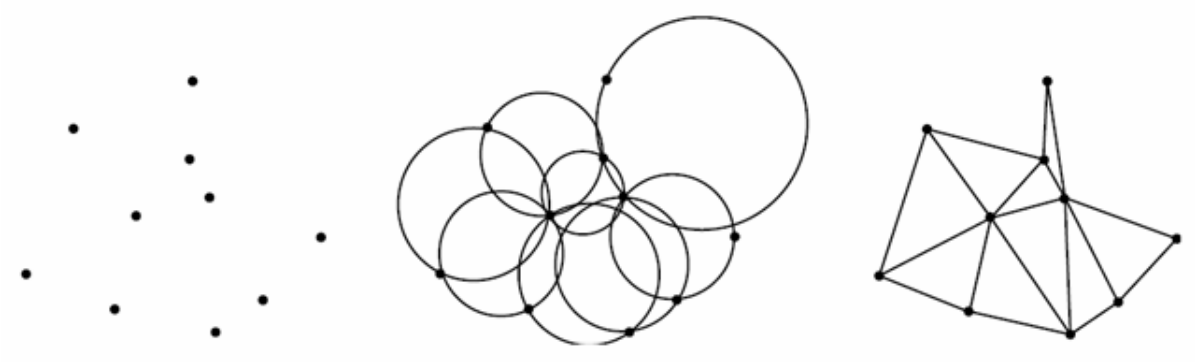


Figure 34: Les cercles (milieu) et la triangulation (droite) de Delaunay associés à un ensemble de points (gauche)

La Figure 34 présente l'exemple du graphe de Delaunay dont la définition, donnée pour l'exemple, est :

Soit un ensemble de points $P=\{p_1,...,p_n\}$ dans R^2. Définir une relation $\Re$ sur les triplets $\{a,b,c\}\subset P$ telle que $\{a,b,c\}\in\Re \Leftrightarrow$ Il n'y a pas de point à l'intérieur de l'unique cercle C passant par les points a, b et c.

On définit ainsi le graphe G de Delaunay tel que $G=(V,E)$, avec $V=P$ et $E=\{(a,b)|\exists c \quad tq \quad \{a,b,c\}\in\Re\}$.

Cette approche a notamment été exploitée par Efrat et al et implémentée dans le logiciel d'analyse d'image de gels d'électrophorèse CAROL [18]. Cette implémentation se base sur les listes des taches protéiques de chaque gel, avec, associée à chacune des taches, sa surface normalisée par sa valeur d'intensité. Ensuite, il s'agit de trouver la mise en correspondance entre les listes qui concerne le maximum de taches tout en respectant la géométrie de l'image et la cohérence des intensités relatives. Le critère géométrique est basé sur la similarité des segments imaginaires reliant les taches, c'est-à-dire sur la similarité des graphes (obtenus par triangulation de Delaunay) associés.

Une fois optimisés, ces algorithmes fournissent une solution efficace pour la mise en correspondance inter gels des taches protéiques, et ils ont d'ailleurs été adoptés par de nombreux logiciels dédiés. Nous verrons cependant par la suite que, dans le cas de la technologie DIGE, il est possible de s'affranchir de cette étape de mise en correspondance.

3.3.3.7 Approche statistique

Cette partie, basée sur les travaux de R. D. Appel et al. [5], se focalise sur l'approche statistique implémentée dans le logiciel Mélanie II. Le matching entre deux images est accompli selon plusieurs étapes.

Tout d'abord, pour chaque tache protéique d'un gel, on constitue une liste des taches voisines. On associe donc une tache centrale, dite « primaire », à une liste de taches voisines dites « secondaires » (voir Figure 35). Le tout forme un groupe dit « cluster ». Une tache protéique fait partie d'un groupe si son centre est à l'intérieur d'un cercle de rayon fixé, centré sur la tache primaire. Le rayon est estimé en consi-

dérant la dimension de l'image, le nombre total de taches protéiques et le nombre minimum de taches par groupe. Les groupes sont décrits en coordonnées polaires et sont établis avant la mise en correspondance afin d'être comparés rapidement.

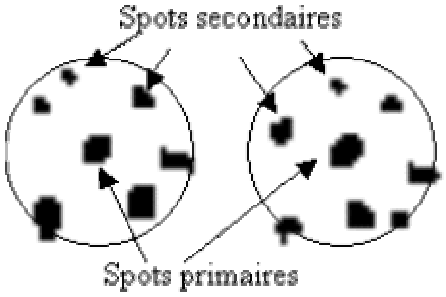


Figure 35: Illustration de la manière dont sont considérées les taches protéiques ("spots") à l'intérieur de deux régions issues de deux gels à mettre en correspondance.

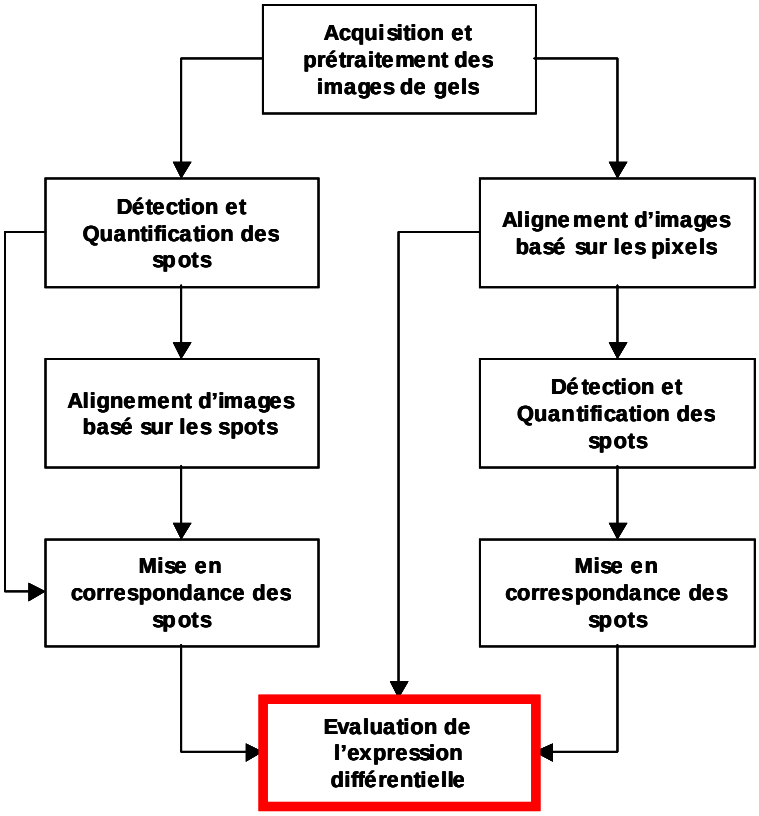
La procédure commence alors par mettre en correspondance les clusters correspondant à des paires déjà créées (une paire correspond à un couple de taches primaires).

Cette mise en correspondance est effectuée par mesure de similarité des constellations formées par les taches secondaires. Cette méthode qui tient compte de l'aire des taches et de leur disposition est basée sur une méthode probabiliste dont l'aspect théorique est détaillé dans l'article de R. D. Appel et al. [5]

Les taches secondaires qui ont pu être appariées, deviennent à leur tour des taches primaires. La méthode progresse ainsi de proche en proche.

Dans certaines situations la comparaison des groupes associés aux spots primaires peut conduire à des erreurs d'appariement identifiées suivant une méthode détaillée dans l'article. Dans ces cas là, les groupes associés aux taches secondaires sont comparés. Ainsi, davantage de taches protéiques sont prises en considération.

IV.3.4 Analyse des données pour l'évaluation de l'expression différentielle



Cette dernière étape est commune à tous les processus envisagés. Elle consiste en l'exploitation des données brutes et doit aboutir à l'identification de marqueurs potentiels du cancer colorectal.

La nature et la structure des données brutes varient selon les méthodes utilisées au préalable pour l'analyse et selon le schéma expérimental adopté. Les données peuvent être complètes (c'est-à-dire sans données manquantes) et concernent alors les groupes de polypeptides pour chacun desquels on a accès à une quantification. Le schéma expérimental impacte également la structure et la façon d'aborder l'interprétation des données. Certains schémas expérimentaux comportent ainsi des répétitions qui permettent l'évaluation d'un ou de plusieurs types de variabilité (variabilité liée au marquage, à la migration ou encore celle liée aux erreurs de mise en correspondance des spots).

Grâce à la technologie DIGE, les variations d'intensités des spots dues aux facteurs expérimentaux spécifiques au gel comme les pertes protéiques, lors de la pénétration des protéines dans les IPG strips, seront identiques pour chaque échantillon au sein d'un même gel. C'est-à-dire que la quantité relative d'une protéine d'un échantillon par rapport à un autre, dans le même gel, ne sera pas affectée.

IV.3.4.1 Normalisation

Le problème de la normalisation des données produites lors des expériences DIGE est souvent abordé dans la littérature, mais n'a pas été autant exploré que pour les données issues des expériences avec des puces à ADN. La nature et la structure des données issues de ces deux technologies sont similaires et il est intéressant de faire le rapprochement, particulièrement en ce qui concerne la normalisation. Nous considérerons, d'ailleurs, la définition la plus répandue dans le domaine des puces à ADN, qui attribue à la normalisation la mission de supprimer les biais systématiques qui masquent les informations biologiques pertinentes.

Au premier rang de ces biais figurent ceux lié aux particularités du gel, de la migration et de l'acquisition. Viennent ensuite les biais liés à l'usage de fluorophores différents nécessaires au multiplexage. Par la suite nous parlerons de « canaux » pour désigner les différentes images issues d'un même gel et correspondant aux différents fluorophores utilisés. Dans le cas de la DIGE, nous considérerons donc trois canaux correspondants aux fluorophores : Cy2, Cy3 et Cy5.

Il existe deux grandes familles non exclusives de méthodes de normalisation. La première famille se base sur les propriétés supposées et inhérentes à chaque jeu de données. Il s'agit d'une normalisation « interne », propre à chaque gel considéré. On impose une cohérence aux données en se basant sur des hypothèses ayant un sens biologique. L'hypothèse émise pour les jeux de données des puces à ADN est transposable pour les jeux de données DIGE : les moyennes d'intensité (de volume dans le cas de la DIGE) doivent être identiques entre deux canaux, c'est-à-dire que la majorité des gènes (des protéines dans le cas de la DIGE) s'exprime dans des quantités comparables, non différentielles, dans les deux échantillons comparés. Dans cette famille se classent la méthode globale, la régression linéaire ainsi que la correction LOWESS [62].

La deuxième famille concerne les méthodes basées sur des éléments de contrôle introduits dans l'expérience et qui correspondent souvent à un schéma expérimental particulier. Dans cette deuxième famille, on retrouve les schémas expérimentaux comportant une référence sur un des canaux ainsi que le « dye-swap » que nous définirons dans le paragraphe 4.3.1.2.

3.4.1.1 Normalisation par auto-cohérence (« self-consistency »)

La technologie DIGE permet une quantification, au sein d'un gel donné, de milliers (environ 2000 de façon courante) de protéines issus d'un échantillon. Au même titre que pour les données de microarray et au regard du grand nombre de valeurs considérées, les biologistes admettent que la majorité des protéines ne présentera pas de changement significatif d'expression [33] pour une étude comparative d'échantillons biologiquement proches comme par exemple les fragments pathologique et normal issus d'un même organe. Suivant cette hypothèse, les changements sur l'intensité globale ne peuvent être que la conséquence de variations non biologiques (biais) et l'on doit corriger les données de telle manière à ce que les distributions des ratios d'expression soient centrées sur la valeur 1. Généralement on emploie la valeur

médiane comme valeur estimée de la tendance centrale de la distribution et donc les données sont corrigées de manière à ce que

$$médiane_{i=1,\ldots,n}\left(\frac{V_{Cy3,i}}{V_{Cy5,i}}\right)=1$$

ce qui est équivalent à

$$\log_2\left(médiane_{i=1,\ldots,n}\left(\frac{V_{Cy3,i}}{V_{Cy5,i}}\right)\right)=0,$$

avec $V_{Cy3,i}$ le volume du spot i sur le canal $Cy3$ et $V_{Cy5,i}$ son volume sur le canal $Cy5$ pour un gel comprenant n paires de spots.

Réaliser cette normalisation consiste à estimer la transformation f adéquate telle que :

$$V_{Cy3}=f\left(V_{Cy5}\right)$$

Trois estimations de f sont généralement proposées dans la littérature et qui résultent d'une approche globale, par régression linéaire et par régression linéaire locale (LOWESS).

a) Méthode globale

La méthode globale est la plus simple à mettre en œuvre. C'est celle qui est la plus souvent proposée dans les outils informatiques dédiés à l'analyse des images issues de la DIGE, comme Decyder [33]. On parle de correction de l'échelle (« scale normalization ») car il s'agit de déterminer le facteur qui lie les deux canaux $Cy3$ et $Cy5$. Le choix le plus courant [62] pour la fonction f est :

$$V_{Cy5}\xrightarrow{f}médiane_{i=1,\ldots,n}\left(\frac{V_{Cy3,i}}{V_{Cy5,i}}\right)\times V_{Cy5}$$

Kreil et al. [33] ont montré que cette méthode n'est pas satisfaisante car le biais lié au fluorophore n'affecte pas seulement la dynamique des volumes des deux canaux mais entraîne aussi un décalage (« offset ») qu'ils jugent constants et proposent donc de lui préférer un modèle affine estimé par régression linéaire.

b) Régression linéaire

Le principe de cette méthode, proposée notamment par Kreil et al. [33], consiste à ajuster une droite affine $V_{Cy5}=a.V_{Cy3}+b$ à la figure de dispersion des volumes issus des deux canaux d'un même gel $\left(V_{Cy3},V_{Cy5}\right)$. Suivant l'hypothèse que la majorité des protéines ne présente pas de changement significatif d'expression, la droite de régression doit être linéaire et de pente égale à 1. Il s'en suit que l'on déduit la fonction f des paramètres a et b de la droite que l'on vient d'ajuster :

$$V_{Cy5}\xrightarrow{f}\frac{1}{a}\times V_{Cy5}-\frac{b}{a}$$

Il existe autant de variantes à cette méthode que de façons d'ajuster une droite à un nuage de point. On préfèrera les méthodes robustes par rapport aux valeurs extrêmes et aberrantes, courantes pour le type de données que l'on considère ici. Huber et al. [28] utilisent la régression des moindres carrés élagués itératifs (Least

Trimmed Squares) qui consiste en une minimisation de la somme des carrés des erreurs les « plus petites ».

c) Correction LOWESS

La correction LOWESS est largement utilisée pour les données de microarray (Yang et al. [62]). Elle permet de supprimer les effets dépendants de l'intensité. En effet, l'hypothèse que la majorité des protéines ne présente pas de changement significatif d'expression reste vraie quelle que soit la plage de volume considérée. La correction LOWESS se base donc sur la figure de dispersion (A,M), où

$$M = \log_2\left(\frac{V_{Cy3}}{V_{Cy5}}\right) \text{ et}$$

$$A = \frac{1}{2}\left(\log_2 V_{Cy3} + \log_2 V_{Cy5}\right)$$

Par ailleurs, il faut noter que cette visualisation est très pratique et très répandue. A est une image de la quantité moyenne des protéines et M est une manière de considérer le ratio des quantités à comparer de telle sorte que la distribution soit centrée sur 0 (quantités égales) et que les ratios double et moitié correspondent aux valeurs +1 et –1.

La fonction de correction est estimée, sur cette représentation, par régression polynomiale locale et pondérée sur un certain sous-ensemble de protéines au voisinage de chaque protéine i. Le voisinage étant défini comme des protéines dont les quantités moyennes sont proches. La fonction LOWESS que l'on note c attribue une valeur corrective à chaque protéine de telle sorte que pour la majorité des protéines :

$$M = \log_2\left(\frac{V_{Cy3}}{V_{Cy5}}\right) \cong c(A)$$

C'est à dire,

$$\log_2\left(\frac{V_{Cy3}}{V_{Cy5}.2^{c(A)}}\right) \cong 0$$

La correction LOWESS se traduit donc sur les quantités protéiques initiales de la manière suivante :

$$V_{Cy5,i} \xrightarrow{f} 2^{c(A_i)}.V_{Cy5,i}$$

Cependant, cet estimateur robuste ne prend pas en compte une réelle information biologique si ce n'est certaines hypothèses sur la distribution des quantités mesurées. Comme nous allons le voir, l'introduction d'éléments de contrôle au sein même de l'expérience peut apporter une information supplémentaire sur les biais systématiques et donc permettre de mieux les appréhender.

3.4.1.2 Deuxième famille : normalisation à partir d'éléments de contrôle inclus dans l'expérience

a) Référence interne

Le multiplexage apporté par la technologie DIGE améliore grandement la détection des changements du niveau d'expression entre les échantillons. En effet, les variations du volume des spots dues aux facteurs expérimentaux comme les pertes protéiques potentielles durant l'incorporation des échantillons dans les « strips », sont identiques pour chaque échantillon d'un même gel. Ainsi, les quantités relatives d'une protéine entre les échantillons d'un même gel restent inchangées alors que pour l'électrophorèse bidimensionnelle classique, les migrations réalisées séparément affectent différemment le volume des spots d'une même protéine. La technologie DIGE améliore donc la confiance sur la détection et la quantification des différences d'expression détectées au sein d'un même gel. Les différences d'expression observées au sein d'un même gel sont donc de vraies différences biologiques. En fait, seul le biais dû à la différence entre les fluorophores utilisés sur chacune des voies entre en compte.

Cependant cette amélioration n'est valable à condition de se cantonner à comparer les échantillons ayant migrés dans le même gel. Cette restriction peut être contournée en incluant dans chaque gel un échantillon standard sur un troisième canal, à l'aide du marquage en fluorescence par une troisième cyanine (Cy2). Basé sur ce principe, le schéma expérimental proposé par Alban et Al. [1] s'est largement imposé pour ce type d'analyse et est aujourd'hui pris en compte par les principaux logiciels du marché (ImageMaster 2D Platinium, Decyder…). Ce schéma consiste à utiliser comme standard, le mélange stoechiométrique de tous les échantillons de l'expérience. Le standard est qualifié d'interne puisqu'il est élaboré à partir des échantillons de l'expérience. La puissance de ce standard interne réside dans le fait que chaque protéine de chaque échantillon est représentée dans des quantités identiques sur la voie Cy2 de chaque gel. Ainsi, chaque protéine de chaque échantillon est censée (sauf défaut de détection) être présente dans le jeu de spots détectés sur ce standard interne et, comme l'illustre la Figure 36, servir de référence quantitative. Le standard interne permet ainsi de normaliser les mesures d'abondance protéique entre tous les gels de l'expérience. En outre, les images issues de ce standard présentent des patrons de spots très semblables pour tous les gels de l'expérience et donc facilitent et améliorent la mise en correspondance des spots inter-gels.

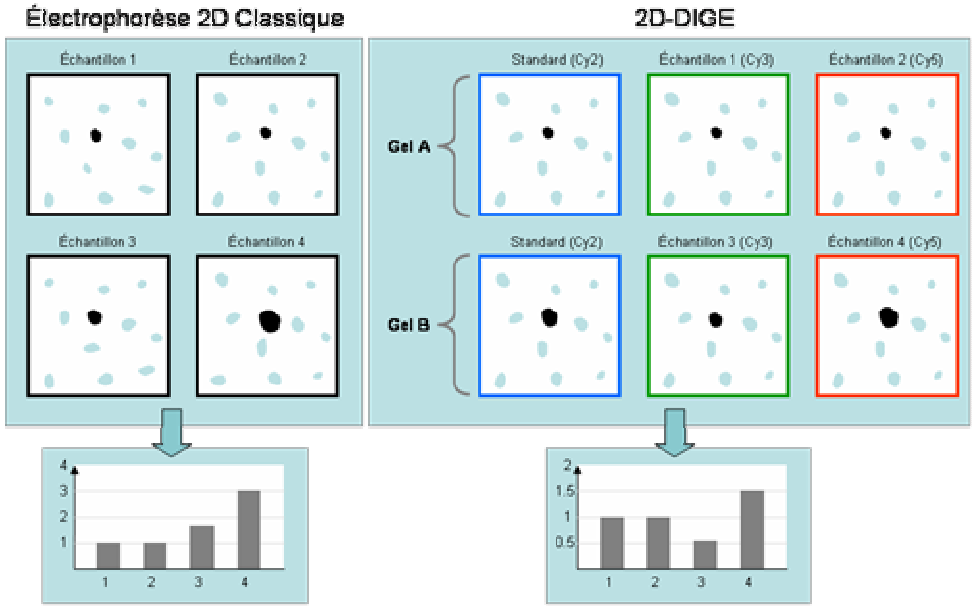


Figure 36 : Principe d'utilisation du standard interne dans une expérience DIGE. Dans cet exemple, le standard interne permet de constater que, sur le gel B, la quantification du spot en gras est surestimée par rapport au gel A. Ce biais est corrigé en rapportant l'abondance mesurée pour chaque échantillon à l'abondance mesurée sur le standard interne du gel correspondant. Comme le montrent les histogrammes des abondances mesurées, cette situation conduit à une erreur d'interprétation pour l'échantillon 3 dans le cas de l'électrophorèse bidimensionnelle classique.

Il faut cependant noter que la présence de ce standard interne peut se révéler superflue pour certains types d'analyse. Par exemple, dans le cadre d'analyse d'échantillons appariés, comme par exemple dans le cas de la comparaison de deux types de tissus issus du même patient, les quantités d'intérêt peuvent être les ratios protéiques et non pas directement les quantités protéiques. Or, à condition d'avoir préalablement supprimé les biais systématiques, les ratios protéiques réalisés entre deux mesures d'un même gel sont directement comparables et l'usage du standard interne comme élément de normalisation n'est pas nécessaire. C'est le cas dans notre projet de recherche de marqueurs protéiques potentiels où l'analyse différentielle concerne deux classes (tumeur et muqueuse) d'échantillons appariés patient par patient : la variable d'intérêt est le ratio observé pour chacun des patients.

b) Dye-Swap

Deux principales propriétés diffèrent d'un fluorophore à l'autre : le rendement quantique et l'affinité spécifique à chaque protéine. La différence de rendement quantique est un biais systématique quantifiable car il affecte toutes les protéines d'un échantillon de la même façon. Ce biais est supprimé lors des étapes de normalisation par auto-cohérence. L'affinité spécifique à chacune des protéines est, quant à elle, beaucoup plus difficilement quantifiable car elle peut être associée à un vrai changement d'expression de la protéine et donc se confondre. Si cette différence d'affinité est prépondérante devant la variabilité expérimentale, elle influence significativement l'estimation des quantités protéiques. Il est possible de tenir compte de ce phénomène en réalisant un schéma expérimental particulier, dit en « Dye-swap »,

qui consiste à répéter une fois chaque gel de l'expérience en inversant le marquage en fluorescence des deux classes d'intérêt.

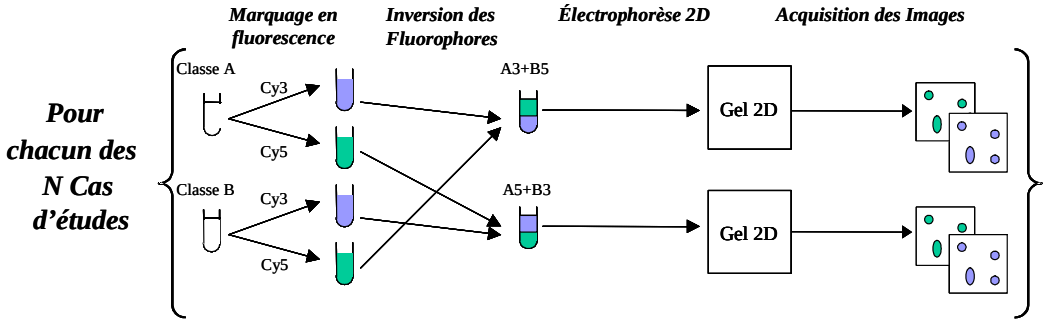


Figure 37 : Principe du Schéma expérimental en « Dye-Swap » . Pour chacun des cas d'étude on réalise deux gels. Ces deux gels ne se distinguent que par l'inversion du marquage en fluorescence sur les deux classes étudiées.

Afin de mieux comprendre le bénéfice apporté par un tel schéma expérimental, considérons une protéine i dont le niveau d'expression dans deux échantillons différents est connu. Nous noterons M et T (pour Muqueuse et Tumeur, deux classes très souvent comparées) les deux types d'échantillon biologique à comparer. Supposons que l'échantillon M soit marqué avec le fluorophore $Cy3$ et l'échantillon T avec $Cy5$. L'expression suivante permet alors la comparaison des deux valeurs d'expression :

$$M_i = \log_2\left(\frac{V_{Cy3,i}}{V_{Cy5,i}}\right)$$

De la même manière, à partir des mêmes échantillons M et T avec inversion du marquage, la comparaison peut également être faite grâce l'expression suivante :

$$M_i' = \log_2\left(\frac{V'_{Cy3,i}}{V'_{Cy5,i}}\right)$$

À partir de ces équations, il est possible d'écrire :

$$M_i = \log_2\left(\frac{V_{Cy3,i}}{V_{Cy5,i}}\right) = \log_2\left(\frac{m_i}{t_i}\cdot k_i\right) = \log_2\left(\frac{m_i}{t_i}\right) + \log_2(k_i) = \log_2\left(\frac{m_i}{t_i}\right) + c_i$$

$$M_i' = \log_2\left(\frac{V'_{Cy3,i}}{V'_{Cy5,i}}\right) = \log_2\left(\frac{t_i}{m_i}\cdot k'_i\right) = -\log_2\left(\frac{m_i}{t_i}\right) + \log_2(k'_i) = -\log_2\left(\frac{m_i}{t_i}\right) + c'_i$$

avec m_i la quantité de protéine i dans l'échantillon M, et t_i dans l'échantillon T. Le facteur k_i modélise la différence entre la protéine i et chacun des deux fluorophore. L'objectif est d'estimer la valeur

$$\log_2\left(\frac{t_i}{m_i}\right)$$

à partir des données disponibles : M_i et M'_i ainsi que du système :

$$\begin{cases} \log_2\left(\frac{m_i}{t_i}\right) + c_i = M_i & (1) \\ -\log_2\left(\frac{m_i}{t_i}\right) + c'_i = M'_i & (2) \end{cases}$$

où les deux inconnues c_i et c'_i représentent les biais liés au marquage pour la protéine i.

À partir de là, nous pouvons combiner les valeurs relatives M_i et M'_i des répétitions « dye-swap ». Pour ce faire, Yang et al. [62] proposent, sous le nom de « self-normalization », de considérer $c_i \approx c_i'$ quelle que soit la protéine i considérée. Sous cette hypothèse, en faisant (1)-(2), on obtient :

$$\log_2\left(\frac{m_i}{t_i}\right) = \frac{1}{2}\left(M_i - M'_i\right) .$$

Cette valeur correspond à la moyenne des 2 ratios logarithmiques T/M (calculés pour les 2 gels du « dye-swap ») des niveaux d'expression de la protéine i.

Yang et al. [62] proposent également de vérifier cette hypothèse à l'aide de « gènes ménagers », c'est à dire de gènes dont le niveau d'expression est connu pour être constant d'un échantillon à l'autre. De la même manière, il faudrait connaître des protéines de ménage dont le niveau d'expression serait constant pour tous les échantillons au sein de l'expérience.

IV.3.4.2 Analyse différentielle

Dans le cadre de la recherche de marqueurs protéiques potentiels, l'analyse différentielle entre une classe d'échantillons pathologiques et une classe d'échantillons témoins doit permettre de désigner les protéines dont l'expression est modifiée. Certains spots peuvent ainsi être induits, inhibés ou bien voient leur niveau d'expression varier d'une classe à l'autre.

L'étape de normalisation décrite précédemment permet la suppression d'un maximum de biais systématiques d'ordre expérimental pouvant masquer ces changements d'expression. La normalisation rend donc possible une première comparaison quantitative des populations des expressions protéiques que ce soit pour des patients, des classes ou bien des gels différents. Cependant, les imprécisions dans la détection et la quantification des spots, et les divers bruits liés à l'acquisition et au procédé expérimental font que l'analyse différentielle passe nécessairement par une analyse statistique du jeu de données. C'est pour cela que les jeux d'échantillons pathologiques et témoins doivent comprendre autant de gels que possible afin de permettre de distinguer des vraies expressions différentielles de celles issues du bruit, en apportant une information sur la variabilité biologique et expérimentale. Cette comparaison de classes doit se traduire par une valeur de significativité attribuée à chacune des protéines détectées par l'électrophorèse bidimensionnelle. Pour cela, elle doit tenir compte non seulement de la stabilité de la mesure, du ratio des valeurs interclasses mais aussi de la quantité effective de la protéine concernée. En effet, les mesures correspondantes aux faibles quantités protéiques sont davantage affectées par la variabi-

lité de la mesure et il en résulte un phénomène d'hétéroscédasticité du ratio inter-classe sur le domaine des quantités protéiques. Ces aspects doivent évidemment être pris en compte dans l'estimation de la significativité des changements d'expression.

Nous allons maintenant présenter les approches les plus généralement employées pour l'analyse différentielle dans le cadre de la comparaison de deux conditions en électrophorèse bidimensionnelle DIGE.

La technologie DIGE, permet le multiplexage de trois populations protéiques. Plusieurs schémas expérimentaux sont donc possibles. La pratique la plus répandue est de réaliser la comparaison entre les deux conditions biologiques d'intérêt directement au sein de chaque gel, en gardant la plus grande homogénéité possible. Par exemple, dans notre cas d'étude qui concerne l'échantillon issu du tissu sain et l'échantillon issu du tissu pathologique d'un même patient, ceux-ci doivent être étudiés au sein d'un même gel de manière à travailler sur des données appariées et de simplifier les interprétations.

Par ailleurs, la valeur de significativité d'une protéine, que peuvent apporter les tests statistiques, ne prend en compte aucune information a priori d'ordre biologique et pratique. Cette significativité ne traduit donc pas parfaitement l'intérêt que doit lui porter le biologiste. Il est donc intéressant d'introduire, au-delà de la significativité, une notion différente, celle de l'intérêt biologique. L'évaluation de cet intérêt biologique doit faciliter le travail d'investigation du biologiste en le guidant vers les protéines pertinentes non seulement en terme de significativité statistique mais également en terme biologique et pratique. En effet, si un ratio important et stable traduit une variation significative de l'expression, il peut cependant correspondre à des quantités protéiques faibles. Ceci peut poser problème pour l'identification en spectrométrie de masse mais aussi pour l'exploitation de la protéine en tant que marqueur de la maladie. En effet, certains biologistes considèrent que les protéines fortement exprimées ont davantage de chance de se retrouver dans le sang et donc d'y être détectée. La difficulté de la mise en place d'un indicateur de l'intérêt biologique vient du fait qu'il repose sur des informations difficilement quantifiables. En effet, il est par exemple impossible à un biologiste de connaître la probabilité de retrouver une protéine dans le sang en fonction de la quantité observée par électrophorèse à partir d'un échantillon tissulaire car le phénomène est très complexe et différent d'une protéine à l'autre. Cependant, à choisir, le biologiste va d'abord s'orienter vers les protéines fortement exprimées. Le biologiste est capable d'exprimer cette préférence en terme de risque. Ainsi, le biologiste jugera peut-être opportun de s'intéresser aux spots de faibles intensités à condition que leur significativité dépasse un seuil plus élevé que pour les spots de hautes intensités.

IV.3.5 Méthode de validation biologique

L'analyse différentielle vise à déterminer des marqueurs potentiels. La validité de ces marqueurs doit être vérifiée. En effet, l'objectif premier du projet NODDICCAP étant la mise au point d'un test de diagnostic sanguin, les marqueurs, découverts à partir des échantillons de tissu, doivent se retrouver dans la circulation

sanguine et s'y comporter de manière semblable. D'autres pistes peuvent être envisagées tel qu'un test sur les selles.

Pour rechercher la présence des marqueurs protéiques potentiels dans le sérum, l'immunodosage est l'approche la plus adaptée. Pour cela, il est nécessaire de disposer des anticorps spécifiques des protéines identifiées comme marqueurs potentiels. Si les anticorps dirigés contre ces protéines cibles ne sont pas disponibles dans le commerce il peuvent être obtenus par immunisation.

Après avoir caractérisé et sélectionné ces anticorps, il faut les utiliser pour doser des sérums de patients et ainsi déterminer si les marqueurs potentiels associés présentent un intérêt clinique. Pour cela, il faut mettre au point les tests ELISA permettant le dosage sérique de ces protéines chez les sujets sains et malades. Il faut noter que la complexité et la diversité des molécules, présentent dans un milieu biologique tel que le sérum, peut interférer avec la détection du marqueur et le rendre inutilisable. De plus, certaines protéines sont naturellement présentes dans le sérum des sujets sains. Il faut donc détecter un taux significativement modifié dans les sérums des sujets atteints d'un cancer colorectal. Le marqueur idéal serait donc une protéine absente dans le sérum de sujets sains. Mais, il n'est pas exclu que l'association de plusieurs protéines (dosage multi-paramétrique) soit une solution. Il pourrait y avoir deux ou plusieurs paramètres qui se présentent à des taux différents mais caractéristiques de l'apparition de la tumeur colorectale à ses tout premiers stades. Seule l'étude différentielle, sur un nombre représentatif de sérums, du taux sérique d'une protéine donnée permettra réellement de mettre en évidence son intérêt réel en tant que marqueur de diagnostic du cancer colorectal. C'est également cette approche qui permettra de doser l'éventuel marqueur à grande échelle.

IV.4 L'outil informatique

IV.4.1 Introduction

Comme nous l'avons vu, l'électrophorèse bidimensionnelle offre la possibilité de séparer des milliers de taches protéiques, de les quantifier et de les identifier par spectrométrie de masse si elles ne sont pas déjà répertoriées. Les approches protéomiques utilisant cette technologie supposent l'emploi d'un nombre souvent important de gels afin de révéler des différences d'expression des protéines. La quantité et la complexité des données sont telles que l'usage de méthodes puissantes et automatisées d'analyse se révèle indispensable.

IV.4.2 Évaluation du logiciel MELANIE IV

Afin d'orienter les travaux de la thèse, nous avons, dès le début, chercher à cerner dans le procédé d'analyse des images de gels d'électrophorèse bidimensionnelle, les problématiques trouvant une réponse satisfaisante dans les solutions logicielles et celles pouvant être améliorées. Pour cela, nous disposions d'un des logiciels de référence de l'époque, Mélanie 4.

Il s'agit d'un logiciel développé par l'Institut Suisse de Bioinformatique qui propose une solution tout en un pour l'analyse et l'annotation des images de gel d'électrophorèse 2D. Il offre de nombreuses possibilités telles que:

- L'affichage dynamique : zoom, menus contextuels, contrôle des couleurs, formes des spots, annotations visibles...
- La comparaison visuelle intuitive : transparence et affichage des vecteurs pour lier les spots du gel au premier plan aux spots de gels en arrière-plan.
- La détection automatique des spots, quantification, matching.
- Des outils et tests statistiques.
- Une classification par intelligence artificielle.
- Une solution pour la normalisation permettant de corriger certains facteurs affectant la reproductibilité des images.
- L'affichage d'un ou plusieurs gels.
- Les différentes possibilités d'annotations : étiquetage des taches avec liens hypertextes.
- La comparaison de gels, grâce à des analyses statistiques avancées et une classification.
- La quantification des taches protéiques.

L'évaluation du logiciel a été menée dans la continuité du travail effectué par Raman B et al [46] qui concernait les logiciels Mélanie 3.0 et Z3. Une description détaillée de la méthode utilisée ainsi que les images de travail sont disponibles sur le site internet : http://www.umbc.edu/proteome/. La Figure 38 présente les deux gels (a et b) choisis pour l'étude. Ils sont représentatifs de ceux couramment obtenus.

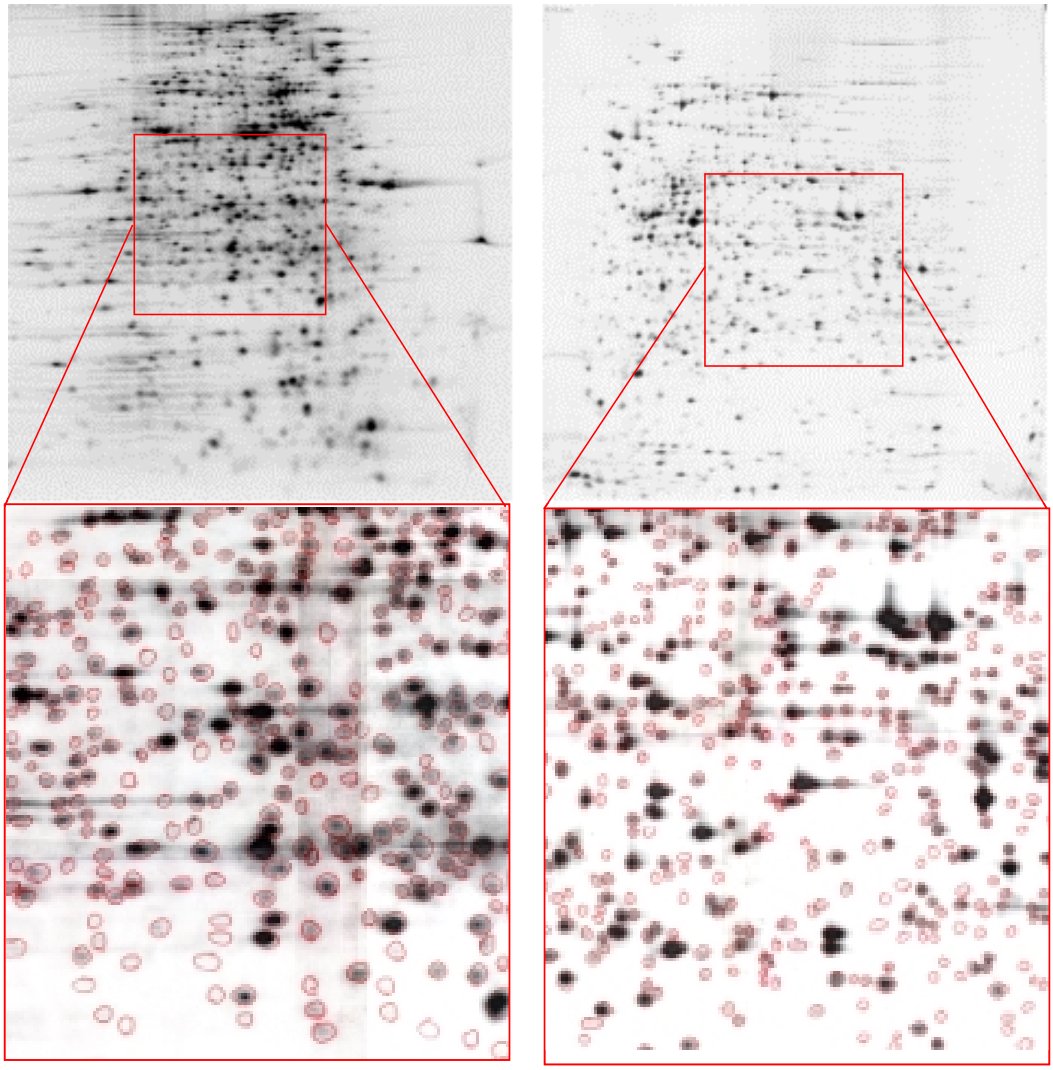


Figure 38: Gels choisis pour l'étude. Le gel à (gauche) et le gel b (droite) sont présentés ici, accompagnés chacun d'un détail de la détection des spots (entourés en rouge) faite manuellement par un expert.

IV.4.2.1 Évaluation de la détection des spots

Afin de pouvoir effectuer une estimation du nombre de spots faussement détectés ou omis par la détection, un comptage manuel a été réalisé par un expert sur chacun des deux gels de référence (Figure 38). Pour chacun de ces deux gels, la détection par défaut proposée par le logiciel est effectuée, puis les paramètres de détection, spécifiques à chacun des logiciels, sont ajustés de façon à obtenir une détection jugée optimale. Les résultats, exprimés en pourcentage du nombre de spots détectés manuellement, sont résumés dans le Tableau 1.

Gel a	% spots vrais détectés	% spots non détectés	% spots faux détectés
Melanie4 defaut	**88,3**	**11,7**	**32,7**
Melanie4 ajusté	**85,2**	**14,8**	**15,9**
Z3 defaut	89,0	11,0	14,0
Melanie3 defaut	93,0	7,0	350,0
Melanie3 ajusté	87,0	13,0	138,0

Gel b	% spots vrais détectés	% spots non détectés	% spots faux détectés
Melanie4 defaut	**89,7**	**10,3**	**11,7**
Melanie4 ajusté	**89,3**	**10,7**	**11,4**
Z3 defaut	89,0	11,0	14,0
Melanie3 defaut	93,0	7,0	165,0
Melanie3 ajusté	87,0	13,0	31,0

Tableau 1 : Évaluation de la qualité de la détection de 3 logiciels dédiés. Les pourcentages sont relatifs aux nombres de taches protéiques détectés manuellement par l'expert, ce qui constitue la référence idéale.

On constate que Mélanie IV est globalement beaucoup plus performant que Mélanie III. Le principal défaut de Mélanie III est le nombre très élevé de fausses détections, même une fois les paramètres ajustés.

Il apparaît que sur le gel a davantage que sur le gel b, Mélanie IV a lui aussi tendance à détecter de nombreux « faux spots ». Sur les deux gels la non détection reste importante.

IV.4.2.2 Évaluation de la mise en correspondance des gels

Pour cette évaluation, des images plus où moins déformées des gels a et b ont été soumises à l'algorithme de Mélanie IV avant d'établir la mise en correspondance de leurs taches protéiques avec celles de l'image originale du gel. Deux types de déformation ont été employés :

- une distorsion appelée « Center Pull » qui correspond à une courbure des lignes horizontales vers le bas telle que le montre la Figure 39
- et une distorsion appelée « Height decrease » qui correspond à un aplatissement vertical.

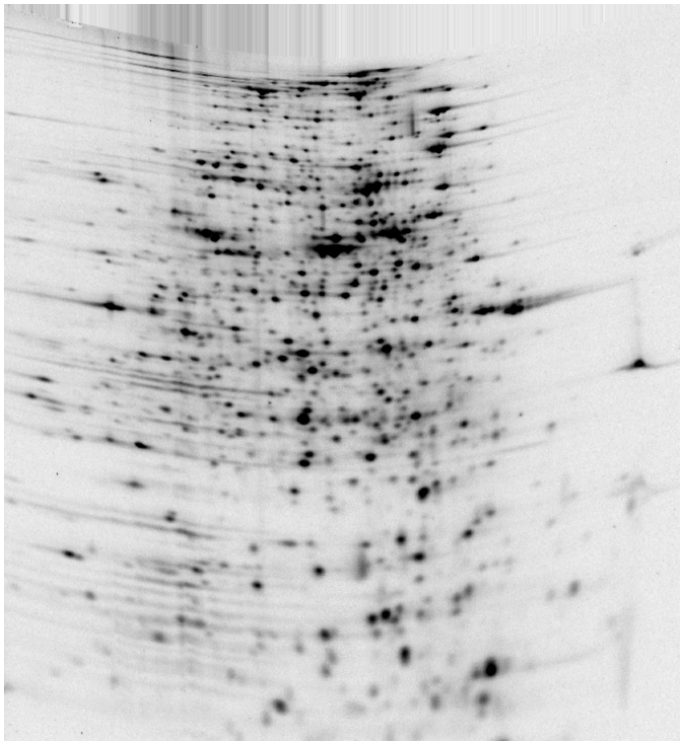


Figure 39 : Visualisation, sur le gel a, de la distorsion « center pull » (distorsion de 7.4%).

Pour les logiciels Mélanie III et Z3, un ajustement est possible. Une fois la détection des taches protéiques et l'alignement des gels effectués, la mise en correspondance sous Mélanie IV n'est pas ajustable. L'application des différents algorithmes de mise en correspondance produit les résultats résumés dans le Tableau 2, pour le premier type de distorsion (« Center Pull distorted »), et dans le Tableau 3, pour le deuxième type de distorsion (« Height decrease »). Les pourcentages associés aux distorsions sont arbitraires mais croissants avec l'importance de la distorsion.

Center pull	Gel a: Spots non matchés (%)			
Distorsion (%)	Melanie 4	Melanie 3 ajusté	Z3 défaut	Z3 ajusté
1.7	3	2	1.5	0.5
3.9	4	2.9	4	2.5
5.8	5	3	6	3.7
7.4	7	3	7	4
Center pull	Gel b: Spots non matché (%)			
Distorsion (%)	Melanie 4	Melanie 3 ajusté	Z3 défaut	Z3 ajusté
1.6	2	2.3	2.9	1
2.9	2	3.1	7	2
4.4	3	3	8	3.8
6.6	4	3.5	10.5	4

Tableau 2 : Résultats de l'évaluation, pour les 3 logiciels de l'étude, de la qualité de mise en correspondance des taches protéiques après différentes distorsions par « center pull » (courbure des lignes horizontales vers le bas) des images.

Height decrease	Gel a: Spots non matchés (%)		
Distorsion (%)	Melanie 4	Melanie 3 ajusté	Z3 ajusté
1	2.4	3	0.2
2	2.8	3	0.2
2.5	2.4		
4	3.1	4	0.4
5	3.3	4.3	0.6
Height decrease	Gel b: Spots non matchés (%)		
Distorsion (%)	Melanie 4	Melanie 3 ajusté	Z3 ajusté
1	1.1	1.5	0.2
2	1.6	3.5	0.3
2.5	1.1		
4	1.8	4.9	0.5
5	1.8	6	0.5

Tableau 3 : Résultats de l'évaluation, pour les 3 logiciels de l'étude, de la qualité de mise en correspondance des taches protéiques après différentes distorsions « height decrease » (aplatissement vertical) des images.

Il ressort de cette comparaison que, même s'il est faible, le nombre de taches protéiques n'ayant pas été appariées par Mélanie 4 reste supérieur à celui obtenu grâce au logiciel Z3, quelle que soit la nature de la distorsion et son amplitude. Par ailleurs, dans la réalité des études, les images mises en correspondance proviennent de populations protéiques différentes. La variabilité biologique inhérente à ces études crée des situations de mise en correspondance compliquées, même pour un œil expert. Par ailleurs, les distorsions observées au sein d'un gel et liées à son procédé de fabrication et à la migration des protéines, sont beaucoup plus complexes que celles créées artificiellement par distorsion d'une seule et même image. Il en résulte que les conclusions de cette étude ne sont valables que dans le cadre d'une comparaison.

IV.4.2.3 Évaluation de la quantification des taches protéiques

À partir d'images générées grâce au logiciel Matlab, il est possible de tester l'efficacité de la quantification des spots. Pour chacune des 11 images générées, visi-

bles sur la Figure 40, on connaît exactement le volume du spot central de profil gaussien (dont on connaît les paramètres).

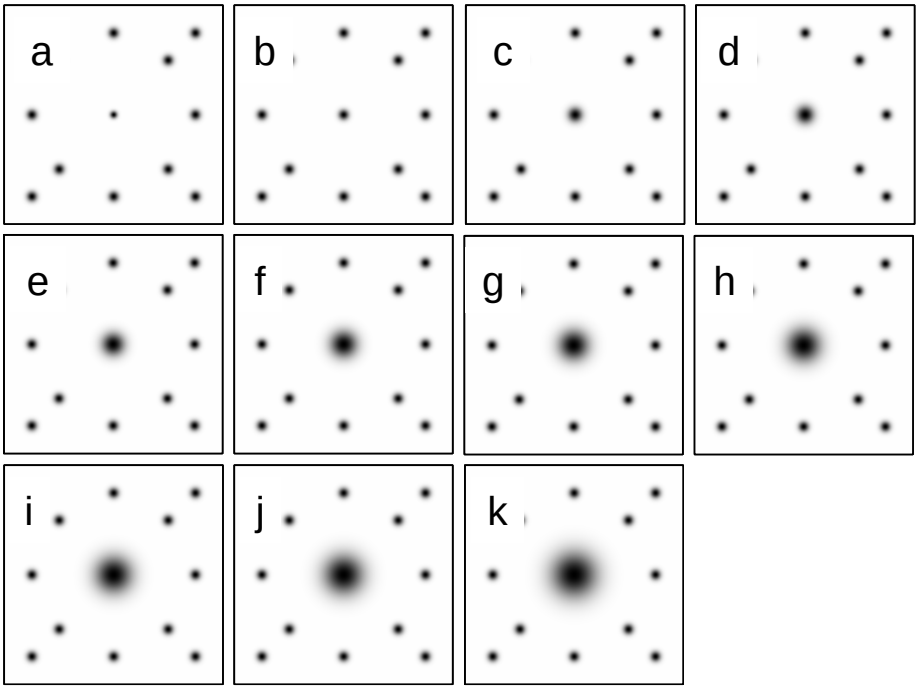


Figure 40 : Jeu d'images tests générées sous Matlab et permettant une évaluation objective de la quantification réalisée par un logiciel.

Pour notre évaluation, le ratio d'intérêt correspond au rapport du volume du spot central, sur le volume d'un spot de référence. Ce ratio est connu pour chacune de ces 11 images annotées de la lettre « a » à la lettre « k ». Les ratios théoriques sont : 2, 4, 6, 10, 14, 18, 22, 26, 30 et 40.

Evaluation de la quantification			
	Volumes observés	Ratios Théoriques	Ratios observés
a	93.4	1.0	1.0
b	186.1	2.0	2.0
c	371.5	4.0	4.0
d	556.5	6.0	6.0
e	917.6	10.0	9.8
f	1284.7	14.0	13.7
g	1651.5	18.0	17.7
h	2017.9	22.0	21.6
i	2384.0	26.0	25.5
j	2751.3	30.0	29.4
k	3668.5	40.0	39.3

Tableau 4 : Résultats du test de quantification des taches protéiques

Après détection des spots, la quantification effectuée par Mélanie IV permet d'établir les ratios. Les valeurs de ces ratios observés sont reportées dans le Tableau 4. Au regard de ces résultats, il apparaît que Mélanie IV répond de manière très satisfaisante au test de quantification de spots d'allure gaussienne puisque les ratios ob-

servés sont très proches des ratios attendus et qu'il n'y a aucune dérive suivant la valeur du volume associée.

IV.4.3 Autres solutions logiciels

Au cours des dernières décennies de nombreux logiciels ont été développés afin d'apporter des méthodes puissantes et automatisées pour l'analyse des images des gels d'électrophorèse bidimensionnelle. Le Tableau 5 présente les plus communs d'entre eux.

Logiciels	Société	année
Delta 2D	DECODON GmbH	2000
GELLAB II	Scanalytics	1999
Melanie IV	Geneva Bioinformatics S.A.	2002
PD Quest	Bio-Rad Laboratories Inc.	1998
Phoretix 2D Advanced	Nonlinear Dynamics Ltd.	1991
AlphaMatch 2D	Alpha Innotech Corporation	1999
Image Master 2D Elite	Amersham Pharmacia Biotech	2001
Investigator HT Analyser	Genomic Solutions Inc.	2000
Progenesis	Nonlinear Dynamics Ltd.	2001
Z3	Compugen	2000
Decyder2D Differential Analysis Software v6.5	Amersham Biosciences	...

Tableau 5: Quelques logiciels pour l'analyse des images de gel d'électrophorèse bidimensionnelle

Par ailleurs, de nombreux articles sont parus sur le thème de l'outil informatique dédié à l'électrophorèse bidimensionnelle. Ainsi Miller et Merril [39] proposent dès 1989, des méthodes pour l'évaluation des performances de ces outils.

Quelques articles présentent des comparaisons entre logiciels : Rosengren, Salmi et al. [49] mettent en concurrence PDQuest et Progenesis alors que Raman, Cheung et Marten opposent Z3 et Mélanie 3.0. Cependant, l'évaluation de ces logiciels est rendue difficile par la présence pour chacun d'un certain nombre de paramètres que l'utilisateur doit régler afin d'optimiser les différentes étapes de l'analyse (alignement des gels, détection, mise en correspondance et quantification des taches protéiques).

Une évaluation objective des algorithmes de détection des taches protéiques, des logiciels les plus courants, a été réalisée par Rogers, Graham et Tonge [48]. Un nouveau modèle statistique des taches protéiques, décrit dans l'article, fondé sur l'apprentissage à partir de jeux de gels réels et auquel on ajoute du bruit permet la création de gels de synthèse réalistes. La sensibilité de la détection des taches protéiques y est illustrée par des courbes FROC (Free-response Receiver Operator Characteristic) qui permettent de visualiser la relation entre le taux de vrais positifs et le nombre de faux positifs lorsqu'un paramètre change. Dix images de rapports signal sur bruit différents ont été utilisées pour tester la robustesse au bruit. La sensibilité au chevauchement a été évaluée à partir de 40 images sur lesquelles des chevauchements plus ou moins importants ont été simulés. ImageMaster a été jugé comme le système le plus efficace avec 85.1% de vrais positifs. Z3 est le logiciel s'accommodant le mieux du bruit tandis que PDQuest est le plus robuste au chevauchement. Mélanie 3 présente de bonnes performances pour les différents critères. Enfin, Progenesis offre l'avantage d'une détection non paramétrée (automatique) des taches protéiques ainsi

que de bons résultats pour la plupart des critères. Il ressort des résultats de l'étude que les logiciels testés, relativement récents, sont encore loin de permettre une analyse complètement automatisée. Si depuis la parution de cet article en 2003, les algorithmes de ces logiciels n'ont pas fondamentalement évolué, il faut cependant noter l'apparition de la notion de « co-détection » qui doit permettre une détection simultanée des taches protéiques sur les images issues d'un même gel lors d'une expérience DIGE. Ces co-détections sont basées sur des techniques de fusion d'images, et de détection de motifs. Ces co-détections nécessitent un alignement parfait de l'ensemble des images d'une expérience DIGE. Pour cela, Nonlinear Dynamics a enrichi son logiciel Progenesis d'un module nommé SameSpots Ce module, disponible indépendamment de Progenesis, permet un alignement supervisé par l'expert. Son usage est ergonomique et exploite l'alignement implicite des images d'un même gel DIGE.

La dernière version du logiciel ImageMaster base la co-détection des spots pour la DIGE sur la technique de ligne de partage des eaux à partir d'une fusion des images de gels.

IV.4.4 Conclusion

L'évaluation de Mélanie IV semble mettre en évidence que l'étape la détection des taches protéiques est l'étape la plus décisive et la plus perfectible. Cependant comme nous l'avons expliqué précédemment, l'évaluation de la qualité de l'appariement n'est valable qu'à titre de comparaison et l'usage montre que les défauts d'appariement en conditions réelles demeurent importants et rédhibitoires pour l'interprétation.

Par ailleurs, les différentes comparaisons et évaluations de logiciels rapportées dans la littérature tendent à montrer que les dernières versions des logiciels permettent une détection satisfaisante des taches protéiques. Le logiciel ImageMaster dispose d'un algorithme de détection efficace et fournit donc la solution que nous adopterons dans notre procédé d'analyse. La mise en correspondance des taches protéiques est également une étape critique de l'analyse et la technique de co-détection des spots semble la solution la plus prometteuse puisqu'elle permet une mise en correspondance parfaite.

En s'appuyant sur le logiciel ImageMaster dont la détection est performante et sur le logiciel TT900 (SameSpot) permettant un alignement efficace des images, il est possible de mettre en place un procédé performant. Ce procédé nécessitera cependant l'implémentation sous Matlab de certaines fonctions telles que la fusion d'image. Donc pour réaliser des analyses différentielles pertinentes en DIGE, il faut à la fois plusieurs logiciels et un programme ad hoc pour les mettre en relation. Le développement de ce procédé et l'implémentation du « workflow » correspondant ont été l'objectif principal de cette thèse. Cela a débouché sur une application qui permet de dérouler ce workflow : « ProDIGE ».

IV.5 Bilan de l'état des lieux

Cet état des lieux nous a tout d'abord conduit à placer la technologie DIGE comme héritière de l'électrophorèse bidimensionnelle classique afin d'exploiter les acquis d'une technologie largement employée depuis plus de 30 ans. Il en est ressorti une connaissance du contexte biologique et physico-chimique nécessaire pour la compréhension de l'origine des divers biais qui affectent les images de gels.

Ensuite nous avons présenté la technologie DIGE, ses avantages et ses limitations. L'atout majeur de la technologie DIGE est le multiplexage de populations protéiques différentes au sein d'un même gel. Lors d'une étude réalisée au sein d'un gel, il est alors possible de s'affranchir des biais de l'électrophorèse classique. Cependant, dans le cas d'une étude menée sur plusieurs gels, ce bénéfice est perdu.

Comme le montre notre tour d'horizon des principaux logiciels ainsi que l'évaluation de Melanie IV, l'outil informatique dédié à l'électrophorèse bidimensionnelle bénéficient aujourd'hui d'algorithmes de traitement et d'analyse d'images performants. Cependant, même les tout derniers d'entre eux n'exploitent encore que partiellement le bénéfice du multiplexage. Il apparaît donc aujourd'hui clairement une voie de recherche portant sur une approche stratégique qui optimiserait cette possibilité de multiplexage.

Nous avons donc pu cerner l'état de l'art et identifier une voie de recherche prometteuse. Nous allons maintenant chercher à utiliser ces connaissances afin de d'appliquer un procédé de traitement performant, novateur et adapté aux besoins et aux contraintes liés à notre projet de recherche de marqueurs tumoraux.

V. Élaboration d'un procédé d'exploitation optimal de la technologie DIGE

V.1 Introduction

Dans cette deuxième partie nous allons mettre en place le procédé développé, puis appliquer l'ensemble des méthodes qui le constituent au sein d'un « workflow » ou processus. Ce workflow doit permettre d'exploiter au mieux la technologie DIGE dans le contexte lié au projet NODDICCAP. L'application qui en permet le déroulement a été baptisée ProDIGE.

Dans un premier temps le travail a consisté à définir les attentes, les contraintes, ainsi qu'à établir l'inventaire des ressources et des moyens disponibles. Ensuite, à la lumière de l'état des lieux réalisée ci-dessus, ce cahier des charges a permis d'envisager les méthodes les mieux adaptées à chaque étape du traitement ainsi que plusieurs approches stratégiques. Il s'est agi alors de déterminer la meilleure des approches sur des critères objectifs. Enfin, nous présenterons la mise en application avec ProDIGE de ce workflow, sur un jeu de données réelles et nous discuterons des résultats obtenus ainsi que des perspectives.

V.2 Définition du cahier des charges

V.2.1 Introduction

Tout le travail effectué durant cette thèse se devait d'être applicable dans le contexte du laboratoire de protéomique de bioMérieux afin non seulement de pouvoir lui apporter une légitimité, mais aussi afin de contribuer à l'avancée du projet NODDICCAP. Il a donc été essentiel de définir un cahier des charges afin de prendre en compte ce contexte dans l'élaboration du procédé de traitement des données issues de la DIGE.

Ce cahier des charges précise non seulement les attentes des biologistes et les grandes étapes du traitement, mais également les moyens expérimentaux, matériels et logiciels ainsi que leurs contraintes afférentes. Il a pu être défini grâce aux nombreux échanges avec les biologistes acteurs du projet. Ces échanges ont eu lieu lors de fréquentes réunions de laboratoire, mais aussi et surtout au quotidien, par compagnonnage. Ainsi, dès que cela était possible, le travail se faisait en collaboration. Il m'a notamment été permis de manipuler les gels et de procéder à leurs acquisitions, avec l'assistance des biologistes. Cette proximité a permis de cerner au mieux les attentes, les moyens à disposition et les diverses contraintes. Il faut noter, par ailleurs, que ce cahier des charges a évolué avec le temps avec notamment l'acquisition par le laboratoire de nouveaux logiciels.

V.2.2 Les attentes des biologistes

La motivation première de l'utilisation de la technologie DIGE dans le contexte de notre projet est de pouvoir comparer des échantillons pathologiques à des échantillons sains afin de mettre en évidence des marqueurs protéiques potentiels.

La technologie DIGE est parfaitement adaptée à ce type d'étude. En effet, la comparaison de deux échantillons protéiques ayant migré dans un même gel se révèle beaucoup plus juste, en terme de quantification et de mise en correspondance des taches protéiques, que s'ils avaient migré dans deux gels séparés. Par ailleurs, les échantillons dont nous disposons dans le cadre du projet NODDICCAP sont appariés puisque que pour chaque patient nous disposons d'un prélèvement de tissu sain et un autre de tissu pathologique.

Cependant, l'objectif n'est pas une simple étude patient par patient. Le risque serait en effet de passer à coté de phénomènes moins visibles en terme d'intensité, mais pourtant récurrents chez un grand nombre de patients et donc potentiellement très intéressants. L'objectif est de pouvoir mener une étude globale sur l'ensemble des patients d'une expérience et de pouvoir ensuite recouper les résultats entre plusieurs expériences.

Cette étude globale doit permettre de faire ressortir tous les marqueurs protéiques potentiels et d'évaluer pour chacun d'eux une valeur de significativité. Cette valeur doit permettre aux biologistes de s'orienter vers les protéines les plus intéressantes et légitimer leurs choix.

La robustesse du procédé, si elle est limitée par un certain nombre de critères de qualité imposés aux données, doit être suffisante pour permettre son application à des expériences séparées.

Par ailleurs, le procédé doit pouvoir être appliqué par un biologiste et les résultats doivent être facilement interprétables et réexploitables.

V.2.3 Les grandes étapes du traitement

Le rapprochement des attentes des biologistes et de la nature des données brutes permet de dégager les étapes essentielles à la base du procédé mis en place (voir Figure 41).

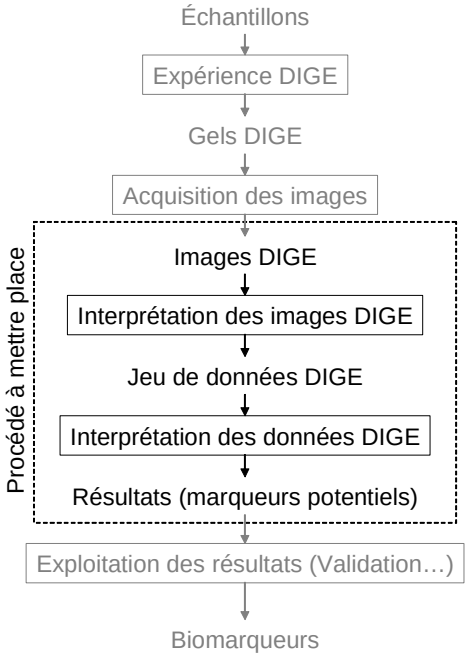


Figure 41 : Schéma macroscopique des grandes étapes nécessaires à l'exploitation de la technologie DIGE

D'un point de vue macroscopique, le procédé de traitement des images DIGE est classique. Il se compose en effet d'une étape d'interprétation lors de laquelle on cherche toujours à extraire l'information pertinente contenue dans l'image afin de l'exploiter dans une seconde étape. Il est donc nécessaire de bien identifier les informations à extraire et la manière de les interpréter.

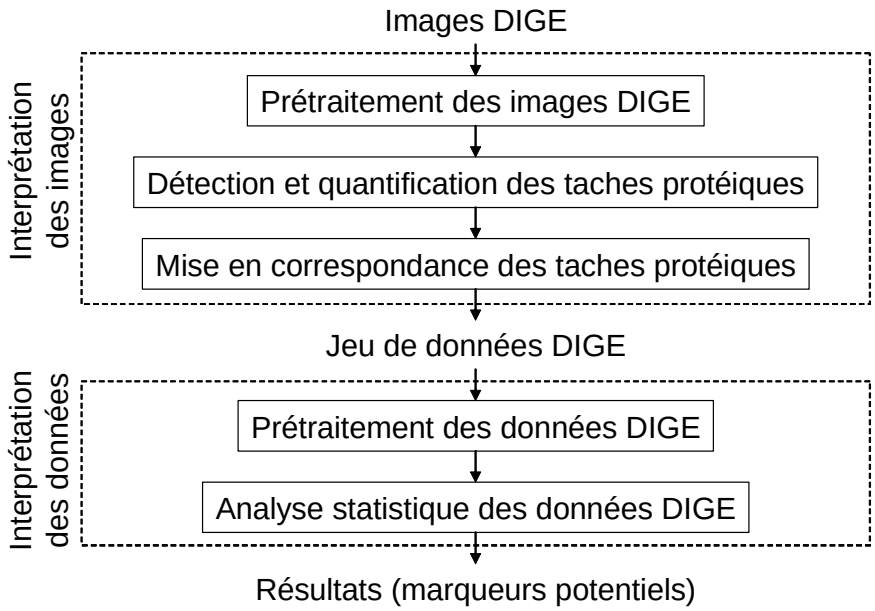


Figure 42 : Schéma des grandes étapes nécessaires à l'exploitation de la technologie DIGE

Pour l'étude différentielle, il nous faut disposer de la quantification d'un maximum de protéines pour chacun des échantillons de l'expérience. Il est donc nécessaire de détecter et de quantifier chacune des taches protéiques présentes sur chaque image de l'expérience. Mais cela n'est pas suffisant car il faut également mettre en relation les taches correspondantes à une même protéine, d'une image à l'autre, afin de pouvoir comparer leur quantification. Par ailleurs, afin de tenir compte des divers biais liés à l'expérience, une étape de prétraitement d'image et une autre de prétraitement des données doivent faire partie du procédé. Au final, comme l'illustre la Figure 42, il est maintenant possible de se faire une idée un peu plus précise des étapes incontournables du procédé à mettre en place.

V.2.4 Les moyens à disposition pour l'analyse et les contraintes afférentes

BioMérieux dispose d'une plateforme technologique dédiée à la DIGE qui comprend notamment :

- un système d'imagerie, le ProXPRESS, présenté dans le paragraphe IV.2.2, permettant d'acquérir les images des gels en fluorescence
- ainsi qu'un ensemble de logiciels de traitement d'image (ProSCAN, ImageMaster 2D platinium, TT900 S2S).

Par ailleurs, l'implémentation de nos propres méthodes a été réalisée grâce au logiciel Matlab (Mathworks) et à sa boite à outils de traitement d'image. Les logiciels de traitement d'images ont chacun des rôles distincts. ProSCAN permet de commander le ProXPRESS et de paramétrer l'acquisition des images. ImageMaster 2D Platinium est un logiciel très complet pour l'interprétation des images de gels. Il n'est cependant pas entièrement adapté à notre contexte et à nos attentes. En effet, ImageMaster 2D Platinium, bien que performant pour la détection des taches protéiques et pour la manipulation et l'annotation des gels, ne permet pas de prendre en

considération la structure de nos données issues d'échantillons appariés avec répétition par inversion des fluorophores (« dye-swap »). Enfin, le logiciel TT900 permet aux biologistes de déformer les images de gels de manière à pouvoir les aligner et ensuite de pouvoir les exporter. Ce logiciel est adapté à la structure des images DIGE, c'est-à-dire qu'il tient compte de l'alignement intrinsèque des 3 images issues d'un même gel. L'intégration ou non de ces logiciels dans le workflow mis en place est dictée par des critères de performance, d'ergonomie et par la stratégie choisie pour le traitement des images. Ces logiciels ont chacun leurs points forts et présentent une bonne ergonomie, ce qui nous a poussé à les exploiter le plus souvent possible. La contrepartie de ce choix est d'avoir du adapter, grâce à Matlab, le format des données entre chacun d'entre eux afin de pouvoir les faire cohabiter.

D'un point de vue expérimental, il est important de noter que des limitations existent en terme de quantité d'échantillon et de gels réalisables par expérience. En effet une expérience DIGE ne pourra compter plus de 12 gels.

Enfin, le choix a été fait, très tôt, de réaliser deux gels pour chaque patient avec simplement une inversion du marquage en fluorescence. Il s'agit du schéma expérimental communément appelé « dye-swap ». Le procédé à développer doit proposer une prise en compte de schéma afin d'en exploiter les bénéfices décrits dans le paragraphe IV.3.4.1.2b).

V.3 Mise en place du processus de traitement « ProDIGE »

V.3.1 Les problématiques à résoudre

Chacune des grandes étapes du traitement décrites plus haut a pour objectif de répondre à une problématique bien précise. Certaines problématiques trouvent leur solution dans des traitements déjà proposés par les logiciels ou bien implémentés à l'aide de Matlab. Ces problématiques, sur lesquelles nous reviendrons plus loin sont celles, du prétraitement des images (pour la suppression du bruit et de biais systématiques) ainsi que de la détection et de la quantification des taches protéiques. En effet, au début du travail de thèse un diagnostic a été dressé sur les points faibles et les points forts du traitement classique des images de gels d'électrophorèse. Ainsi, il avait été constaté que la partie de l'interprétation de l'image qui consiste en la détection et la quantification des taches protéiques était efficace et que la source d'erreur majeure dans l'interprétation provenait d'une mise en correspondance souvent défectueuse de ces taches. Il en résultait des jeux de données incomplets et peu fiables. C'est-à-dire que la quantification d'une protéine ne pouvait pas être réalisée sur chacun des gels de l'expérience et quand elle l'était il subsistait un doute sur l'identité réelle de la protéine. Avec l'apparition de logiciels d'alignement d'images performants et ergonomiques, une nouvelle option est apparue. En effet le travail sur des images parfaitement alignées ne nécessite plus d'algorithmes de mise en correspondance des taches protéiques. Ces algorithmes sont parfois compliqués dans leur mise en oeuvre et sont souvent inefficaces. Cette mise en correspondance est implicitement contenue dans la situation géographique de la tache sur l'image. Mais l'emploi de cette technique amène de nouvelles problématiques. Il s'agit de déterminer si le concept est généralisable à des images issues d'expérience différentes. Il s'agit également de savoir comment considérer un maximum de taches protéiques car aucune image n'est assurée de contenir toutes celles présentes sur l'ensemble des images de l'expérience.

Les données issues de l'interprétation des images nécessitent à leur tour un prétraitement appelé normalisation qui permet de s'affranchir des biais systématiques les affectant. L'observation de ces données a montré que les techniques habituelles pouvaient être améliorées en s'appuyant notamment sur celles employées dans le domaine des puces à ADN. En effet, la stratégie de mise en correspondance implicite des taches protéiques, par alignement préalable des images, permet de disposer de jeux de données sans valeur manquantes similaires aux jeux de données des puces à ADN. Par ailleurs, les observations ont également montré une dépendance entre la variabilité des changements d'expression et la gamme d'intensité dans laquelle ils sont observés. Les changements d'expressions concernant les taches protéiques de faible intensité sont plus variables que ceux concernant les taches de forte intensité. Les techniques communément employées ne tiennent pas compte de ce phénomène.

Apporter une solution à cette problématique revient à affiner un peu plus le choix des biomarqueurs potentiels.

V.3.2 Aspects stratégiques de l'analyse

L'objectif du projet NODDICCAP est de découvrir des marqueurs protéiques du cancer colorectal et le choix a été fait d'établir une liste de marqueurs potentiels à partir de l'étude différentielle entre tissus sains et tissus pathologiques. Afin de garder une meilleure cohérence des données et faciliter l'interprétation des résultats, le tissu sain et le tissu pathologique d'un même patient sont traités dans un même gel. Il faut également rappeler que, pour chacun des patients, une forme de répétition est réalisée qui consiste en l'usage d'un deuxième gel sur lequel le marquage est inversé, c'est le schéma dit en « dye-swap ». Ces choix expérimentaux ainsi que l'objectif du projet NODDICCAP constituent une base à partir de laquelle il nous faut construire le workflow le mieux adapté aux problématiques présentées précédemment.

La construction de ce workflow relève à la fois du choix des méthodes qui le composent mais également du choix de la stratégie liée à l'usage de ces méthodes. Ces choix stratégiques concernent d'une part l'analyse des images et d'autre part l'analyse des données.

En ce qui concerne l'analyse des images, il faut déterminer la stratégie permettant d'exploiter pleinement le bénéfice de la technologie DIGE. Différentes solutions ont été envisagées et ont fait l'objet de l'étude comparative présentée dans le paragraphe V.3.3.3.

Pour ce qui est du traitement des données, la volonté de conserver une homogénéité en affectant un seul patient par gel, afin de faciliter l'interprétation, conduit à considérer le rapport entre la quantification réalisée pour le tissu pathologique et celle pour le tissu sain, pour chacune des protéines discernables sur les deux images d'un même gel. Il s'agit alors de sélectionner les protéines correspondant aux ratios les plus significatifs. Pour définir une méthode de sélection, il est donc tout d'abord nécessaire de s'interroger sur le sens que l'on donne à la significativité ou tout au moins à l'intérêt porté à une protéine. En effet, par exemple, pour des valeurs de ratios identiques, une protéine peut être plus intéressante qu'une autre car plus abondante et donc possédant de plus grandes chances d'être exploitable dans le cadre de la recherche de marqueurs. Elle sera par exemple davantage susceptible de se retrouver dans la circulation sanguine, qui est une voie de recherche pour la mise au point d'un test de dépistage rapide. Ces questions seront abordées dans le paragraphe V.3.3.5 et permettront la mise en place d'une stratégie pour la fouille des données.

V.3.3 Justification et description des méthodes à chaque étape du traitement

La stratégie d'analyse a été mise en place afin de répondre aux exigences décrites précédemment et constitue donc un compromis. Il en va de même pour le choix des méthodes qui composent le processus d'analyse. Depuis l'acquisition des images jusqu'à l'analyse des données, le choix des méthodes est dicté par la performance, les spécificités de l'environnement et du protocole et la simplicité de mise en œuvre.

Ces différents aspects sont très liés. En effet, pour chaque étape du traitement, la littérature fournit de nombreuses méthodes souvent très performantes dans le cas précis de l'application présentée. Malheureusement, le protocole et les différentes contraintes propres à nos recherches sortent bien souvent de ces cadres d'application. Par ailleurs, les logiciels informatiques à disposition sont performants et ergonomiques ; s'en passer signifierait un développement conséquent qui n'assurerait pas une amélioration des traitements. Il s'agit en fait de composer avec tous ces éléments et de les compléter afin d'optimiser l'utilisation des ressources quand elles sont disponibles et jugées performantes, et de choisir les meilleures méthodes afin de compléter le processus.

V.3.3.1Prétraitement des images

Le prétraitement des images a pour but de corriger les bruits et les biais affectant directement l'image, concernant sa géométrie et l'intensité de ses niveaux de gris. Ce prétraitement nécessite de travailler sur l'information brute, c'est-à-dire sur les images issues de l'acquisition des gels.

3.3.1.1 Filtrage des images

Les bruits et ces biais, seulement identifiables sur l'image, sont constitués :

- des bruits de haute fréquence spatiale tels que les bruits électroniques, les poussières, les cassures,
- ainsi que des bruits de basse fréquence spatiale qui se traduisent par l'inhomogénéité du fond sur la surface de l'image.

Les gels obtenus en routine par le laboratoire ne présentent pas tous les défauts classiquement observés. Comme illustré sur la Figure 43, les bruits de hautes fréquences sont peu importants et les spots même faibles s'en détachent largement.

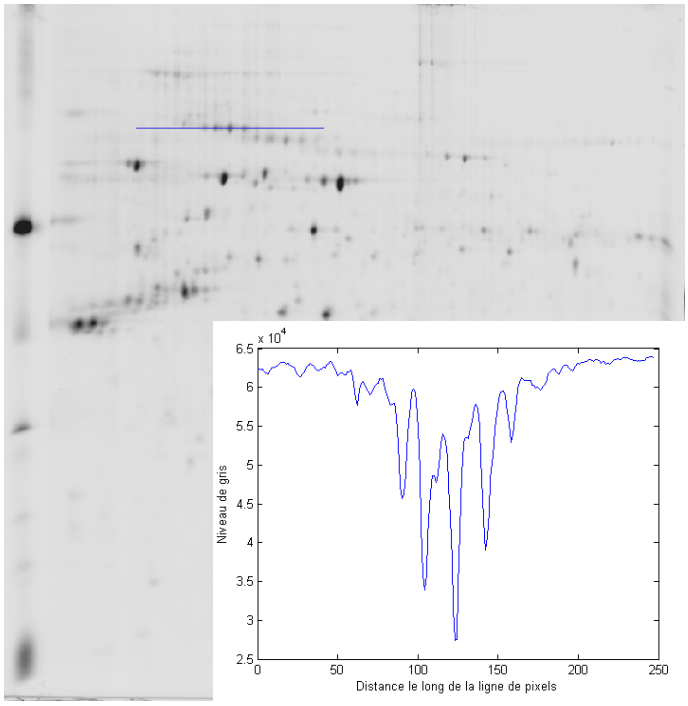


Figure 43: Profil de niveau de gris le long d'un segment (en bleu) sur l'image d'un des gels produits en routine par le laboratoire.

Il a donc été choisi de filtrer l'image de manière douce et respectueuse de la morphologie des spots. Pour cela, nous avons opté pour un filtrage médian. Comme nous le présentions au paragraphe IV.3.2.1.1b), le filtrage médian est un filtre d'ordre indépendant de la gamme dynamique employée et respectueux de la morphologie des éléments significatifs de l'image. La taille du masque de convolution la mieux adaptée a été déterminée de façon pragmatique sur les images obtenues en routine au sein du laboratoire. Ainsi, un masque de 3x3 permet de respecter au mieux la morphologie des taches protéiques tout en supprimant les bruits de plus hautes fréquences. Prendre un masque de taille supérieure c'est prendre le risque d'affecter les taches protéiques les plus petites. En effet, les taches les plus petites sont très proches du bruit, en terme de fréquence spatiale.

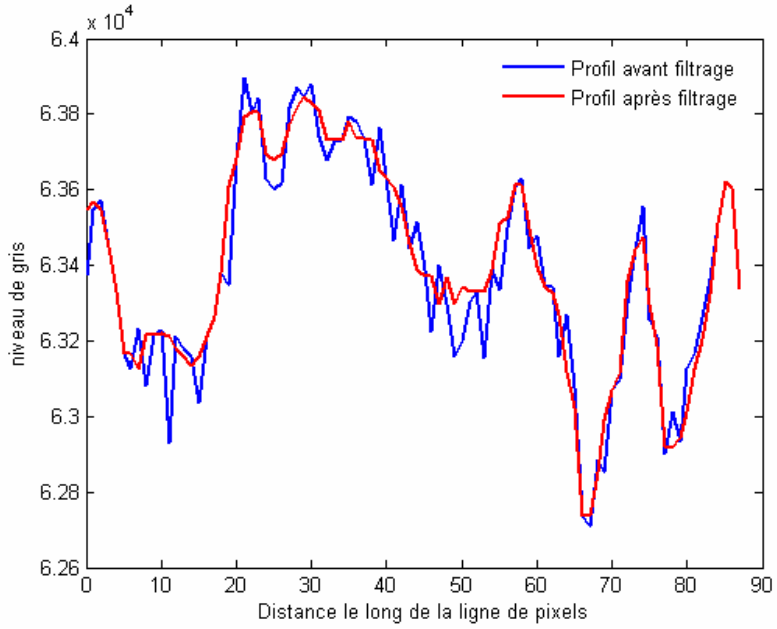


Figure 44: Visualisation de l'effet d'un filtrage médian sur un profil des niveaux de gris du fond de l'image.

Si l'effet est peu visible à l'échelle dynamique des spots, il est clairement visible à l'échelle du fond de l'image (voir Figure 44).

Par ailleurs le bruit de basse fréquence observé est faible et, surtout, il a été jugé comme étant potentiellement porteur d'information par les experts biologistes. Des essais menés au laboratoire de protéomique consistant en l'acquisition de gels « vierges » (n'ayant subi aucune migration protéique) ont confirmé l'impact négligeable de la technologie sur l'inhomogénéité du fond. Lors des expériences DIGE, cette inhomogénéité est principalement due à des phénomènes biologiques tels que la glycosylation. Les protéines glycosylées génèrent des trains de spots avec de légères variations de point de masse et de point isoélectrique. Ces spots peuvent être adjacents ou même confondus, produisant une augmentation du signal sur l'ensemble de la région du gel concernée. Le fond possède donc un sens biologique et n'est donc pas supprimé.

L'acquisition est faite de façon indépendante pour chaque fluorophore de chaque gel. Le temps d'acquisition est optimisé pour chaque longueur d'onde en opérant systématiquement une pré-acquisition de chaque canal afin d'estimer la durée d'acquisition permettant d'occuper au mieux la gamme dynamique du scanner. Les écarts d'intensité des images obtenues ne reflètent donc pas les rendements quantiques des différents fluorophores. Le fait d'augmenter la durée d'acquisition correspond simplement à une transformation linéaire des données, les valeurs relatives des intensités d'une même image sont conservées et surtout, les spots de faibles intensités peuvent être mieux quantifiés. Cependant, l'étape de normalisation reste nécessaire. En effet, la maximisation de la gamme dynamique de chaque image ne correspond pas à un alignement d'une quantité qui serait la traduction d'une intensité moyenne de chaque population de taches protéiques mais plutôt à l'alignement des maximums détectés sur chaque image. Ces maximums correspondent en général à une des taches

protéiques du marqueur de poids moléculaire qui est présent sur la gauche de tous les gels réalisés. Bien que réalisé avec les mêmes quantités protéiques, la mise à niveau des intensités observées pour une même tache protéique issue de ce marqueur de poids moléculaire n'est pas fiable car fondée sur une seule mesure, sujette à tous les biais et bruits possibles de l'image. D'autant plus que, de part sa position sur le gel, la tache protéique correspondante et donc la quantification qui en est faite, subissent des effets de bords.

3.3.1.2 Alignement des images

L'électrophorèse bidimensionnelle permet de séparer de nombreuses protéines contenues dans un échantillon. Théoriquement, chacune de ces protéines se situe sur la surface du gel selon sa masse moléculaire et son point isoélectrique. En pratique, on observe que les positions relatives des protéines sont différentes d'un gel à l'autre. Ces variations de la cartographie des protéines sont liées à des facteurs physiques et biologiques :

- un gradient de concentration inhomogène de la matrice du gel de polyacrylamide,
- des fuites de courant lors des étapes de migration,
- des modifications post traductionnelles (glycosylation ou phosphorylation).

Comme nous l'avons évoqué lors de l'étude bibliographique, certaines méthodes se proposent de corriger les variations dues aux fuites de courant. Malheureusement, le modèle est insuffisant car ces fuites ne sont pas le seul facteur intervenant. Par ailleurs, dans le cadre de l'étude différentielle d'échantillons protéiques, le cœur du problème n'est pas de situer une tache protéique à ses coordonnées réelles de masse moléculaire et de point isoélectrique. Le cœur du problème est de s'assurer de pouvoir désigner les taches protéiques correspondant à une même chaîne polypeptidique quelle que soit l'image de gel considérée. L'identité de la protéine n'est pas révélée par ses coordonnées sur le gel mais par une étude ultérieure en spectrométrie de masse.

De la même manière, les différentes méthodes d'alignement automatique des images ne peuvent pas intégrer les facteurs biologiques affectant les cartographies. C'est pour cela que, pour cette étape du traitement des images, l'expertise d'un biologiste s'avère indispensable.

Il s'agit alors de faciliter la tache de l'opérateur biologiste, en mettant à sa disposition un outil ergonomique et optimisé pour le traitement des images issues de la DIGE. Dans le cadre du projet NODDICCAP, le choix s'est porté sur le logiciel dédié TT900 de Nonlinear Dynamics. Ce logiciel permet de se contenter de l'alignement des images de la population protéique du standard interne (lignée cellulaire de Caco-2 dans notre cas) présent sur pour le Cy2 de chacun des gels. Selon le principe DIGE, après l'alignement des Cy2 à l'aide de TT900, les 3 images issues d'un même gel étant parfaitement et naturellement alignées, le logiciel en déduit l'alignement de l'ensemble des images de l'expérience.

L'alignement des images est donc plus qu'un simple prétraitement des images. Ce traitement constitue déjà une première interprétation des images puisqu'il nécessite une expertise et qu'il enrichit l'information portée par l'ensemble des images.

V.3.3.2Fusion des images

La fusion des images a pour finalité la constitution du jeu de données complet des intensités de l'ensemble des taches protéiques présentes sur l'ensemble des images de l'expérience considérée. Elle est rendue possible grâce à l'alignement préalable des images. Les deux étapes d'alignement et de fusion se substituent à l'étape de mise en correspondance des taches protéiques systématiquement présente sur la première génération de logiciels dédiés. Pour rappel, la détection des taches protéiques se faisait alors, indépendamment, sur chacune des images. Ensuite, l'utilisateur imposait la mise en correspondance de quelques taches protéiques en plaçant des « landmarks » dont un algorithme se servait ensuite comme conditions initiales pour une mise en correspondance automatisée de l'ensemble des taches protéiques. Malheureusement cette méthode conduisait systématiquement à des jeux de données incomplets et comprenant un certain pourcentage d'erreur de mise en correspondance.

Comme on l'a vu plus haut, l'alignement préalable des images permet l'utilisation d'un unique patron pour la quantification de toutes les taches protéiques sur toutes les images. Cette nouvelle approche assure un jeu de données complet ainsi que l'absence d'erreur de mise en correspondance. Tout l'intérêt de la fusion des images est dans l'obtention de ce patron. La fusion se doit de conduire à l'obtention d'une image représentative d'un ensemble d'images de gels d'électrophorèse. Chaque tache protéique présente sur une de ces images de gel d'électrophorèse doit avoir une tache la représentant sur l'image de fusion.

Avant de fusionner les images, il faut s'assurer qu'elles sont semblables en terme de dynamique des niveaux de gris d'intérêt, c'est-à-dire des niveaux de gris des taches protéiques. L'étalement de la dynamique, qui pourrait être réalisée lors du prétraitement, n'est pas suffisant dans l'optique d'une telle mise à niveau et une égalisation spécifique doit être réalisée.

3.3.2.1 Égalisation des images avant fusion

Certains phénomènes intervenants de manière ponctuelle, tels que le bruit (rare) et surtout la forte expression biologique de seulement quelques protéines, peuvent modifier l'étendue de la gamme dynamique sur les images où ils sont présents (cf Figure 45).

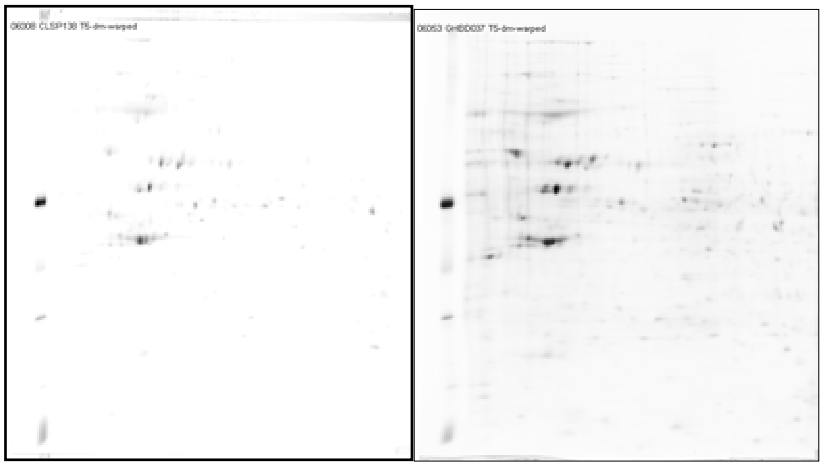


Figure 45: Observation de la différence de contraste pouvant exister, malgré l'étalement de la dynamique opéré lors du prétraitement, entre les différentes images à exploiter. Pour des quantités protéiques équivalentes, les taches apparaissent à des niveaux d'intensité supérieurs sur l'image de droite par rapport à l'image de gauche.

Comme l'illustrent les deux images présentées sur la Figure 45, ces phénomènes causent alors, entre les différentes images, des écarts entre les niveaux moyens d'intensité des taches protéiques. Il existe alors des différences importantes en terme d'intensité (niveau de gris) entre deux taches correspondant pourtant à une même quantité protéique. L'objectif de l'égalisation des images est donc de rapprocher les niveaux de gris correspondants à des quantités protéiques équivalentes.

Pour cela il est tout d'abord nécessaire de définir les valeurs d'intensité d'intérêt. Le choix se porte sur les valeurs du niveau de gris des centroïdes des taches protéiques car elles sont le reflet des quantités protéiques.

Afin de réaliser cette mesure, un algorithme de détection basé sur la recherche de minimums locaux a été implémenté sous matlab. L'algorithme retient les valeurs de niveaux de gris associées aux centroïdes des taches protéiques les plus intenses.

```
%Filtrage alterné séquentiel nécessaire avant détection
se = strel('disk',2,0);
imgf2{i}=graydil(img{i},se);
imgf2{i}=grayero(imgf2{i},se);
% détection des sommets de spots par recherche des minimums locaux
BW{i} = imextendedmin(imgf2{i},0);
L=bwlabel(BW{i},4);
% recherche des centroides des minimum locaux
s = regionprops(L, 'centroid');
centroids = cat(1, s.Centroid);
% évaluation de l'intensité de chacun des sommets
I{i}=diag(img{i}(floor(centroids(:,2)), floor(centroids(:,1))));
T=[I{i} centroids(:,1), centroids(:,2)]; % On regroupe les informations utiles dans un tableau
T_sorted{i} = sortrows(T,1); % on ordonne le tableau selon l'intensité décroissante
```

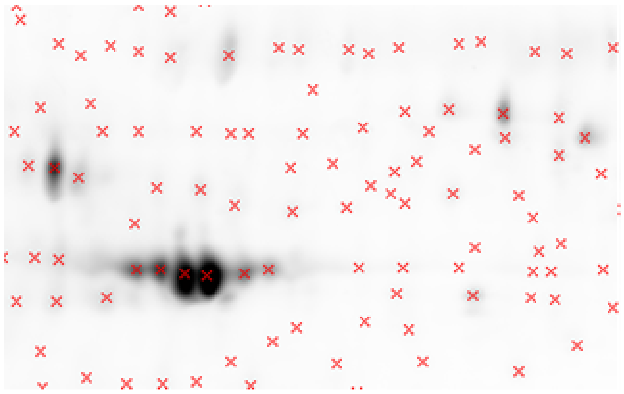


Figure 46: Visualisation de la détection des centroïdes des taches protéiques sur un détail d'une image de gel d'électrophorèse.

L'égalisation des images se base ensuite sur l'hypothèse selon laquelle les distributions des quantités protéiques associées aux taches protéiques sont semblables d'un échantillon à l'autre. Cette hypothèse s'applique de la même manière aux distributions des valeurs de niveau de gris des centroïdes. Or, comme l'illustre la Figure 47, ces distributions présentent des différences significatives. En effet, l'observation des écarts importants entre les 1ers déciles de chaque population est le signe de la présence d'une valeur de fond propre à chaque image. De même, les écarts, encore plus importants, entre les valeurs des 9ème déciles laissent à penser que chaque population est également affectée par un facteur d'échelle propre. L'idée est donc de corriger les images de manière à rapprocher les distributions les unes des autres.

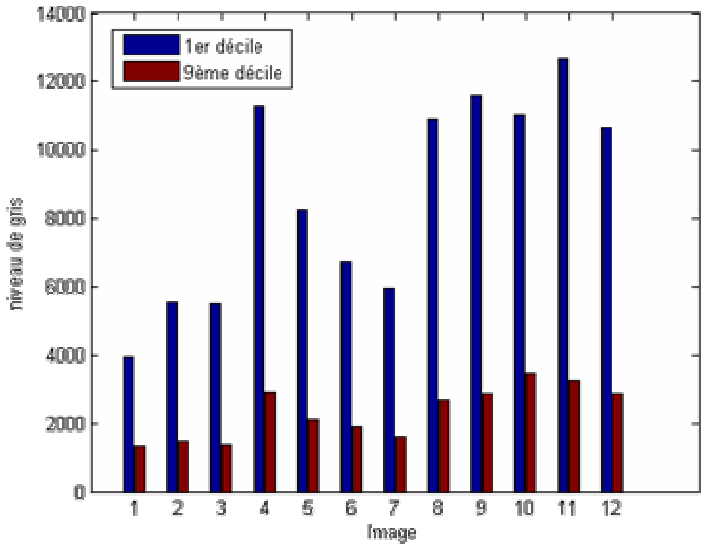


Figure 47: Visualisation du 1er et du 9ème décile avant égalisation pour chacune des 12 populations d'intensité de taches protéiques correspondant aux 12 images de gel d'électrophorèse (expérience DIGE de juin 2006).

Afin de quantifier les écarts entre les distributions et donc de déterminer les corrections à appliquer, on utilise une comparaison par quantile (qq-plot) entre chacune des distributions et une distribution médiane. La population médiane est définie à partir de l'ensemble des populations d'intensité d'intérêt de telle sorte que la nième

intensité la plus forte de la population médiane soit la valeur médiane des nièmes valeurs d'intensité la plus forte de chaque distribution. La Figure 48 présente un exemple de comparaison par quantile entre la distribution associée à une image particulière et la distribution médiane associée à l'ensemble des images de l'expérience considérée.

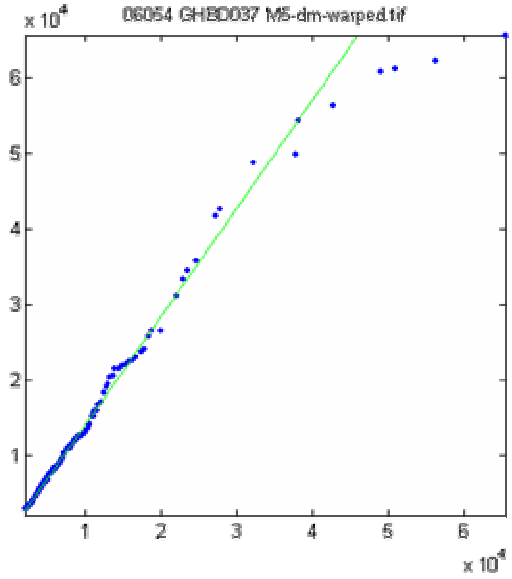


Figure 48: Graphique quantile-quantile permettant la comparaison entre la distribution d'une des 12 images d'une expérience DIGE (juin 2006) et la « distribution médiane » associée à ces 12 images. La régression LTS est également visible (en vert).

Il est alors possible, par régression linéaire, d'estimer la transformation linéaire permettant de rapprocher les différentes distributions de la « distribution médiane ». La régression utilisée est la régression LTS (Least Trimmed square), qui permet d'écarter les valeurs aberrantes. Les valeurs aberrantes correspondent aux taches protéiques les plus intenses. Ceci s'explique par le fait qu'il suffit qu'il y ait, sur certaines images, davantage de taches de forte intensité pour que les 1ers quantiles soient très différents. Sur la Figure 48, la droite de régression LTS est tracée en vert. Elle fournit les éléments nécessaires à la correction des intensités des images : son ordonnée à l'origine et sa pente. L'égalisation est faite en ramenant les droites de régression à la première bissectrice.

Le code nécessaire à l'égalisation est montré ci-dessous :

```
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
% Recherche des coefficients pour l'égalisation à partir d'une régression linéaire LTS
% 1ère étape: création d'une population A d'intensités représentative de l'ensemble
% des populations
for i = 1:nb_img
    AA(:,i) = double(T_sorted{i}(1:nb_spots,1));
end
A = median(AA,2);
Imed = median(AA,1);

% 2ème étape: Régression LTS (robuste) entre chacune des populations et la
% population A afin de déterminer les coeff pour l'égalisation
A = double(65536-A);
for i = 1:nb_img
    B = double(65536-T_sorted{i}(1:nb_spots,1));
```

```
    options.alpha=0.99;
    res(i) = fastlts(A, B, options); % régression LTS
end
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
% égalisation des images
for i = 1:nb_img
    %les coeff ont été déterminés sur le complément des intensités donc :
    img{i} = double(imcomplement(img{i}));
    % on opère la correction LTS
    img{i} = uint16((img{i}-res(i).coefficients(2))/res(i).coefficients(1));
end
```

La Figure 49, à rapprocher de la Figure 47, montre les valeurs des 1ers et 9èmes déciles des populations des valeurs d'intensité des taches protéiques mesurées après la correction par transformation linéaires des niveaux de gris de chacune des images de l'expérience (juin 2006). Les 1ers déciles, tout comme les 9èmes sont maintenant voisins les uns des autres, ce qui permet de vérifier le rapprochement des distributions d'intensité des taches protéiques. L'effet de l'égalisation est également mis en évidence par l'observation des profils de niveaux de gris avant et après le traitement. Ainsi les profils de niveaux de gris de la Figure 50 montrent le rapprochement des lignes de fond ainsi que l'ajustement dynamique des taches protéiques. Comme l'atteste la Figure 51, cette égalisation permet à l'algorithme de fusion (décrit plus loin dans le paragraphe 3.3.2.2) d'assurer une meilleure représentativité des taches de faibles intensités. Il faut également noter l'amélioration générale du contraste des images.

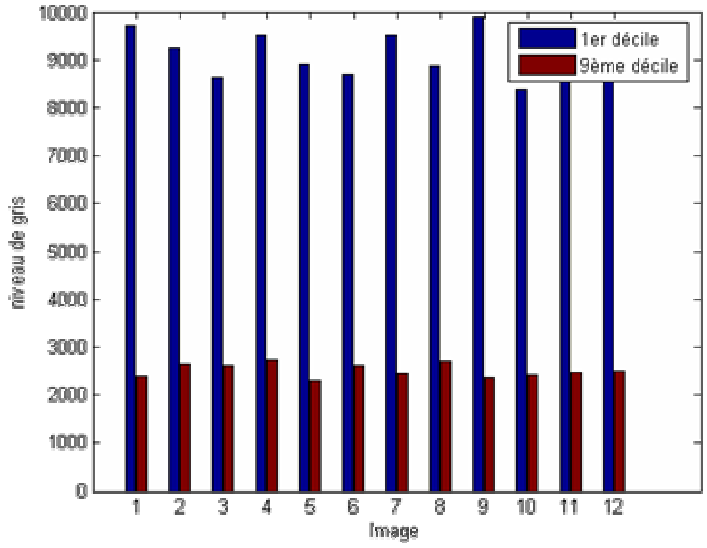


Figure 49: Visualisation, après égalisation, du 1er et du 9ème décile pour chacune des 12 populations d'intensité de taches protéiques correspondant aux 12 images de gel d'électrophorèse (expérience DIGE de juin 2006).

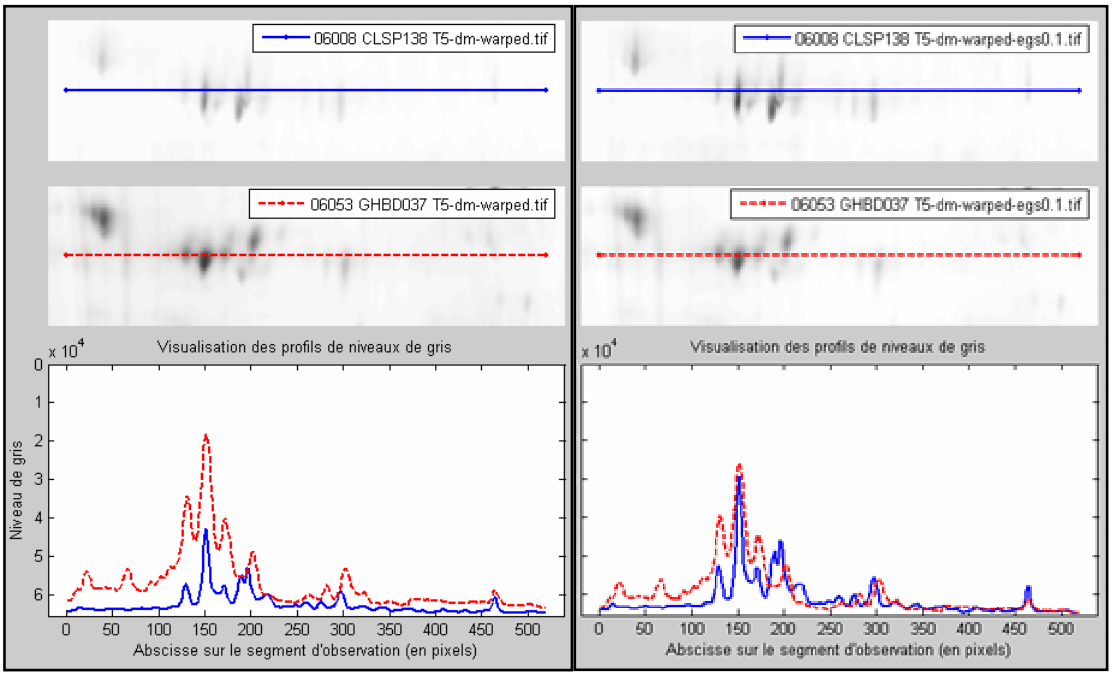


Figure 50 : Visualisation d'un profil de niveaux gris observé sur les images 06008T5 et 06053T5 avant (gauche) et après (droite) l'égalisation.

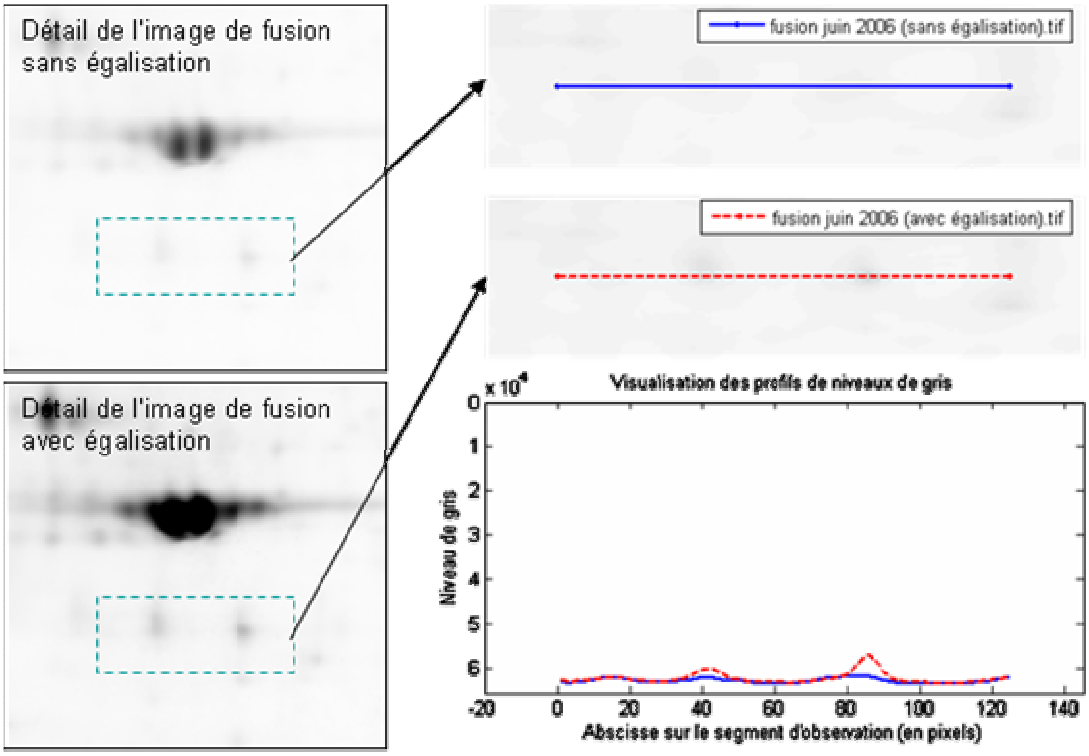


Figure 51: Amélioration de la représentativité de l'image de fusion. Les 2 taches protéiques dans le cadre en pointillés sont uniquement visibles sur l'image de fusion avec égalisation. Le bénéfice de l'égalisation est encore davantage mis en évidence par l'observation du profil des niveaux de gris de ces deux taches.

3.3.2.2 Algorithme de fusion des images

Une fois les images prétraitées et égalisées et afin d'exploiter le bénéfice de l'alignement des images de DIGE, il est nécessaire de produire une image « moyenne » à partir d'un ensemble d'images de départ. La manière d'employer cette image moyenne sera notamment débattue dans le paragraphe suivant. Dans le cadre de la protéomique, l'image « moyenne » ou « fusionnée » se doit d'être représentative de toutes les populations protéiques (les taches) visibles sur l'ensemble des images sources.

Soit $V(x,y)$ la valeur affectée au pixel positionné en (x,y) et considérant que plus la tache protéique est sombre plus la valeur $V(x,y)$ est importante. $V(x,y)$ est une valeur sur l'échelle de gris dont la dynamique est généralement de 16 bits : $V(x,y) \in [0,65535]$.

La première idée est d'affecter à chaque pixel $V(x,y)$ de l'image fusionnée, la moyenne des valeurs des pixels correspondants sur les n images sources :

$$V(x,y) = \frac{1}{n} \times \sum_{i}^{n} V_i(x,y)$$

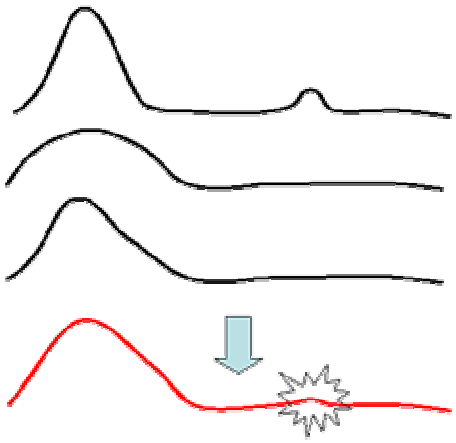


Figure 52 : Le profil de niveaux de gris en rouge est le résultat obtenu en moyennant les trois profils noirs. La moyenne des profils ne permet pas de conserver le plus petit des deux motifs pourtant clairement présents sur le premier profil noir.

L'image obtenue par cette simple moyenne arithmétique est visuellement très proche des images sources. Par ailleurs, la moyenne entraîne la suppression des divers bruits résiduels et l'image est visuellement agréable. Cependant, comme le montre la Figure 52, cette technique présente un défaut majeur : elle ne permet pas la représentativité de l'ensemble des taches protéiques visibles sur l'ensemble des images sources.

Afin de contrer cet effet de dilution des pixels de faible valeur, une deuxième idée est d'affecter à chaque pixel $V(x,y)$ de l'image fusionnée, la valeur maximale parmi les pixels correspondants sur les n images sources :

$$V(x,y) = \underset{i}{Max}(V_i(x,y))$$

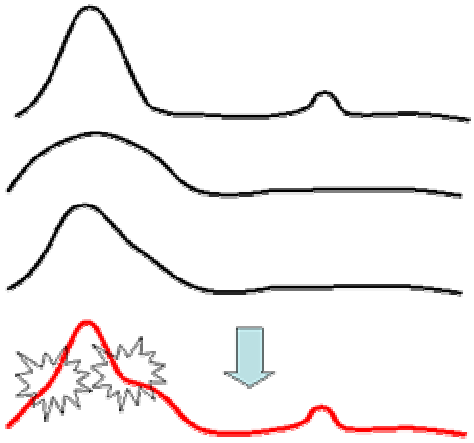


Figure 53 : Le profil de niveaux de gris en rouge est obtenu en conservant la valeur maximale prise sur les trois profils noirs. Cette technique permet de conserver le motif de faible intensité présent sur le premier profil. Cependant, elle entraîne des problèmes l'apparition d'épaulements indésirables lorsque les motifs présentent des formes différentes comme c'est le cas, ici, pour le motif de forte intensité.

L'image obtenue s'éloigne considérablement d'une image de gel d'électrophorèse bidimensionnelle classique. L'image est chargée et très bruitée. Certes l'ensemble des taches protéiques présentes sur l'ensemble des images source est bien présent mais la qualité de l'image en empêche l'exploitation. On observe notamment une distorsion des taches protéiques. Leur géométrie spatiale est très affectée avec notamment l'apparition d'épaulements artificiels indésirables (voir Figure 53).

La solution retenue est une solution intermédiaire qui doit permettre une bonne représentativité tout en conservant une géométrie spatiale des taches protéiques proche de la réalité. Pour parvenir à ce compromis, nous avons mis en place le calcul d'une moyenne pondérée. Ainsi, on affecte à chaque pixel $V(x,y)$ de l'image fusionnée la moyenne pondérée des valeurs des pixels correspondants sur les n images sources. La pondération est la valeur même du pixel considéré. Ainsi, même une tache de faible intensité apparaîtra sur l'image fusionnée puisque les autres valeurs contribuant au calcul seront encore plus faiblement pondérées.

$$V(x,y) = \frac{\sum_{i}^{n} (V_i(x,y) \times V_i(x,y))}{\sum_{i}^{n} V_i(x,y)}$$

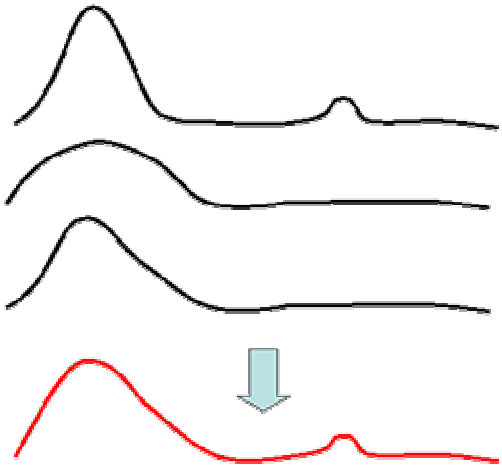


Figure 54 : Le profil de niveaux de gris en rouge est le résultat obtenu en opérant une moyenne pondérée des trois profils noirs. Cette technique permet de conserver le plus petit des deux motifs présents sur le premier profil noir tout en préservant l'aspect naturel des spots (pas d'épaulement indésirable).

Un extrait de l'implémentation de cet algorithme sous Matlab est donné ci-dessous :

```
function fusion_selection(selection_filename, selection_pathname,nom_fusion)
nb_img = size(selection_filename,2);
for i = 1:nb_img
  img{i} = imread([selection_pathname selection_filename{i}],'TIFF');
end
M = cell2mat(img);
h = waitbar(0,'Fusion: Please wait...');
for i=1:size(img{1},2)
  img_fusion(:,i) = sum((65536-
double(M(:,i:size(img{1},2):end))).*double(M(:,i:size(img{1},2):end)),2)./sum(65536-
double(M(:,i:size(img{1},2):end)),2); % moyenne pondérée
  waitbar(i/size(img{1},2),h);
end
close(h)
imwrite(uint16(img_fusion),[selection_pathname nom_fusion],'TIFF', 'Compression','none');
end
```

L'image de fusion ainsi obtenue est un bon compromis qui permet une très bonne représentativité de l'ensemble des taches protéiques de l'ensemble des images sources tout en préservant l'aspect naturel des spots.

V.3.3.3Analyse des images : comparaison de différentes approches

Il existe différente manière d'employer l'algorithme de fusion mis en place dans le paragraphe précédent. Ce paragraphe se propose de comparer différentes approches stratégiques d'analyse des images afin de mettre en évidence l'intérêt de la fusion ainsi que son meilleur usage.

Concernant les quatre premières approches étudiées, les étapes d'alignement des images, de détection et de mise en correspondance des taches protéiques sont ré-

alisées à l'aide de logiciels commerciaux : TT900 de Non-Linear Dynamics (alignement) et ImageMaster 2D Platinium de Genebio (détection et mise en correspondance). Ces méthodes diffèrent selon la stratégie employée qui peut être est plus ou moins adaptée à la technologie DIGE. La dernière des 5 méthodes considérées correspond à l'usage d'un logiciel commercial dédié à la technologie DIGE : Progenesis de Non-Linear Dynamics.

La technologie DIGE permet en théorie de produire pour chaque gel un jeu de 3 images parfaitement alignées puisque ayant subi les mêmes biais expérimentaux. Afin d'exploiter cette particularité, une première idée consiste à utiliser une seule détection des spots, c'est-à-dire un seul et même « patron de détection » pour chacune des trois images d'un gel. Ceci doit permettre non seulement une mise en correspondance parfaite des spots d'une population à une autre au sein d'un même gel mais aussi une amélioration de la quantification qui est alors réalisée à partir de contours de spots rigoureusement identiques. Le potentiel lié à l'utilisation d'un pattern commun intra-gels nous a conduit à étendre cette idée inter-gels. Mais ressurgit alors le problème de l'alignement des images entre différents gels puisqu'ils subissent des biais expérimentaux différents. Une solution consiste à aligner préalablement l'ensemble des images d'une étude grâce au travail conduit par un expert et rendu possible par un logiciel adapté (TT900 de Non-Linear Dynamics).

Ces deux approches doivent être comparées entre elles mais doivent également être mises en concurrence avec l'approche classique sans pattern commun déclinée avec et sans alignement préalable des images. Enfin, de la même manière, le logiciel Progenesis de Non Linear Dynamics sera évalué, afin de situer nos différentes approches par rapport à un logiciel commercial dédié.

Au final et pour résumer, l'étude comparative présentée ici concerne 5 approches différentes :

- Approche classique sous ImageMaster, sans alignement préalable (« csw »)
- Approche classique sous ImageMaster, avec alignement préalable (« caw »)
- Approche patron commun intra-gel avec alignement préalable (« pc1 »)
- Approche patron commun inter-gels avec alignement préalable (« pc2 »)
- Approche logiciel commercial: Progenesis (« pg »)

3.3.3.1 Présentation des approches concurrentes

Les différentes approches se distinguent par le degré de généralisation du « patron de détection ». L'approche classique consiste à utiliser un patron différent pour chacune des images de l'expérience tandis que l'approche par patron commun intra-gel profite de la particularité de la technologie DIGE afin de n'employer qu'un seul et même patron pour les images d'un même gel. L'approche par patron commun inter-gels généralise ce concept à toutes les images d'une expérience grâce à un alignement préalable des images. Cet alignement, présenté en Figure 55, est supervisé par un expert et conditionne la qualité de la suite des analyses.

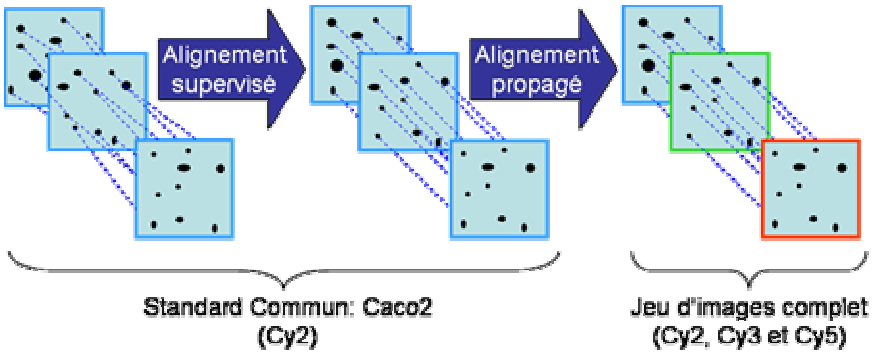


Figure 55 : Alignement de l'ensemble des images d'une expérience de DIGE : l'expert se contente d'aligner les images du standard commun et le logiciel (TT900 de NonLinar Dynamics) en déduit l'alignement pour le reste des images.

La Figure 56, la Figure 57 et la Figure 58 illustrent les principes de ces trois approches que nous détaillons davantage dans les paragraphes suivants.

a) Approche Classique :

L'approche classique est réalisée grâce au logiciel ImageMaster dans sa version dédiée aux gels d'électrophorèse bidimensionnelle classiques tels que les gels à l'argent. Cette version n'exploite donc pas les avantages liés à la technologie DIGE : les images sont considérées indépendantes les unes des autres. Elles sont traitées individuellement jusqu'à l'étape de mise en correspondance des spots (« matching »).

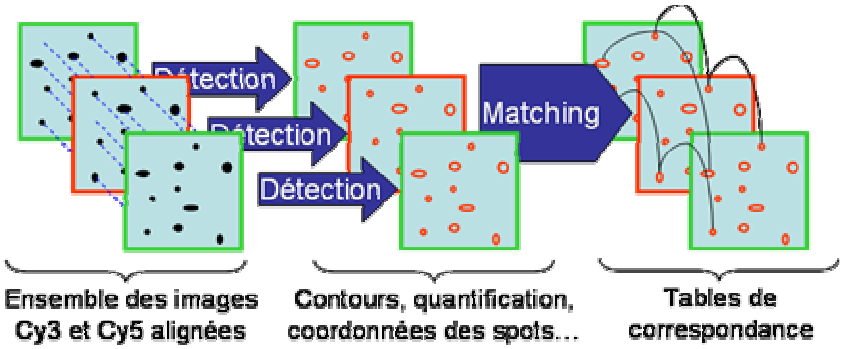


Figure 56 : Approche classique enrichie de l'étape d'alignement des images

Nous déclinons cette approche de deux façons selon que le traitement se fasse à partir des images brutes ou bien à partir des images préalablement alignées via le logiciel TT900. La Figure 56 illustre seulement ce deuxième cas. Le reste du procédé, basé sur l'usage exclusif du logiciel ImageMaster, se compose d'une détection et d'une quantification des spots sur chacune des 12 images de l'expérience ainsi que d'une étape de mise en correspondance des spots de toutes les images de l'expérience.

b) Approche par patron de détection commun intra-gel :

Comme l'illustre la Figure 57, l'approche par patron de détection commun intra-gel est similaire à l'approche classique si ce n'est que le patron de détection pour les deux images d'un même gel est déterminé à partir de la moyenne de ces deux images. Le calcul de cette image moyenne nécessite, au préalable, d'avoir ajusté les niveaux de gris

des deux images concernées. De cette manière, à quantités protéiques égales, deux taches d'un même groupement protéiques apportent une même contribution à l'image moyenne calculée en prenant simplement la moyenne des niveaux de gris, pixel par pixel, de ces deux images. L'ajustement est réalisé par ré-étalement de la dynamique : un certain centile dans les niveaux de gris faible ainsi qu'un autre correspondant aux niveaux élevés sont confondues aux limites extrêmes de la dynamique. Par ailleurs ces valeurs extrêmes sont déterminées sur une portion de l'image excluant les bords sources de valeurs aberrantes pouvant fausser la mise à niveau.

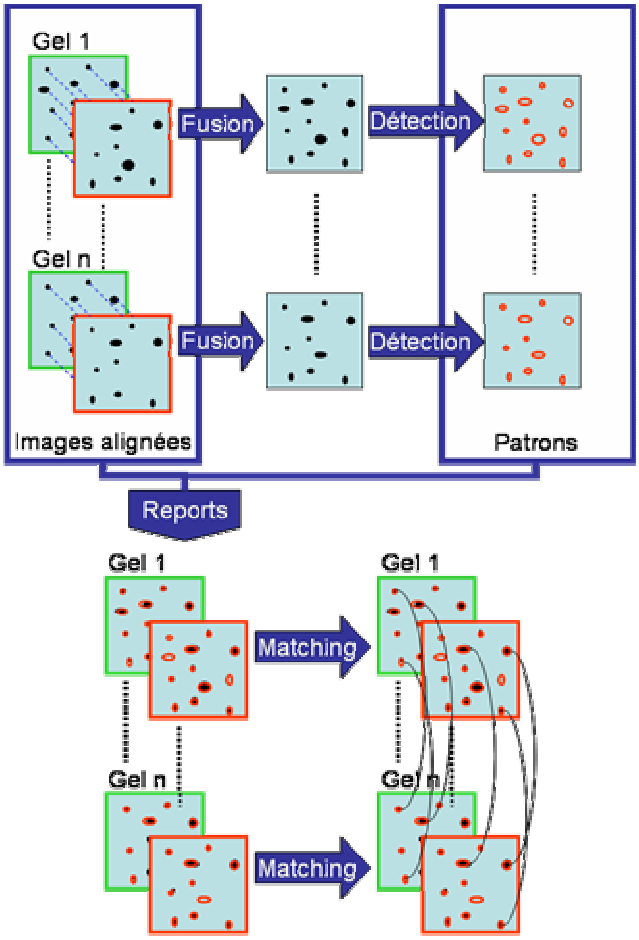


Figure 57 : Approche « patron de détection » commun intra-gel

L'image moyenne sert exclusivement à la détermination d'un patron commun. Ce patron est reporté sur chacune des deux images non ajustées, l'ajustement se faisant, par la suite, par normalisation des données.

c) L'approche par patron de détection commun inter-gels :

Cette approche (Figure 58) exploite non seulement l'alignement implicite lié à l'emploi de la technologie DIGE, mais également l'alignement supervisé par l'expert et réalisé grâce au logiciel TT900 de Non-Linear Dynamics.

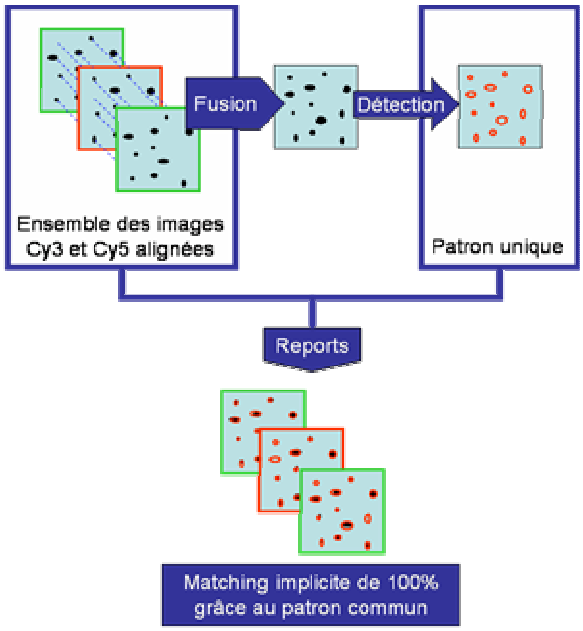


Figure 58 : Approche « patron de détection » commun inter-gels

Toutes les images de l'expérience étant ainsi alignées, il est possible, de la même manière que précédemment, d'établir leur moyenne. Sur l'image moyenne est représenté un grand nombre de taches protéiques. Seules les taches d'intensité faible ou bien présentes sur une minorité d'images sont susceptibles de ne pas apparaître sur l'image moyenne. Mais cela n'est pas un problème puisqu'il s'agit alors de groupements protéiques ne présentant pas un intérêt majeur dans le cadre de l'étude différentielle à laquelle est destiné ce jeu de données.

d) Approche par logiciel dédié (Progenesis) :

La dernière approche à laquelle nous nous sommes intéressés concerne l'utilisation du logiciel Progenesis, adapté à la technologie DIGE et qui utilise, lui aussi, un patron commun à toutes les images d'une expérience. Il s'agit de la méthode nommée « SameSpots » dans le logiciel. Le principe est le même que celui adopté pour notre approche de patron commun inter-gels. Cependant, les méthodes employées pour l'obtention du patron commun ne sont pas communiquées. À la vue des images, il apparaît assez clairement que la détection de ce patron est effectuée par un algorithme du type « ligne de partage des eaux ». Afin de juger de l'efficacité de cette approche, nous effectuons la même étude avec le même jeu de données, avec ce logiciel.

3.3.3.2 Principe de la comparaison

L'objectif des différentes approches comparées ici est d'obtenir, à partir des données brutes (images), le plus grand nombre de quantifications de groupements protéiques différents et donc le minimum de données manquantes sur chacune des images. Bien sûr, ces quantifications doivent avoir un sens biologique, elles doivent être l'image des quantités protéiques réelles ou tout au moins, les ratios établis entre les deux classes étudiées (tumeur et muqueuse) doivent se rapprocher des ratios réels.

Afin d'établir notre comparaison, il est donc nécessaire de déterminer des indicateurs de qualité qui permettent de juger ces critères.

Le nombre de groupements protéiques pris en compte et les données manquantes associées seront mesurées par le nombre de spots détectés sur chacune des images, associé au pourcentage de spots ayant pu être mis en correspondance avec ceux d'une image de référence commune à toutes les approches.

Chacune des approches se compose nécessairement d'une étape de détection des spots sur une ou plusieurs images ainsi que d'une étape de mise en correspondance des spots. Il faut cependant noter que pour l'approche par patron commun inter-gels et pour l'approche Progenesis, la mise en correspondance est implicite et parfaite (sous réserve que l'alignement des images par l'expert biologiste ne comporte pas d'erreur). Afin de légitimer la comparaison il est nécessaire de comparer des détections aboutissant à des patrons similaires en terme de nombre de taches protéiques détectées. Pour ce faire, les différentes détections sont supervisées par l'opérateur de sorte que le nombre de taches, une fois celles aberrantes (taches protéiques situées sur les bords du gel, défauts de l'image ainsi que les taches d'intensité ou de surface très faibles) supprimées, soit comparable d'un gel à l'autre (fourchette de 1000 à 1300 spots par image). La mise en correspondance des taches est réalisée par rapport à une même image de référence (image 06009 M) quelle que soit l'approche considérée. Cette image a été choisie comme référence en raison de la richesse et de la qualité de la détection de ses taches protéiques. Une vingtaine de « landmarks » a été placée afin de faciliter la mise en correspondance des spots quand cela est nécessaire. En effet, les « landmarks » sont le résultat d'une mise en correspondance, supervisée par l'expert, d'un certain nombre de spots à travers toutes les images d'une expérience. Ils ont pour but de servir de germes à l'algorithme d'ImageMaster et donc d'en améliorer la qualité.

L'approche basée sur l'utilisation de Progenesis est guidée par le logiciel. Parmi les options proposées à l'opérateur certaines ont été choisies différentes de la configuration par défaut. Afin de considérer une approche sans donnée manquante, nous imposons que les spots soient présents dans tous les gels. La soustraction du bruit est faite suivant la méthode : « Mode of non spot » en prenant une marge de 45 pixels. La normalisation est réalisée à partir du volume total des spots.

À l'issue de différentes étapes qui composent les différentes approches, nous disposons des informations sur la qualité des détections ainsi que sur les quantifications et les appariements associés à cette première approche. Afin de s'assurer de la validité biologique des quantifications nous nous plaçons dans le cadre d'une analyse différentielle. Les quantifications brutes sont normalisées, c'est-à-dire que les transformations affines nécessaires à l'obtention de distributions de probabilité pseudo-normales centrées et réduites sont appliquées aux volumes correspondant aux populations protéiques de chacune des images. Il est alors intéressant d'observer la dispersion des logarithmes base deux des ratios calculés à partir des quantifications normalisées. Comme nous l'avons vu précédemment, afin de tirer le bénéfice de la DIGE (et ainsi d'obtenir une « normalisation naturelle » des ratios), ces ratios sont calculés à partir des deux quantifications issues des deux images Cy3 et Cy5 d'un même gel,

avec, à chaque fois, la valeur dans la classe « Muqueuse » en dénominateur. Pour chaque gel, le nuage de points des ratios en fonction de l'intensité moyenne des paires correspondantes doit être centré sur 1. Une expertise est ensuite menée afin de critiquer les spots singuliers tels que ceux situés aux limites du nuage de points. Ils doivent correspondre à des sur-expressions ou des sous-expressions confirmées par l'expert ou bien à des biais expérimentaux clairement identifiés.

3.3.3.3 Jeu de données utilisé pour la comparaison

Le jeu de données utilisé pour la comparaison des différentes approches est issu de l'expérience DIGE de janvier 2006 qui concerne trois patients fournissant chacun un échantillon de tissu sain et un autre de tissu tumoral. Ces deux classes de population protéique sont couplées à deux cyanines différentes notées Cy3 et Cy5. Une troisième classe, issue d'une lignée Caco2, nous sert de référence interne et est couplée à une cyanine notée Cy2. Pour chacun des patients, deux gels sont réalisés de manière à pouvoir permuter le couplage (Cy3 et Cy5). Le jeu de données est donc constitué de 18 images issues de 6 gels différents. L'objectif de l'étude étant une comparaison de méthodes, nous pouvons considérer les 6 gels comme indépendants (nous n'exploitons pas l'information apportée par la répétition avec permutation du couplage).

		Couplage Cy2	Couplage Cy3	Couplage Cy5
Patient 1 CLSP138	Gel 06008	Lignée Caco2	Muqueuse	Tumeur
	Gel 06009	Lignée Caco2	Tumeur	Muqueuse
Patient 2 CLSP086	Gel 06051	Lignée Caco2	Muqueuse	Tumeur
	Gel 06052	Lignée Caco2	Tumeur	Muqueuse
Patient 3 GHBD037	Gel 06053	Lignée Caco2	Muqueuse	Tumeur
	Gel 06054	Lignée Caco2	Tumeur	Muqueuse

Tableau 6: Plan d'expérience correspondant au jeu d'images utilisé pour la comparaison des méthodes

Les images issues de la lignée Caco2 permettent de faciliter l'alignement des images inter-gels lorsque la méthode considérée inclut cette étape. Les autres étapes de traitement - la détection et la mise en correspondance des spots - ne concernent que les canaux Cy3 et Cy5 (populations saine et tumorale). Finalement, les données quantifiées telles que le nombre de spots détectés, les pourcentages de matching ou bien les volumes des spots sont issues de 12 images réparties en deux classes et appariées suivant le gel dont elles sont issues (voir Tableau 6). En effet, deux images d'un même gel correspondent au même patient et ont été affectées par les mêmes biais expérimentaux : les quantifications d'un même groupement protéique sur ces deux images sont biaisées de la même manière, ce qui signifie que, pour une modélisation linéaire des intensités et après normalisation de chaque population, le ratio des deux quantités se rapproche du ratio réel.

Le choix de ce jeu de données est justifié par la qualité des images acquises. En effet, ces images ne sont affectées que modérément par les biais classiquement observables sur les gels obtenus en routine. Ainsi, les bruits de haute fréquence spatiale sont modérés et la gamme dynamique du scanner est relativement bien exploitée.

Ce jeu de données fait partie des meilleurs disponibles en terme de qualité d'image (sans juger de l'intérêt et des significations biologiques liés à son interprétation). Par ailleurs, ce niveau de qualité est mis en évidence par la détection réalisée grâce à ImageMaster (paramètres par défaut) qui fait apparaître une population protéique riche avec des contours de spots bien définis ainsi qu'une résolution efficace des chevauchements.

3.3.3.4 Résultats et discussion

L'alignement, par l'expert, d'un ensemble d'images sur une image de référence a pour but de contourner les problèmes de distorsion induits par les disparités des conditions expérimentales de migration. Pour réaliser ce travail, l'expert identifie des motifs communs entre deux images et prend la décision de les confondre ou non suivant le contexte, son expérience et ses connaissances biologiques. Il sera par notamment capable d'identifier un décalage de spots provoqué par des modifications post-traductionnelles (phosphorylation, glycosylation, etc..). Ce type de décalage ne doit pas impacter sur l'alignement, car il a un sens biologique. À l'issue de ce travail d'alignement et dans l'idéal, la superposition des images prend tout son sens : à une localisation précise correspond autant de quantifications d'un groupement protéique que de nombre d'images considérées pour l'alignement. Le bénéfice de l'alignement peut être de deux sortes suivant l'approche adoptée.

Dans le contexte d'une approche classique pour laquelle une détection des spots est effectuée pour chacune des images alignées, le bénéfice de l'alignement se traduit dans l'étape de mise en correspondance des spots qui doit se voir sensiblement améliorée. En effet, l'alignement réalisé, même imparfait, rapproche les images de la situation idéale en supprimant les déformations géométriques et améliore donc l'efficacité de la reconnaissance de motifs de l'algorithme dédié.

Dans le cadre d'une approche de type « patron commun », il s'agit de considérer l'alignement comme étant quasi idéal et de déterminer un seul et même patron pour l'ensemble des images alignées. La mise en correspondance des spots est alors implicite et parfaite. Dans l'idéal, l'usage du patron commun est donc une solution élégante au problème des données manquantes et permet également une quantification des spots plus juste car faite à partir de contours identiques. Malheureusement, malgré tous les efforts de l'expert, certaines zones de certaines images ne peuvent pas être parfaitement alignées car trop dissemblables à l'image de référence. En effet, la faiblesse du signal acquis dans certaines zones de certaines images, la grande disparité dans les populations protéiques, l'importance des déformations ou bien encore la présence de bruits (poussières, cassures, traînées,…) sont autant de causes potentielles empêchant le rapprochement de motifs par l'expert. Le mauvais alignement résulte alors en un décalage du patron commun sur certaines zones de l'image et les quantifications des spots présents sur ces zones sont faussées : les contours qui leur sont associés ne sont pas centrés sur eux.

Les différentes approches présentent chacune des avantages et des inconvénients que les indicateurs de qualité mis en place précédemment permettent de mesurer.

La première information est apportée par la mesure du pourcentage de spots ayant pu être mis en correspondance. Ces mesures sont accompagnées d'une estimation de la proportion de faux positifs, c'est-à-dire la proportion des spots mis en correspondance de manière erronée. L'estimation du taux d'erreur de mise en correspondance des spots a été réalisée par un comptage mené par l'expert sur une fenêtre dont l'aire et la localisation sont identiques quelles que soient l'image et l'approche considérées. Cette estimation permet de légitimer la comparaison des pourcentages d'appariements. Les valeurs obtenues pour ces indicateurs sont rapportées dans les tableaux regroupés dans l'annexe A et sont résumées par la Figure 59 et la Figure 60.

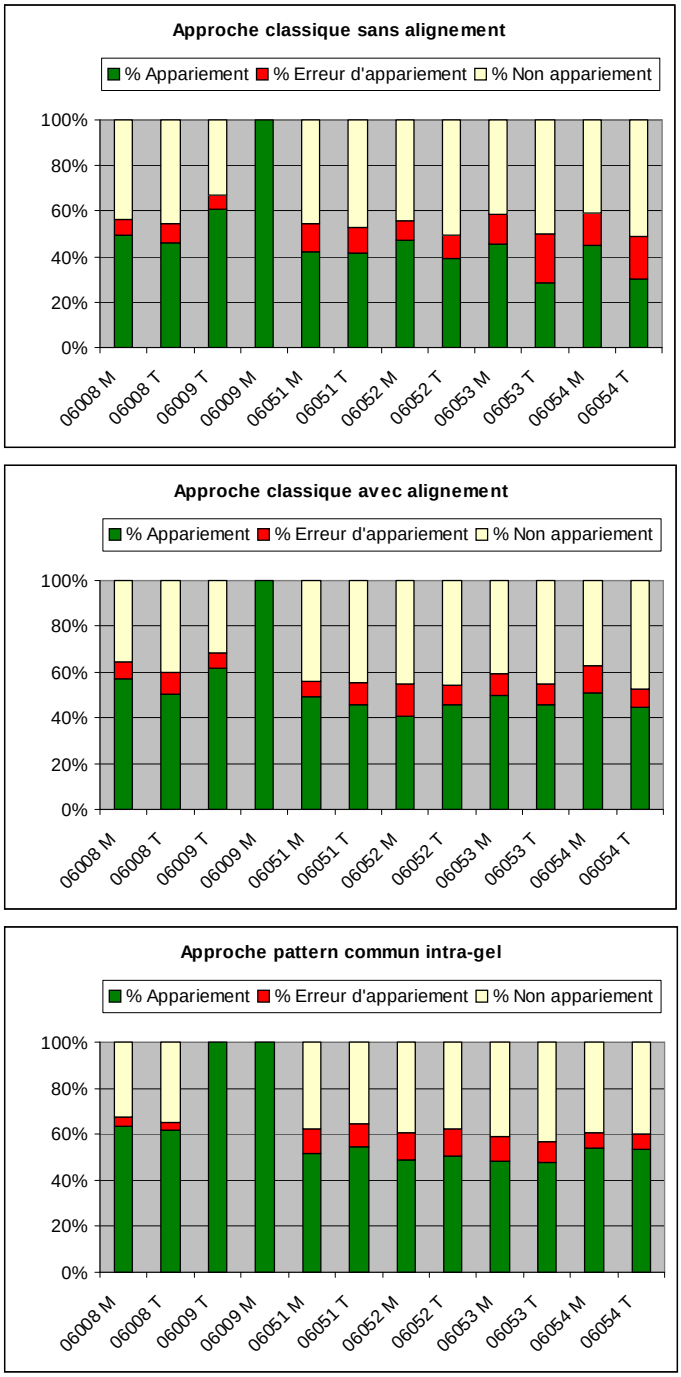


Figure 59: Proportions des taches protéiques détectées ayant été appariées correctement, incorrectement et n'ayant pas été appariées pour les approches classiques avec et sans alignement ainsi que pour l'approche avec patron commun intra-gel (les appariements à 100% correspondent aux gels images ayant servi de référence pour le matching).

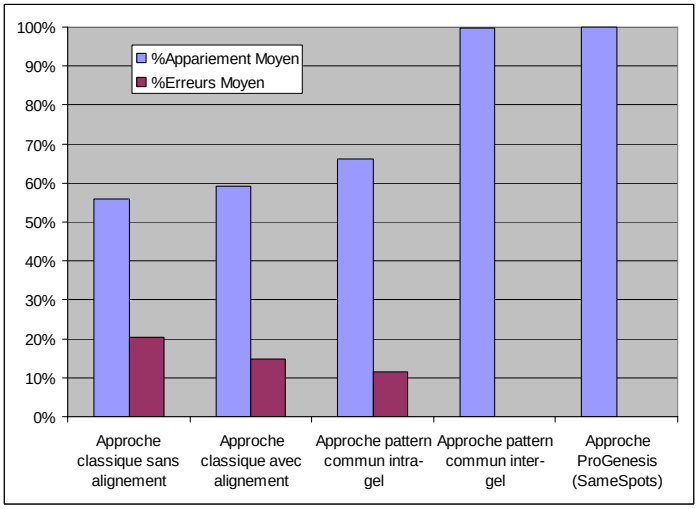


Figure 60: Proportions moyennes des taches détectées ayant été appariées ainsi que l'erreur associée pour chacune des 5 approches étudiées

Les appariements sont réalisés avec l'image 06009 M comme image de référence quelle que soit l'approche. Le premier constat concerne l'impact de l'alignement préalable des images sur la performance de l'algorithme de mise en correspondance des spots d'ImageMaster. Les représentations de la Figure 59 relatives à l'approche classique avec et sans alignement préalable, montrent une amélioration légère mais quasi systématique du pourcentage de mise correspondance en faveur de l'approche avec alignement et ce, pour des taux d'erreur globalement inférieurs. Par ailleurs, l'approche par patron commun intra-gel améliore encore ce pourcentage ainsi que le taux d'erreur associé. Enfin, comme l'illustrent les tableaux donnés dans l'annexe A, les approches par patron commun inter-gels telle que celle mise en place au sein du workflow IDADIGE (décrit plus loin dans la partie I) et celle proposée par le logiciel Progenesis, permettent une mise en correspondance parfaite des spots. Bien entendu, ce critère n'est pas le seul à prendre en compte, sans quoi les conclusions seraient simples et logiquement en faveur de ces dernières approches.

La quantification d'un même spot peut varier suivant l'approche adoptée. Lorsque le patron est défini à partir de l'image considérée, les quantifications sont faites à partir de contours adaptés à la morphologie de chaque spot. Ces contours sont alors toujours parfaitement centrés sur les spots mais un inconvénient est la trop grande souplesse dans la définition de ces contours qui peuvent être très larges autour des spots de faible intensité. Comme l'illustre la Figure 61 pour le spot « 13 », les deux versions de l'approche classique, produisent des contours mal définis et trop larges, ce qui conduit à une surestimation des volumes. Les contours issus des approches par patron commun sont plus réguliers car les images moyennes, sur lesquelles les détections sont effectuées, sont moins bruitées.

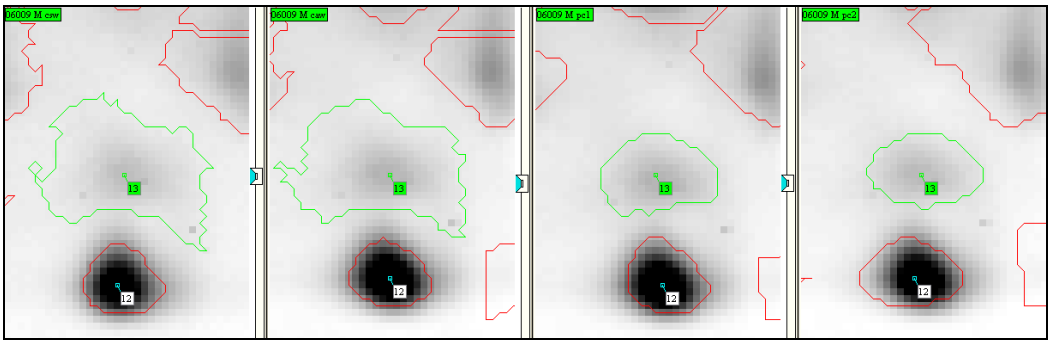


Figure 61: Visualisation des détections logicielles (ImageMaster) obtenues pour le spot 13 suivant 4 approches différentes. De gauche à droite : l'approche classique, l'approche classique avec alignement préalable, l'approche par patron commun intra-gel puis par patron commun inter-gels (approche adoptée pour le workflow IDADIGE mis en place dans cette thèse).

Ainsi, l'avantage d'un patron commun est que la quantification d'un spot est réalisée à partir de contours rigoureusement identiques d'une image à l'autre. Par contre, les imperfections d'alignement entraînent parfois un décalage entre les spots et leurs contours et elles induisent donc des erreurs de quantification. Ce biais est observé sur le spot 7 de la Figure 62. Les contours des deux approches classiques sont naturellement bien adaptés à la morphologie du spot puisque obtenues sur les mêmes images. De même, le contour de l'approche par patron commun intra-gel est bien calé sur le spot. En effet, l'image moyenne, d'où est issue le patron, est formée à partir de deux images parfaitement alignées puisque issues du même gel. Le contour de l'approche par patron commun inter-gels n'est pas calé sur le spot. Ceci traduit soit que cette région est mal alignée entre les différentes images qui composent l'expérience soit que le décalage est d'ordre biologique et correspond, par exemple, à une modification post-traductionnelle de la protéine concernée.

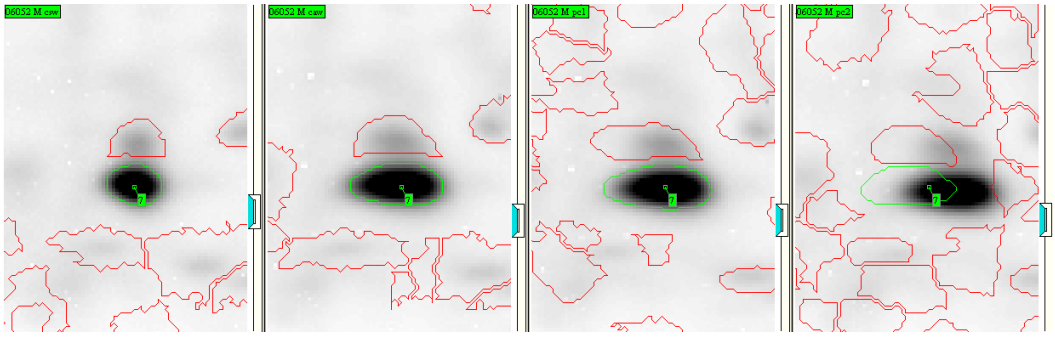


Figure 62: Visualisation des détections obtenues pour le spot 7 suivant 4 approches différentes. De gauche à droite : l'approche classique, l'approche classique avec alignement préalable, l'approche par patron commun intra-gel puis par patron commun inter-gels (approche adoptée pour le workflow IDADIGE mis en place dans cette thèse).

Afin de juger de la qualité des quantifications, nous avons établi un jeu de 48 groupes de taches protéiques témoins (voir Figure 63) pour lesquels nous disposons des quantifications pour chacune des cinq approches. Ces taches protéiques, réparties de façon homogène sur la surface des images, nous permettent en effet de critiquer l'impact des biais identifiés précédemment sur des cas précis et quantifiables. Pour cela, nous nous sommes intéressés, pour chacune des approches, à la dispersion du

ratio tumeur sur muqueuse de chacun des 48 groupes de taches protéiques. Il ressort de cette étude et des boites à moustaches dont certaines sont visibles sur la Figure 64 que l'usage d'un patron commun améliore grandement la stabilité des ratios. Il apparaît également que les résultats obtenus par notre méthode de patron commun inter-gels sont en adéquation avec ceux obtenus par l'intermédiaire du logiciel ProGenesis. Ce dernier présente les dispersions du ratio les plus réduites, mais il est à noter que ce qui semble être une qualité peut en réalité être dû à un biais de quantification. En effet, la détection des taches effectuées par ce logiciel conduit à l'obtention de contours larges. Ceci a pour effet d'englober dans la quantification une partie du fond environnant le spot et donc de noyer un peu l'information portée par la tache protéique. Au final, le logiciel Progenesis conduit à des quantifications légèrement pondérées par le bruit et donc à des ratios moins dispersés et dont la moyenne est plus proche de 1. Les ratios ainsi obtenus sont finalement moins significatifs que ceux établis à l'issue de notre approche par patron commun inter-gels

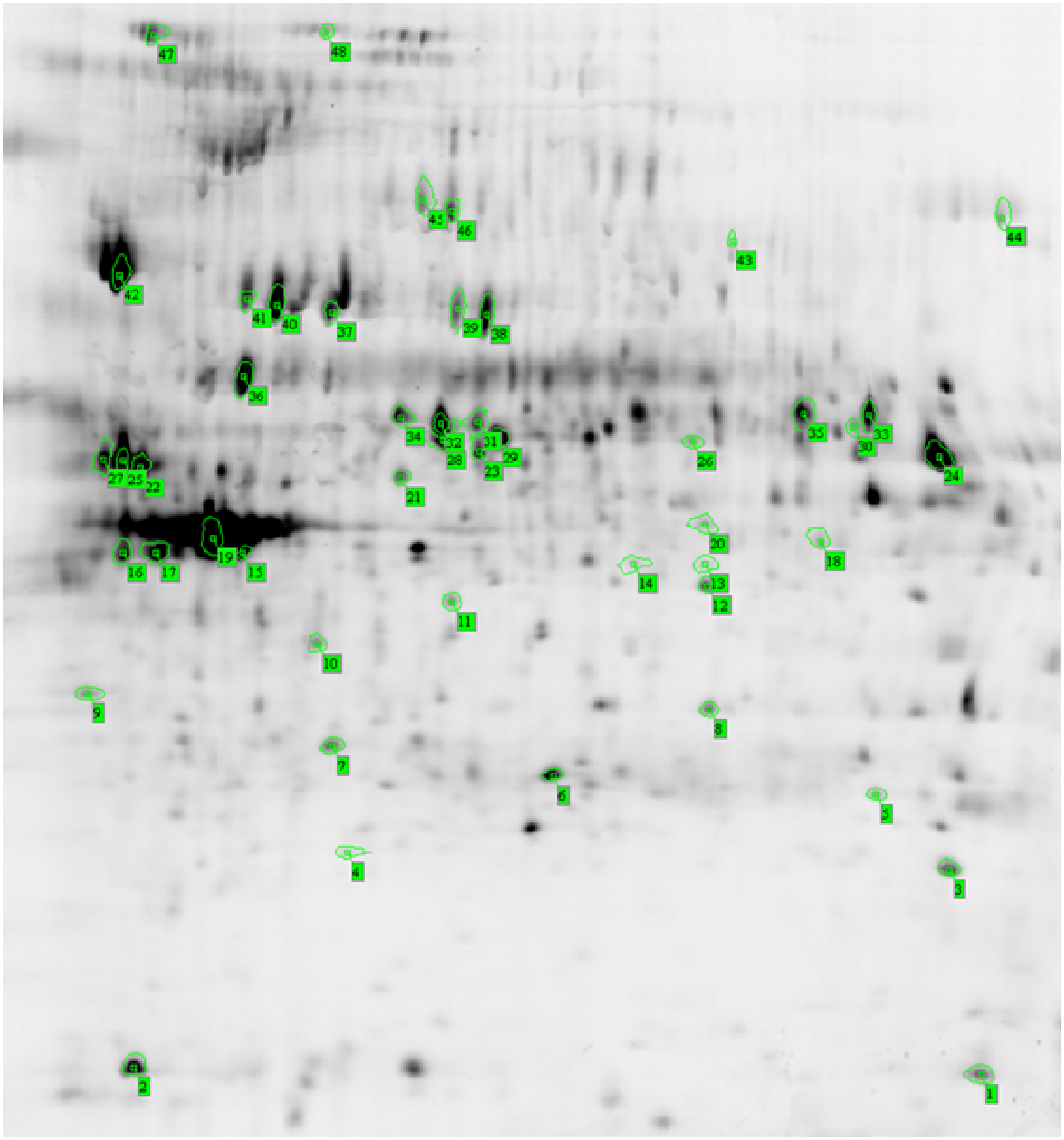


Figure 63: Localisation et identité des groupes de taches protéiques sur lesquels sont réalisées les observations

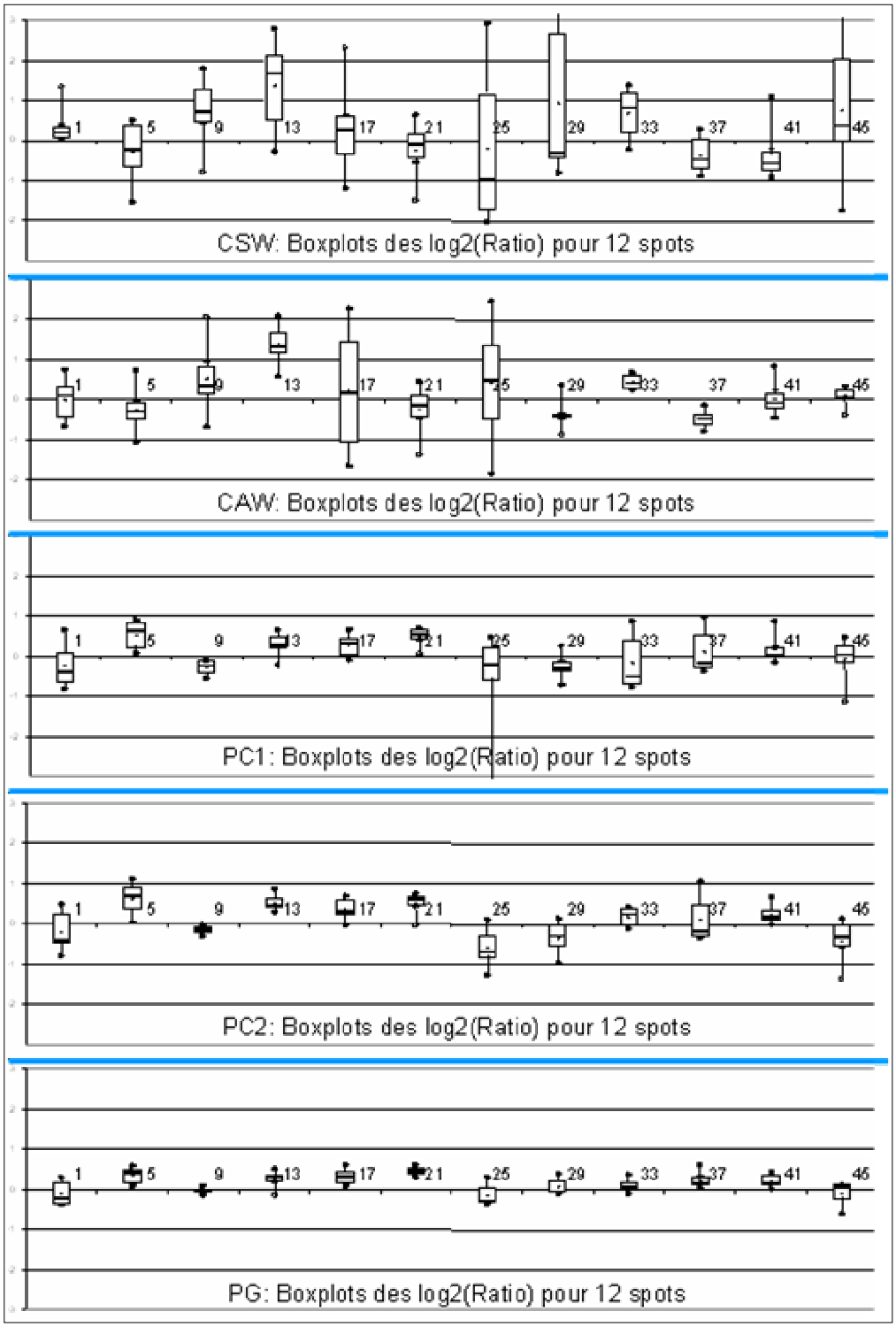


Figure 64: Boites à moustaches (définis) permettant, pour chacune des 5 approches étudiées, de visualiser la dispersion des ratios associés à 12 des 48 groupes de taches protéiques considérés.

V.3.3.4Normalisation des données

L'étape d'analyse des images issues d'une série de gels de DIGE permet l'obtention des données numériques brutes : les volumes des taches protéiques. Ces données reflètent les quantités des protéines contenues dans chacun des échantillons de l'expérience.

Cependant, ces données brutes ne sont pas directement exploitables pour l'interprétation. En effet, l'information qu'elles portent peut être masquée par les divers biais expérimentaux que le traitement d'images n'a pas pu supprimer.

Afin de déterminer la meilleure méthode pour la normalisation (voir partie IV.3.4.1), il est tout d'abord indispensable d'observer les données. Pour cela, nous avons choisi les données brutes issues d'une expérience menée dans des conditions typiques.

Un premier constat établit que les distributions du volume des taches protéiques suivent une loi lognormale (voir Figure 65) alors qu'une distribution normale serait beaucoup plus appropriée dans l'optique de mener des études statistiques et ainsi de les comparer entre elles. De plus, on constate (voir Figure 66) également que la variance du volume des taches protéiques est dépendante de sa moyenne.

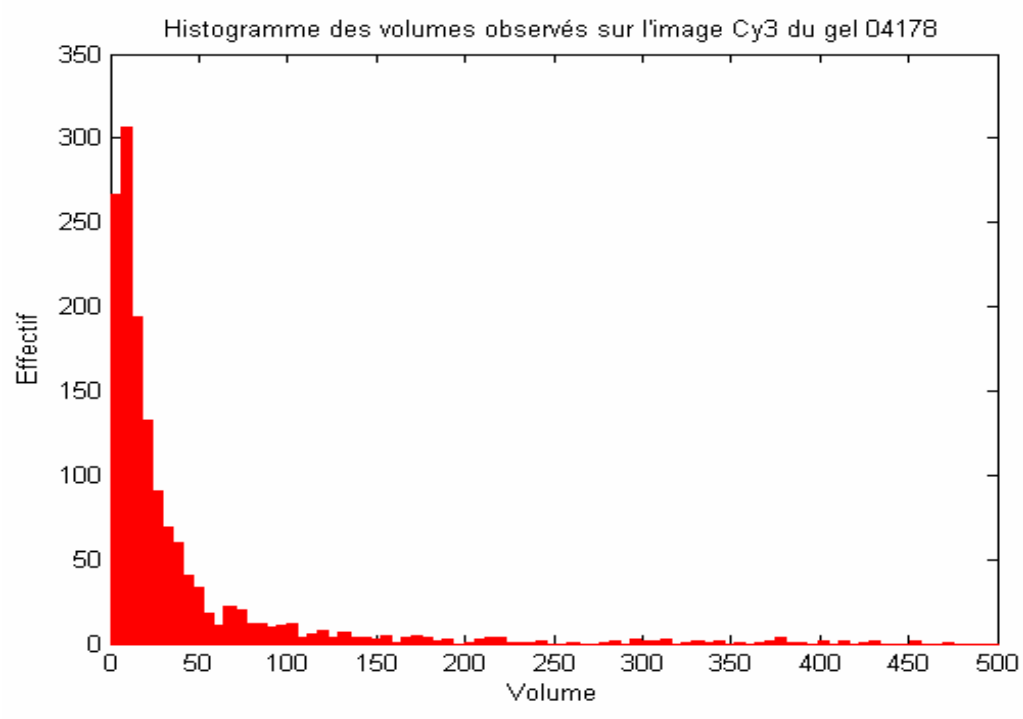


Figure 65 : Distribution classiquement observée des volumes des taches protéiques d'une images de gel DIGE

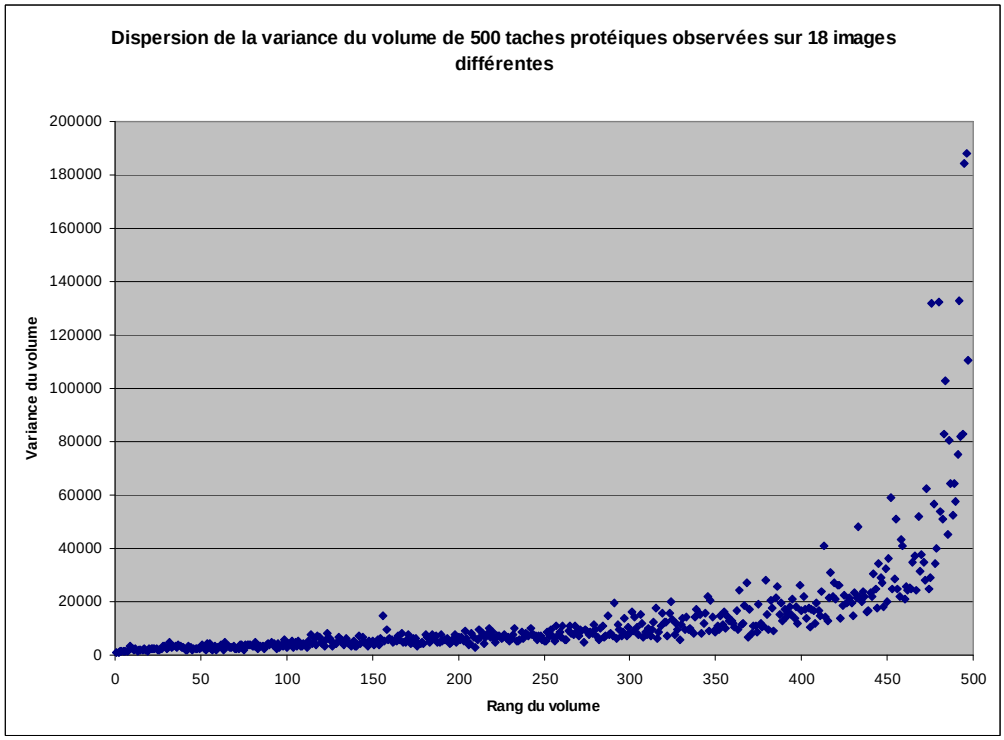


Figure 66 : Dispersion de la variance du volume de 500 taches protéiques observées sur 18 images différentes (jeu de données de Juin 2005)

Par ailleurs, même si les différents rendements quantiques des fluorophores utilisés n'impactent pas directement sur l'intensité des populations des taches protéiques (leur effet a été gommé par la maximisation de la dynamique lors de l'acquisition), il subsiste des écarts entre les distributions des populations. Ces écarts se traduisent par un décalage des tendances moyennes et par une différence dans la dispersion de ces distributions. Les données doivent donc être prétraitées afin de rapprocher les distributions d'une loi normale, de vérifier les hypothèses d'autocohérence et d'exploiter les éléments de contrôle que sont les répétitions croisées (dye-swap).

Dans un premier temps, afin de travailler sur des distributions normales, une transformation logarithmique est appliquée aux données (voir Figure 67). Cette transformation a également pour effet de stabiliser la variance, c'est-à-dire de la découpler de la valeur du volume moyen. En toute rigueur, comme le remarquent Karp et al. [32], il faut considérer non seulement un bruit multiplicatif mais également un bruit additif pour le modèle des volumes. Dans ce cas là, ce n'est pas une transformation logarithmique qu'il convient d'appliquer, mais une transformation sinus hyperbolique inverse (arcsinh). La différence entre les deux transformées n'est prononcée que pour les valeurs proches de zéro et disparaît rapidement lorsque le signal augmente. Comme le montre la Figure 68, pour les données relatives à nos expériences, les valeurs minimales des volumes ne justifient pas l'usage de la fonction sinus hyperbolique inverse. En effet, les résultats sont similaires, aux incertitudes expérimentales près, pour l'une et l'autre des transformées.

Par ailleurs, il est nécessaire de supprimer le signal de fond qui peut subsister après le traitement d'image. Pour ce faire, ce signal est identifié au 0.5^{ème} quantile et il est retranché à l'ensemble des données. Ceci doit être fait avant la transformation logarithmique des données car, du fait des propriétés du logarithme, ce signal de fond impacte sur la symétrie de la distribution des logarithmes des volumes (effet visible sur la Figure 67).

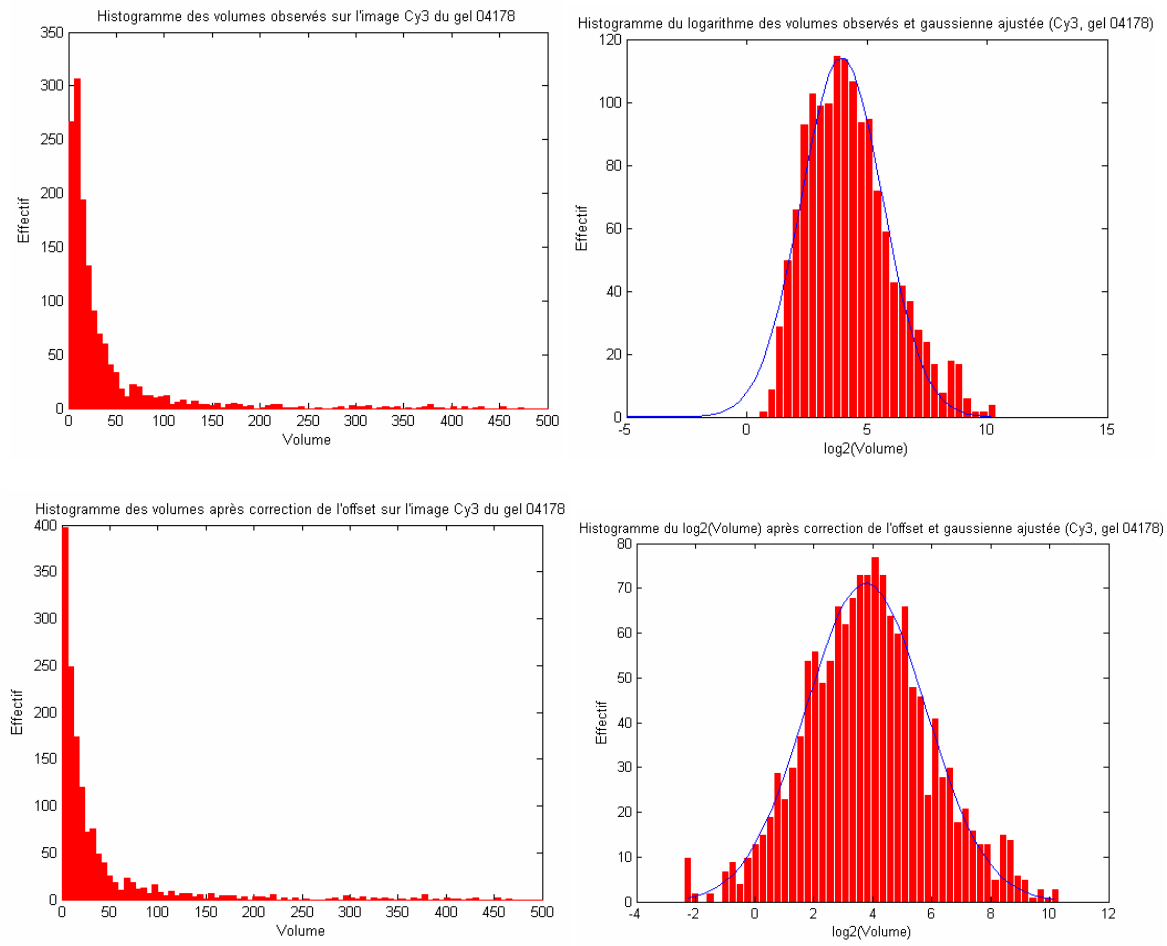


Figure 67: À gauche, la distribution des volumes bruts. À droite, la distribution des logarithmes des volumes. En haut, sans suppression de l'offset et en bas, avec suppression de l'offset. (Cy3, gel 04190).

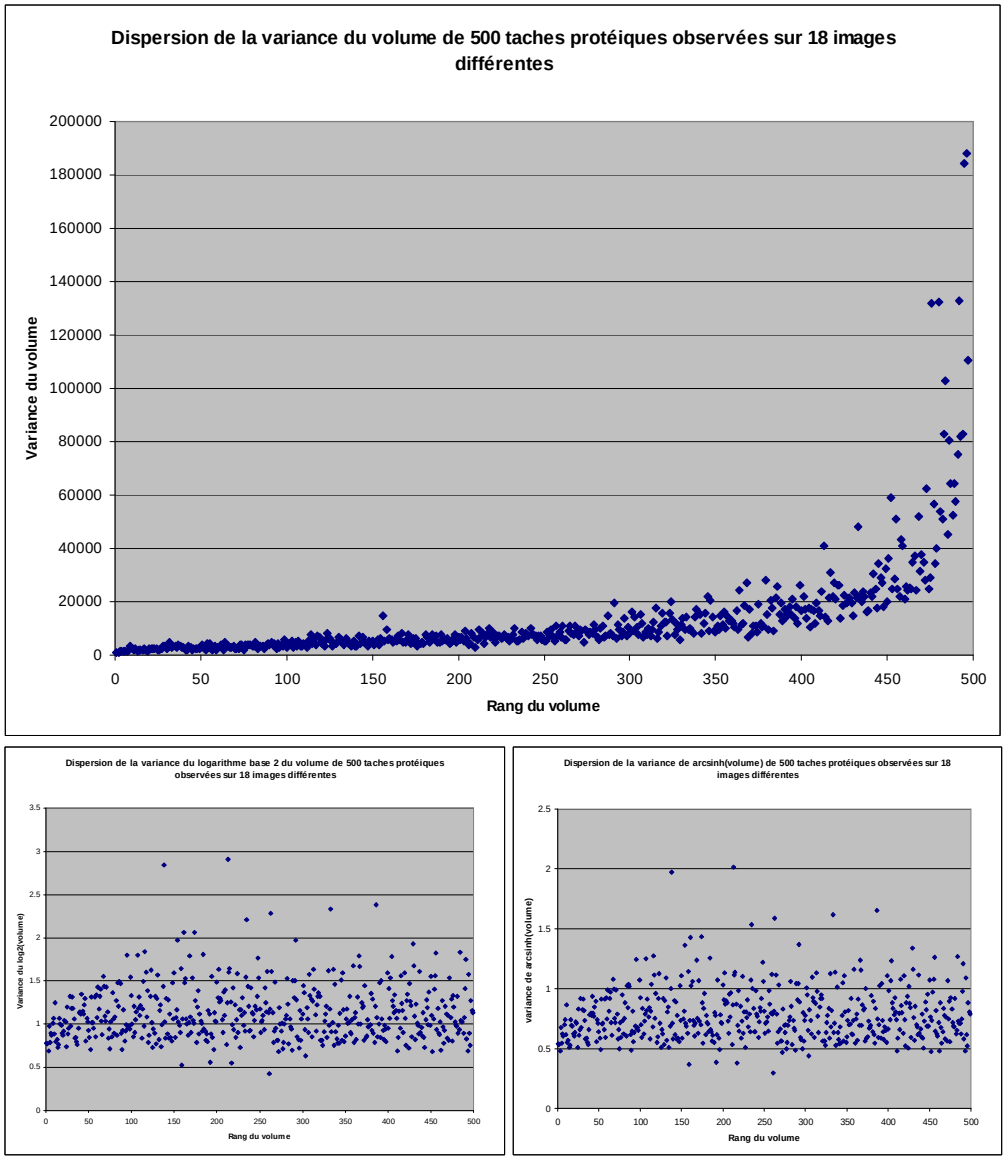


Figure 68: Dispersion de la variance de 500 taches protéiques observées sur 18 images différentes (jeu de données de Juin 2005), pour le volume brute (en haut), après transformation logarithmique base 2 (en bas à gauche) et après transformation sinus hyperbolique inverse (en bas à droite).

Une fois le signal de fond résiduel retranché et la transformation logarithmique appliquée, les distributions du logarithme du volume sont proches d'une distribution normale.

Ces transformations sont motivées par le souci de ramener les distributions de chaque population à un modèle connu et facilement interprétable. Il s'agit d'une simple relecture des données avec des « lunettes logarithmiques ».

L'étape suivante consiste en la suppression des différents biais qui affectent chacune des populations. Cette suppression est réalisée en se basant sur les hypothèses d'auto cohérence formulées à l'échelle des gels :

- Les distributions des volumes des spots sont semblables d'une population à l'autre
- La distribution des ratios volumiques est centrée sur 1 quelle que soit la gamme de volume considérée

En se référant aux distributions des logarithmes avec signal de base soustrait, il est évident que la première des deux hypothèses n'est pas vérifiée. Les biais liés à l'acquisition, au phénomène de blanchiment, mais aussi les différences de rendement quantique sont les principaux facteurs expérimentaux à l'origine des différences dans le centrage et la dispersion des différentes populations. L'hypothèse, basée sur des faits biologiques et physiques, selon laquelle les populations sont censées être semblables nous amène à travailler dorénavant sur les populations centrées réduites. Cette opération est réalisée à partir des paramètres de la gaussienne ajustée par une méthode des moindres carrés (méthode non linéaire de Levenberg-Marquardt [38]). Le résultat est visible sur la figure suivante.

La deuxième hypothèse d'auto-cohérence qui établit que la distribution des ratios volumiques est centrée sur 1 quelle que soit la gamme de volume considérée, fait intervenir, pour la première fois, la notion d'appariement des données. On ne considère plus chacune des populations indépendamment des autres mais par paires. Les paires considérées sont celles constituées par les deux populations Cy3 et Cy5 d'un même patient afin de conserver l'homogénéité en terme de biais expérimentaux et biologiques (même biais expérimentaux et même patient). L'observation, pour un gel, de la répartition des ratios en fonction du volume moyen des paires fait souvent apparaître des tendances générales s'opposant à l'hypothèse. Ce genre de phénomène, observé dans le domaine des puces à ADN est usuellement solutionné à l'aide d'une régression locale de type LOWESS [62].

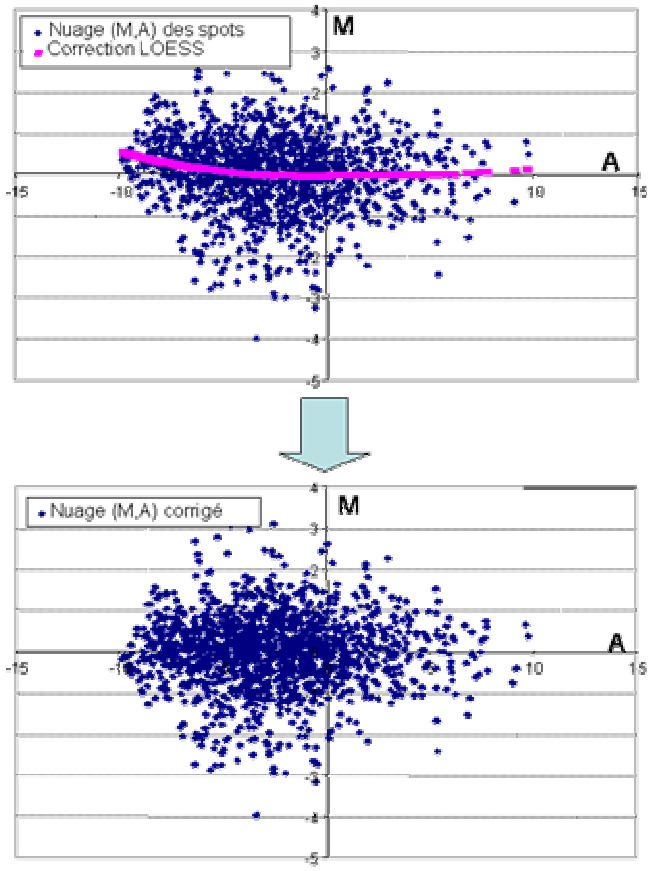


Figure 69: Correction LOWESS

Malheureusement, dans le cas des données DIGE, la densité des données sur les plages d'intensité extrêmes est faible et rend cette technique risquée dans le sens où la régression est faite sur un faible nombre de points et peut conduire à des corrections faussées. On préfère employer une technique non locale qui estime l'orientation générale du nuage de point. La technique utilisée est une régression linéaire LTS (Least Trimmed Squares : moindres carrés élagués). Par rapport à une régression des moindres carrés classique, il ne s'agit plus de minimiser la somme des carrés de toutes les erreurs, mais uniquement la somme des carrés des erreurs "les plus petites", ce qui rend la régression robuste aux données extrêmes et aberrantes. Géométriquement, cette méthode revient à trouver la bande la plus étroite contenant un pourcentage de points défini par l'utilisateur, ce qui, dans notre cas, permet d'estimer l'orientation du nuage.

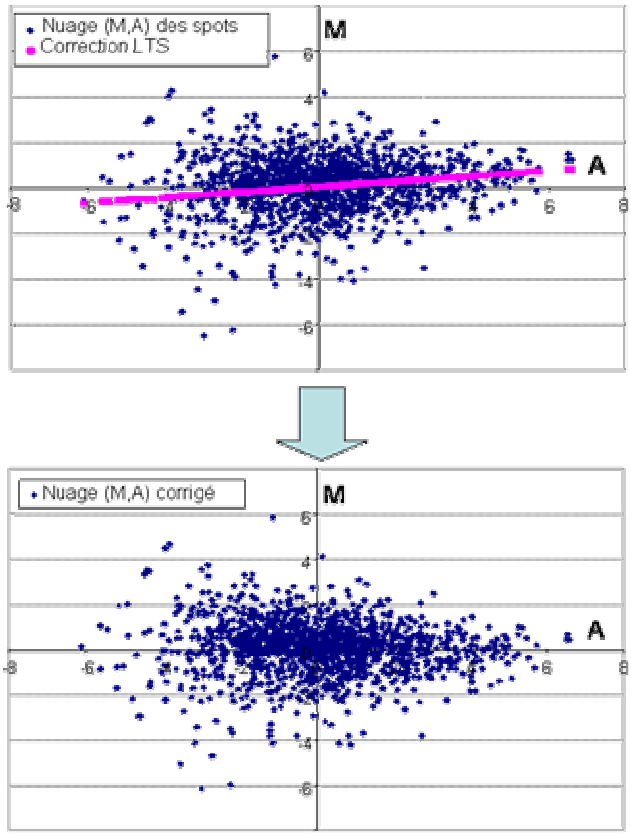


Figure 70: Correction LTS

À l'issue du recentrage du nuage de points, les ratios répondent aux hypothèses d'auto cohérence. Il est encore possible, en se fiant à des éléments de contrôle internes, d'affiner la correction des données. Il s'agit d'exploiter les répétitions avec inversion du fluorophore (« dye-swap ») selon la méthode décrite dans le paragraphe IV.3.4.1.2b). Pour rappel, cette méthode consiste à considérer, pour chaque tache protéique, la moyenne des deux ratios logarithmiques calculés sur les deux gels du « dye-swap », comme ratio logarithmique corrigé.

V.3.3.5Fouille des données

Le processus de traitement « ProDIGE » doit permettre d'aider les biologistes dans la recherche de marqueurs tumoraux spécifiques. Dans leur démarche, les biologistes sont amenés à considérer les échantillons de différentes manières. La première et la plus naturelle consiste à considérer chaque patient individuellement : l'échantillon sain et l'échantillon pathologique d'un même patient présentent des profils protéiques d'une grande proximité et l'interprétation des résultats de l'électrophorèse est plus facilement interprétable. Dans ce cas, la recherche de marqueurs potentiels consiste à recouper les listes des marqueurs potentiels identifiés sur le plus grand nombre de patients. L'inconvénient majeur de cette approche réside dans le fait que, sur un seul patient, une protéine doit s'écarter assez largement d'un seuil de significativité pour être déclaré comme marqueur potentiel. Il y a donc un risque de passer à côté de marqueurs potentiels, certes peu variants, mais pourtant très intéressants s'ils peuvent être confirmés sur un grand nombre de patients. C'est

pour cela qu'il est utile de raisonner également par groupe de patients, en séparant l'ensemble des données d'une expérience mettant en jeu plusieurs patients, en une classe d'échantillons sains et une autre d'échantillons pathologiques. Grâce à la nature du test statistique envisagé, il est également possible de conserver l'information apportée par l'appariement des échantillons.

Dans ce chapitre nous présenterons les deux méthodes retenues qui correspondent à ces deux approches. Dans un premier temps, nous introduirons la méthode du t-test apparié avec régulation de la variance, qui permet une étude par classe et qui tient compte du phénomène liant la variance à la valeur des quantifications (hétéroscédasticité). Dans un second temps nous proposerons une méthode basée sur l'estimation par noyau gaussien de la variance liée aux biais expérimentaux du ratio des quantités protéiques afin d'isoler les marqueurs potentiels patient par patient.

3.3.5.1 Méthode de régulation de la variance

La normalisation a permis la correction des quantifications brutes et rend ainsi possible l'interprétation inter-gels. Pour cela, nous disposons désormais du jeu de données complet des quantifications pour chacune des taches protéiques de l'analyse. Si l'expérience comporte n gels, alors pour chaque tache protéique X nous avons les n valeurs $x_1^A;...;x_n^A$ dans la classe A (par exemple : une classe d'échantillons sains) et les n valeurs $x_1^B;...;x_n^B$ dans la classe B (par exemple : une classe d'échantillons pathologiques) du logarithme de son niveau d'expression. Pour chaque tache protéique, la question est de savoir si ce niveau d'expression varie significativement d'une classe à l'autre. Des méthodes statistiques classiques comme le test de Student semblent toute indiquée dans ce genre de situation. Cependant, nous allons voir qu'il est possible d'évaluer les expressions différentielles avec un meilleur degré de confiance, en nous inspirant des méthodes développées dans le domaine des micropuces ADN.

L'approche communément employée dans les premiers temps de l'électrophorèse bidimensionnelle était de comparer les ratios des expressions moyennes (valeur moyenne de l'expression d'une protéine dans la classe « pathologique » rapportée à sa valeur moyenne dans la classe « saine ») à un seuil typiquement et arbitrairement fixé à 2 pour les sur-expressions et à ½ pour les sous-expressions. Au-delà de ces seuils, le changement d'expression était considéré comme significatif. Malheureusement cette méthode ne peut pas être optimale puisqu'un même ratio peut avoir une significativité très différente suivant le niveau d'expression protéique : dans l'environnement bruité caractéristique de l'électrophorèse bidimensionnelle les ratios 2000/1000 et 2/1 peuvent avoir une significativité très différente.

Une autre approche consiste en l'utilisation du test de Student sur le jeu de données des logarithmes des niveaux d'expression protéique. Dans un test de Student, les moyennes empiriques m_A et m_B de chacune des classes d'une protéine et leur variance s_A^2 et s_B^2 associées sont utilisées pour le calcul d'une distance t normalisée entre les deux populations :

$$t = \frac{(m_B - m_A)}{\sqrt{\frac{s_B^2}{n_B} + \frac{s_A^2}{n_A}}}$$

avec pour chacune des deux populations,

$$m = \frac{1}{n} \times \sum_{i=1}^{n} x_i \text{ et } s^2 = \frac{1}{n-1} \times \sum_{i=1}^{n} (x_i - m)^2 ,$$

les estimateurs classiques de la moyenne et de la variance. Lorsque la distance t dépasse un certain seuil qui dépend d'un niveau de confiance déterminé par le biologiste, les deux populations correspondantes aux deux classes sont considérées comme étant exprimées de manière différentielle. Cette distance étant normalisée par la variance, elle permet d'éviter certains écueils de la simple méthode des ratios présentée précédemment. Ces deux approches restent cependant assez similaires. Le principal problème lié à l'usage du test de Student dans le cadre des expériences DIGE réside dans le fait que le nombre n de gels est faible. À cause des limitations techniques et de la quantité restreinte d'échantillons, le nombre de gels d'une expérience DIGE est de l'ordre de 10. Les divers aléas expérimentaux nous conduisent régulièrement à exploiter $n = 1, 2, 3, 4 \text{ ou } 5$ gels. Dans ces conditions, l'estimation de la variance est biaisée et le test de Student perd de sa puissance. Une approche mieux adaptée est donc nécessaire et consiste en l'utilisation d'un test de Student régulé basé sur les théories bayésiennes pour l'analyse des données.

L'approche bayésienne interprète les probabilités comme des niveaux de confiance en une hypothèse plutôt que comme une mesure de la fréquence d'un événement et spécifie comment affiner ces probabilités en considérant les données dans leur ensemble. Plus précisément, le théorème de Bayes permet de calculer la probabilité P à posteriori d'une hypothèse H sur la donnée D de la manière suivante :

$$P(H|D) = \frac{P(D|H)P(H)}{P(D)}$$

avec $P(H)$ la probabilité a priori de l'hypothèse H (dans notre cas l'hypothèse est qu'il n'y a pas de différence entre les populations des deux classes). Dans le cas des données de micropuces ADN comme dans le cas des données de l'électrophorèse DIGE, le problème est d'obtenir une estimation précise et robuste de la variance à partir d'un faible nombre de mesures. Cependant, dans ces deux domaines, il a été observé l'existence d'une relation entre cette variance et le niveau d'expression : les gènes, comme les protéines, exprimés à des niveaux similaires possèdent également des variances similaires. Cette connaissance a priori est déjà mise à profit grâce à l'approche bayésienne, pour les données d'expression géniques et doit pouvoir l'être également pour les données d'expression protéiques. Ainsi, en tenant compte des protéines de niveau d'expression voisin, il est possible d'obtenir une estimation plus robuste de la variance d'une protéine. C'est le principe de base du test de Student régulé.

À partir de cette constatation et en adoptant l'approche bayésienne, Baldi et Long [7] proposent une nouvelle estimation de la variance :

$$\sigma^2 = \frac{V_0\sigma_0^2 + (n-1)s^2}{V_0 + n - 2}$$

Dans cette formule, les $n-1$ observations empiriques de variance s^2 sont complétées par l'addition d'un jeu de V_0 observations de fond associées à une variance de fond σ_0^2. La Figure 71 donne un aperçu visuel du résultat de cette opération sur un jeu de données DIGE. L'ajout de des observations « de fond » semblent artificiel, mais correspond en fait à la prise en compte de l'information portée par la relation existante entre variance et niveau d'expression. L'implémentation du calcul de cette variance de fond doit rester souple afin de s'adapter aux jeux de données différents d'une fois sur l'autre. Dans le cadre des données de génomique il est possible d'utiliser une fenêtre symétrique de 100 gènes centrés sur le niveau d'expression du gène considéré. Dans le cadre des données de protéines, la relation entre variance et intensité est moins marquée et afin d'éviter une estimation erronée de la variance, la méthode calcul de la variance est légèrement modifiée afin de pondérer davantage la variance empirique par rapport à la variance de fond. Cependant, la variété des situations est telle que l'utilisateur doit toujours se montrer prudent dans l'usage de cette méthode afin d'éviter des conclusions erronées. En effet : donner un poids important à la variance de fond ne peut se faire que si la dépendance entre la variance et le niveau d'expression est forte. En protéomique, ce cas est rare. D'un autre côté, ignorer complètement la variance de fond alors qu'une certaine dépendance est constatée correspond à une perte d'information et de robustesse quant à la désignation des expressions différentielles.

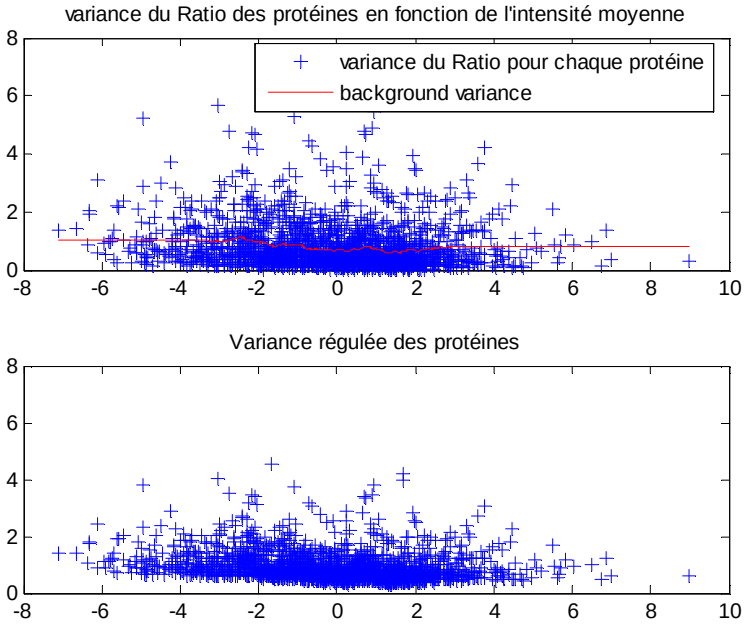


Figure 71 : Visualisation (en rouge, en haut) de la variance de fond associée à la variance des ratio d'une population protéique et de la correction apportée à cette variance (en bas).

Un extrait de l'implémentation de cet algorithme sous Matlab est donné ci-dessous :

```
%Estimation des variances régulées pour chaque spot et chacune des deux classes
%(Differential analysis of DNA microarray gene expression data, Hatfield et al., 2003)


% d'abord il faut estimer la variance de fond (background variance)
w = W/2; %demi-taille de la fenetre centrée sur le spot considéré
q = 0; %percentile à exclure
var_R = var(T-M,0,2);
mean_R = mean(T-M,2);
mean_A = mean((T+M)/2,2); % Attention on observe T+M et non pas T-M afin d'avoir une
estimation de l'intensité moyenne de la protéine
index1 = [1:nlig_gel]'; %index facilitant le tri
sorted_R = sortrows([mean_A var_R index1],1);
for i=w+1:nlig_gel-w
   K = sortrows(sorted_R(i-w:i+w, :),2); % Pour la fenêtre centrée sur i, on trie sui-
vant la variance locale afin de pouvoir retirer les outliers
   bg_var_R(i) = mean(K(1:2*(w-floor(q*w))+1, 2));% background variance  (variance
globale = variance inter + variance intra) On néglige la variance inter
end
bg_var_R(1:w)= bg_var_R(w+1);
bg_var_R(nlig_gel-w+1:nlig_gel)= bg_var_R(nlig_gel-w);


% On revient à l'ordre original des données en réordonnant suivant l'index
U = sortrows([sorted_R bg_var_R'],3);


% Ensuite on calcule l'estimation robuste de la variance associée à chaque à protéine,
par classe
n = ncol_gel/2;
robust_var_R = (V0*U(:,4)+(n-1)*var_R)/(V0+n-2);


% Calcul du t-test régulé
t_reg = mean_R./sqrt(robust_var_R);
% calcul de la p-value associée
ndl = n+V0-2; % nombre de degrés de liberté
p = 1-pt(abs(t_reg),ndl); % pt est une fonction de la toolbox (libre) "stixbox" qui
donne la valeur de la fonction cumulative de la loi de student


% génération d'un tableau regroupant des données
tab_t_reg = [Id mean_A  mean_R  robust_var_R  t_reg  sign(mean_R) p  T-M ];


% Détermination des expressions
for i=1:nlig_gel
   if sign(tab_t_reg(i,3))>0
      expr(i)={'SurEx'};
   elseif sign(tab_t_reg(i,3))<0
      expr(i)={'SousEx'};
   else
      expr(i)={'='};
   end
end


% on ordonne le tableau des données
[tab_t_reg i_ord] = sortrows(tab_t_reg,5);


% Sélection des spots significatifs
liste_id = tab_t_reg(find(tab_t_reg(:,7)<=p_seuil),1)';
```

Dans le logiciel ProDIGE implémenté sous Matlab, la valeur V_0 est laissée au choix de l'utilisateur. Cependant, il est conseillé de choisir V_0 de telle manière à avoir d'une part $V_0 + n$ constant pour toutes les études menées et d'autre part $V_0 < n$ de sorte à éviter tout risque de fausse interprétation puisque non justifiée (la dépendance entre variance et niveau d'expression en électrophorèse DIGE est généralement faible).

Un des grands avantages de cette technique est également de rendre comparable les estimations de significativité d'expression différentielle d'une expérience à l'autre, et ce, même si chacune d'elles ne comporte pas le même nombre de gel. En effet, la significativité est calculée pour un nombre $V_0 + n$ constant. Bien entendu, il faut se montrer prudent et les quantités effectives de gels doivent tout de même rester similaires.

3.3.5.2 Méthode alternative : estimation d'une « courbe enveloppe »

a) Description de la méthode

Soient x la variable aléatoire correspondant au logarithme du volume de la paire de spots et y celle correspondant au logarithme base 2 de son ratio. À noter que par la suite, nous nous contenterons souvent de parler de ratio et d'intensité sans répéter qu'il s'agit de leur logarithme.

Comme l'illustre la Figure 72, la densité de probabilité de y est fonction de x. Il s'agit du phénomène d'hétéroscédasticité.

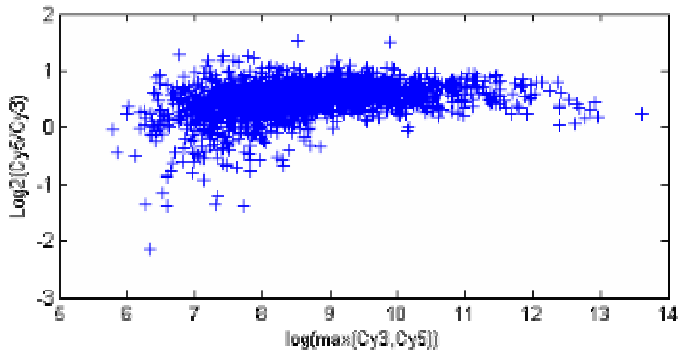


Figure 72 : Observation typique de y (logarithme base 2 du ratio volumique) en fonction de x (grandeur reflétant le volume de la paire de taches protéiques considérée) et mise en évidence du phénomène d'hétéroscédasticité

La méthode d'analyse différentielle présentée dans ce paragraphe a pour objectif de fournir un seuil de significativité sur le ratio volumique qui soit dépendant de l'intensité, afin de pallier à ce phénomène hétéroscédasticité. L'ensemble de ces seuils constitue une « frontière » que l'on appelle ici « courbe enveloppe ».

Le moyen choisi pour la construction de cette courbe enveloppe est d'analyser les données disponibles de manière à pouvoir quantifier l'impact des biais expérimentaux sur le ratio volumique en fonction de l'intensité. La courbe enveloppe ainsi établie à partir des données de répétabilité (cas « A »), voir ci-dessous, permettra de mettre en évidence les spots variants biologiquement (analyse différentielle cas « B »).

Le schéma expérimental adopté dans nos expériences consiste en une répétition avec inversion de fluorophore (« dye-swap »). Les jeux de données issus des expériences menées suivant ce schéma sont affectés par différentes sources de variations. Ces sources de variation sont résumées par le schéma de la Figure 73.

La mesure des ratios volumiques dans le cadre de l'analyse différentielle correspond au cas de figure référencé « B » sur ce schéma. Dans ce cas, la mesure de l'expression différentielle est masquée par les biais relatifs à l'emploi de deux fluorophores différents ainsi que par les biais expérimentaux qui, bien que largement réduits par l'emploi de la technologie DIGE, sont inéluctables.

Pour supprimer de l'analyse différentielle les spots issus de ces biais technologiques, nous avons choisi de construire une courbe enveloppe qui fixe un seuil de significativité biologique, à partir des données de répétabilité. Le choix a été fait d'assimiler l'ensemble des biais affectant le ratio volumique dans le cas de figure « B » à celui du cas de figure « A » qui sont les données de répétabilité inter-gels. En effet dans ce cas de figure « A », les mesures du ratio ne concernent qu'un seul et même échantillon qui a migré dans 2 gels différents. Dans ce cas il n'y a donc pas de variation d'expression biologique, et il est donc possible d'estimer directement la variabilité expérimentale inter-gels. Cette variabilité servira à la détermination de la courbe enveloppe qui sera appliquée au cas « B » pour la détermination des spots variant biologiquement.

Cette assimilation des biais expérimentaux des cas « A » et « B » est justifiée par le fait que la variabilité expérimentale inter-gels (cas « A ») est supérieure à la variabilité intra-gel (cas « B »). Il en ressort que l'estimation de la variabilité expérimentale globale à partir du cas de figure « A » est surévaluée. L'utilisation de cette estimation pour la détermination d'une limite de significativité des ratios volumiques est donc valable même si elle ne permet pas de quantifier cette significativité. Tous les spots sortant de la courbe enveloppe ont une vraie significativité biologique (en fonction du seuil choisi avec un risque d'erreur déterminé). Certains spots biologiquement significatifs seront par contre non observés.

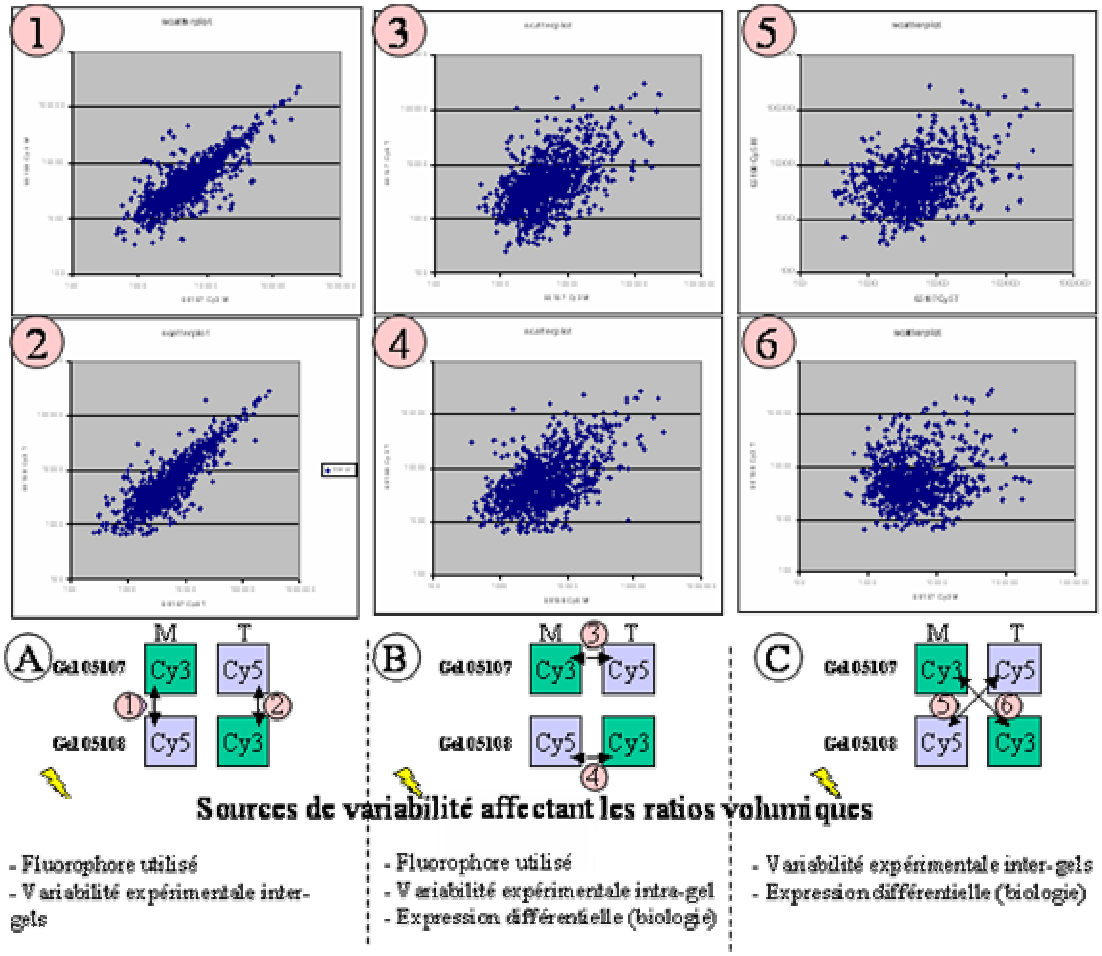


Figure 73: Visualisation sur un cas réel des différentes sources de variabilité du ratio volumique affectant deux images issues de la répétition par inversion de fluorophores de l'électrophorèse DIGE des échantillons muqueuse et tumeur d'un même patient.

Pour l'estimation de la variabilité expérimentale d'une expérience donnée, un jeu de données regroupant l'ensemble des ratios correspondant au cas de figure « A » des données de répétabilité est constitué. On note M(n,g) et T(n,g) les populations protéiques des échantillons respectifs de muqueuse et de tumeur correspondant au patient n et au gel g (g=1 ou 2). On note N le nombre de patients considérés pour l'expérience. Pour un patient n donné, les populations protéiques M(n,1) et M(n,2) sont identiques tout comme les populations T(n,1) et T(n,2) (voir Figure 74). Ainsi, l'estimation de la variabilité expérimentale peut être réalisée à partir d'un seul et même nuage de points issus de tous les couples de populations (M(n,1),M(n,2)) ainsi que des couples (T(n,1),T(n,2)) pour l'ensemble des patients, n allant de 1 à N. La Figure 75 présente l'exemple d'un tel nuage de points. Il faut noter que pour ces couples de populations, il n'y a plus de notion de classe et qu'il est possible de considérer indifféremment le ratio logarithmique ou son opposé. C'est pourquoi, les nuages de points peuvent être symétrisés. Cette symétrie permet notamment l'emploi des opérateurs de convolution décrits par la suite.

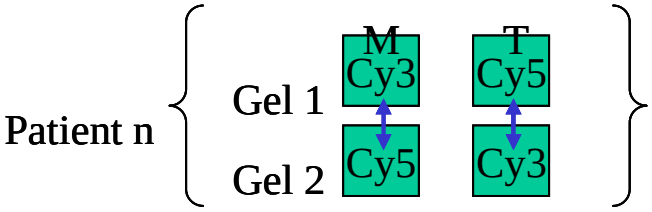


Figure 74 : Constitution du jeu de données pour l'évaluation de la variabilité expérimentale. Les population protéiques M(n,1) et M(n,2) sont identiques. De même pour T(n,1) et T(n,2)

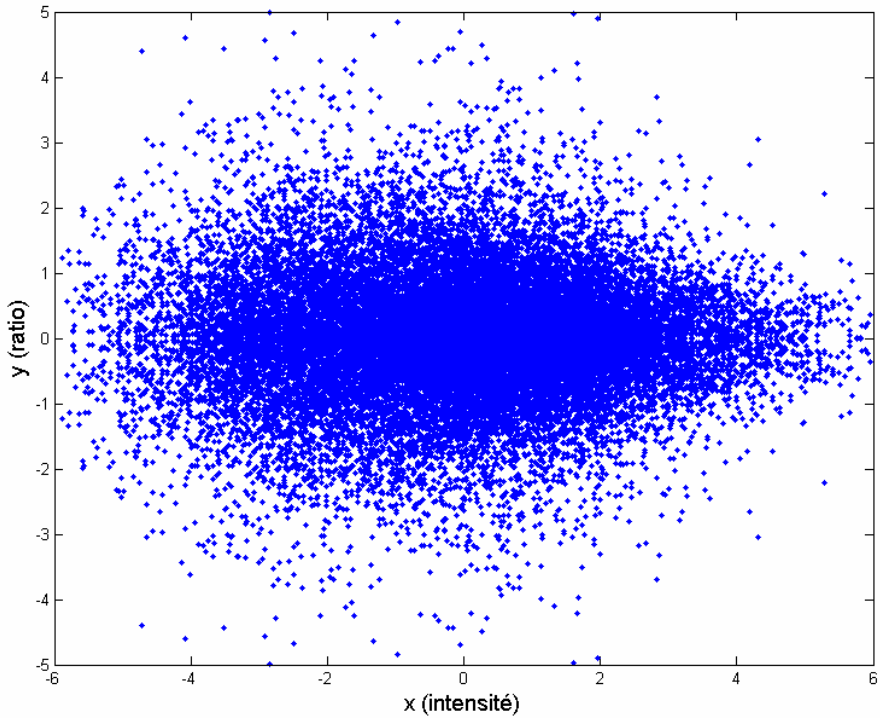


Figure 75:Exemple d'un nuage de points (x_i, y_i) issus des couples de populations (M(n,1),M(n,2)) et (T(n,1),T(n,2)) pour tous les patients n d'une expérience : données de répétabilité.

Afin d'effectuer l'estimation de la variabilité expérimentale, il est nécessaire de réaliser la mesure de la densité de probabilité de y, le ratio volumique, conditionnellement à x, l'intensité associée, sur ces données affranchies des variations biologiques. L'estimation de la densité de probabilité $pdf(y|x)$ se fait par la méthode non paramétrique classique de Nadaraya Watson [40] avec lissage par noyau gaussien sur le nuage de points associé.

$$pdf(y|x) = \sum_{i=1}^{n} p_i(x) \times e^{-\frac{(y-y_i)^2}{2h_y^2}}$$

avec $p_i(x) = \dfrac{e^{-\frac{(x-x_i)^2}{2h_x^2}}}{\sum_{j=1}^{n} e^{-\frac{(x-x_x)^2}{2h_x^2}}}$

et h_x et h_y les paramètres du noyau gaussien choisis pour l'estimation. À partir de cette densité conditionnelle il est alors possible, grâce à une simulation de Monte-Carlo, d'estimer les seuils de significativité pour les ratios en fonction du volume de la paire, et de définir ainsi une « courbe enveloppe ».

Mais avant cela, il est nécessaire d'estimer h_x et h_y en appliquant la méthode des noyaux gaussiens à la distribution de x et à celle de y. Les choix de h_x et h_y doivent permettre de respecter l'allure des distributions dont une première approximation nous est apportée par les histogrammes. Le choix des classes des histogrammes doit également être l'objet d'une attention particulière, en gardant à l'esprit que la distribution attendue est grossièrement gaussienne. Il faut par ailleurs veiller à ce que l'ajustement des paramètres h_x et h_y n'entraîne pas une sur-modélisation. Cette sur-modélisation se manifeste par l'apparition d'un bruit haute fréquence sur la distribution estimée telle que celle présentée sur la Figure 76.

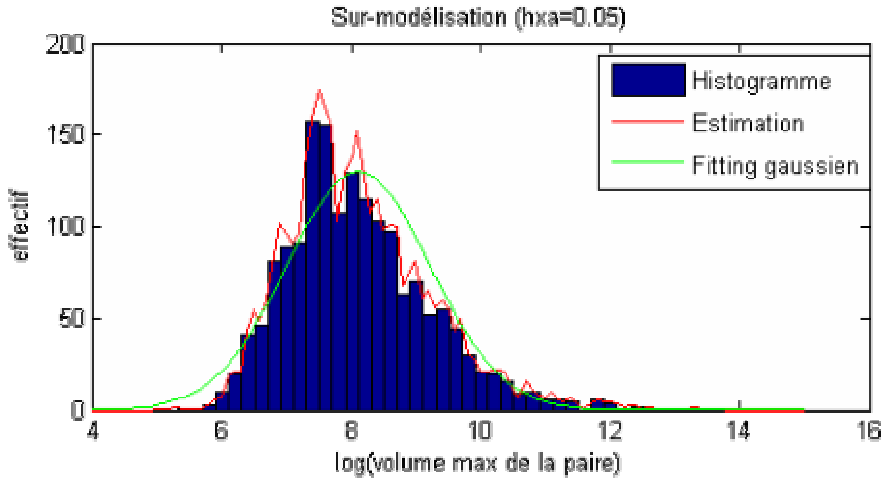


Figure 76 : Visualisation d'un cas de sur-modélisation

L'utilisation d'un histogramme permet donc d'évaluer h_x et h_y de telle sorte que l'estimation des distributions par la méthode des noyaux gaussiens reflète au mieux les distributions. Les valeurs h_x et h_y optimales sont donc déterminées de façon empirique. Elles peuvent être conservées pour des expériences similaires en terme de qualité de gel et de nombre de taches protéiques détectées.

Il est alors possible de visualiser (cf. exemple de la Figure 77), la fonction de densité jointe :

$$pdf(x,y) = \frac{1}{n}\sum_{i=1}^{N}\left(\frac{e^{\frac{-(x-x_i)^2}{2h_x^2}}}{h_x\sqrt{2\pi}} \times \frac{e^{\frac{-(y-y_i)^2}{2h_y^2}}}{h_y\sqrt{2\pi}}\right)$$

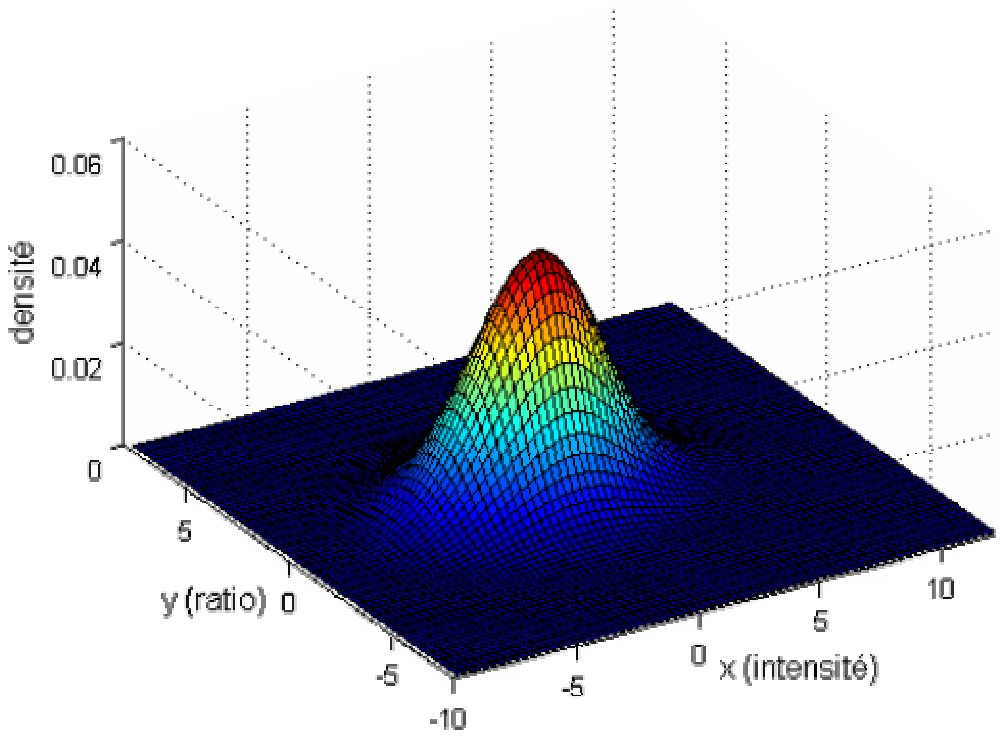


Figure 77 : Représentation graphique de la densité estimée du nuage de point de la Figure 75 par la méthode des noyaux gaussiens ($h_x = 1.5$ et $h_y = 1$).

Comme l'illustre la Figure 78, pour une valeur de x quelconque, on connaît maintenant la fonction densité de probabilité de y. On peut réaliser une simulation de Monte Carlo à partir de cette fonction : on génère suivant cette distribution un nombre M de valeurs y que l'on indice de la manière suivante : $y_1 < y_2 < ... < y_M$

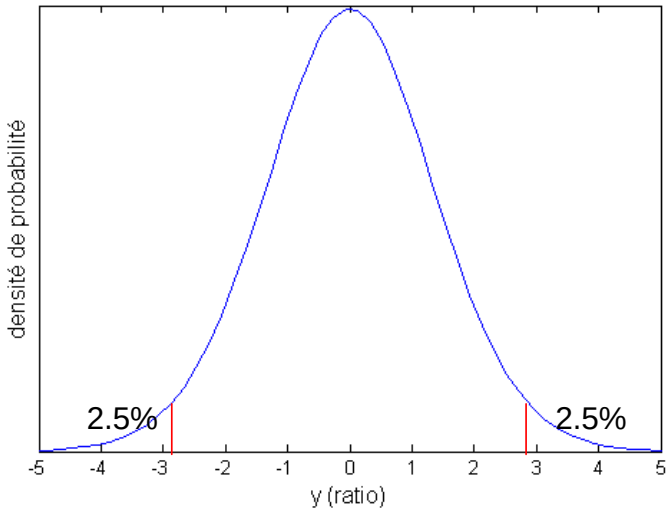


Figure 78: Seuils de significativité au risque de 5% à partir de la densité de probabilité pour une valeur x (intensité) fixée.

Les valeurs $y_{M\times 2.5\%}$ et $y_{M\times 97.5\%}$ tendent vers les seuils recherchés, au risque de 5% suivant l'hypothèse nulle : « les volumes sont égaux », lorsque M tend vers l'infini. Ceci est valable quel que soit le seuil de significativité considéré. Cette méthode permet de lisser la distribution et ainsi d'améliorer l'estimation des seuils de significativité.

Afin d'accélérer les temps de traitement informatique, les seuils sont calculés à partir du calcul de l'aire sous les courbes (histogrammes) de densité conditionnelle. L'aire sous la courbe est calculée en sommant les valeurs de la densité conditionnelle obtenues pour une rampe de valeurs du ratio et en multipliant cette somme par le pas choisi. Il s'agit du calcul de la fonction de distribution cumulée. Cette opération est réalisée sur toute la plage de l'intensité. Au final, elle permet de connaître, pour une intensité x et un ratio y quelconques (dans les plages étudiées), la probabilité d'occurrence, seulement due à la variabilité expérimentale, d'une tache protéique de ratio au moins égal à y, en valeur absolue (voir Figure 79).

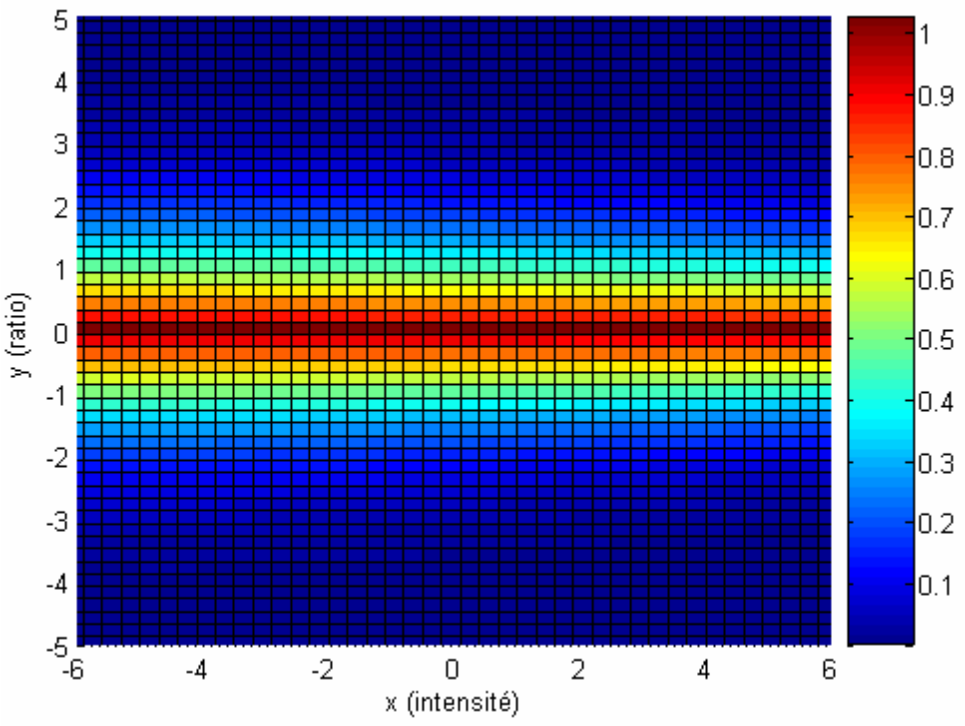


Figure 79: Illustration de la distribution dans le plan (x,y) des probabilités d'occurrence dues à la variabilité expérimentale, d'une tache protéique de ratio au moins égal à y (en valeur absolue).

Sur la Figure 79, la courbe enveloppe recherchée est la ligne de niveau de même probabilité. Les points situés en dehors de cette enveloppe correspondent aux taches protéiques présentant une différence d'expression significative avec un risque connu (typiquement 5%) de se tromper en l'affirmant. Comme l'illustre la Figure 80, la courbe enveloppe est ensuite utilisée sur les données de l'expérience, dans le cadre de l'analyse différentielle entre la classe tumeur et la classe muqueuse. Elle permet de désigner les taches correspondant à des variations biologiques en intégrant le risque que cette variation soit en fait due à la variabilité expérimentale.

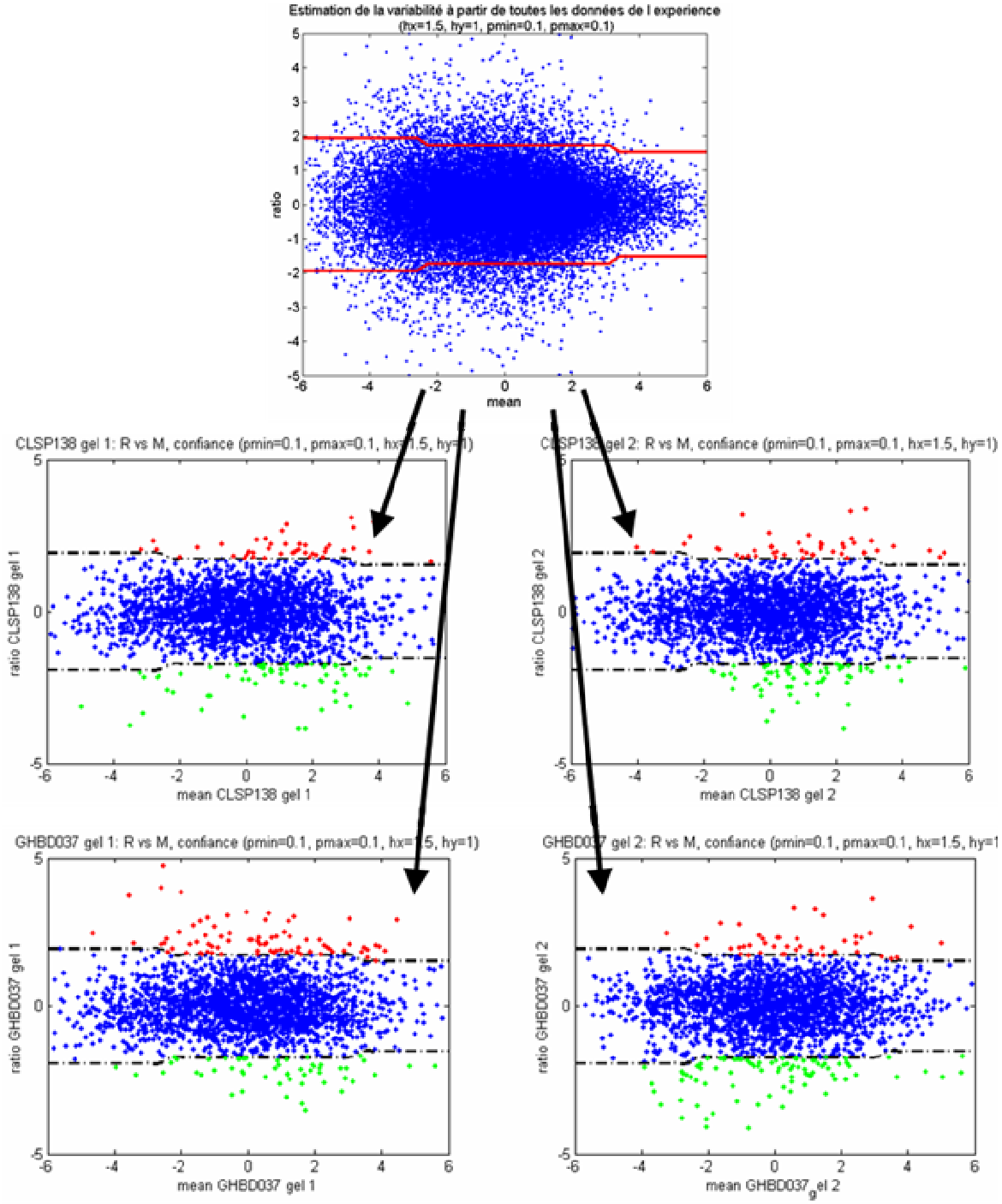


Figure 80 : Visualisation de la courbe enveloppe obtenue à partir du nuage de points des données de répétition biologique (figure du haut) et reportée sur différents cas d'analyse différentielle des classes tumeur et muqueuse pour la désignation tes taches protéiques de sur-expression (en rouge) et de sous-expression (en vert) biologique.

b) Mise en place d'un indicateur de l'intérêt biologique

La volonté des biologistes de porter davantage leurs recherches sur les taches protéiques de volumes (intensités) importants nous a amené à ne plus considérer simplement la courbe enveloppe définie à partir d'une ligne de niveau de la p-value

dans le plan MA (voir Figure 79). Le biologiste est capable de fournir une valeur de risque acceptable *pmin* pour les basses intensités, et *pmax* pour les hautes intensités. Ces deux valeurs permettent de définir un profil de risque fonction de l'intensité, tel que celui présenté en Figure 81.

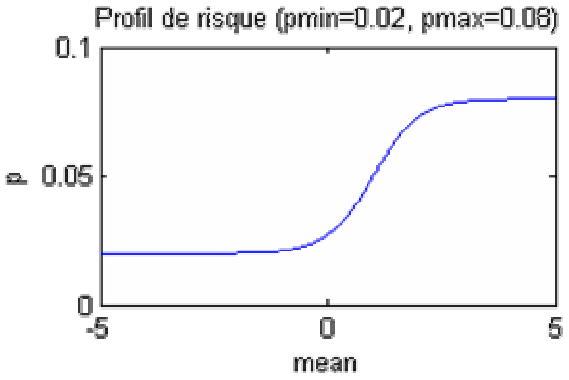


Figure 81: Profil de risque défini par le biologiste servant à définir une courbe enveloppe adapté à l'intérêt biologique différent que représentent des taches de faible ou de forte intensité.

Une première application de ce profil consiste en la redéfinition de la courbe enveloppe telle qu'elle a été présentée dans le paragraphe précédent. La courbe enveloppe n'est plus alors la ligne de niveau d'équiprobabilité : pour qu'une tâche protéique soit considérée comme présentant un intérêt biologique, sa p-value associée doit être en deçà du seuil précisé par le profil pour son niveau d'intensité.

Comme le montre la Figure 82, la prise en compte d'un profil de risque tel que celui présenté en Figure 81, conduit à l'obtention d'une courbe enveloppe possédant un profil en entonnoir qui favorise clairement la prise en compte des taches protéiques les plus intenses.

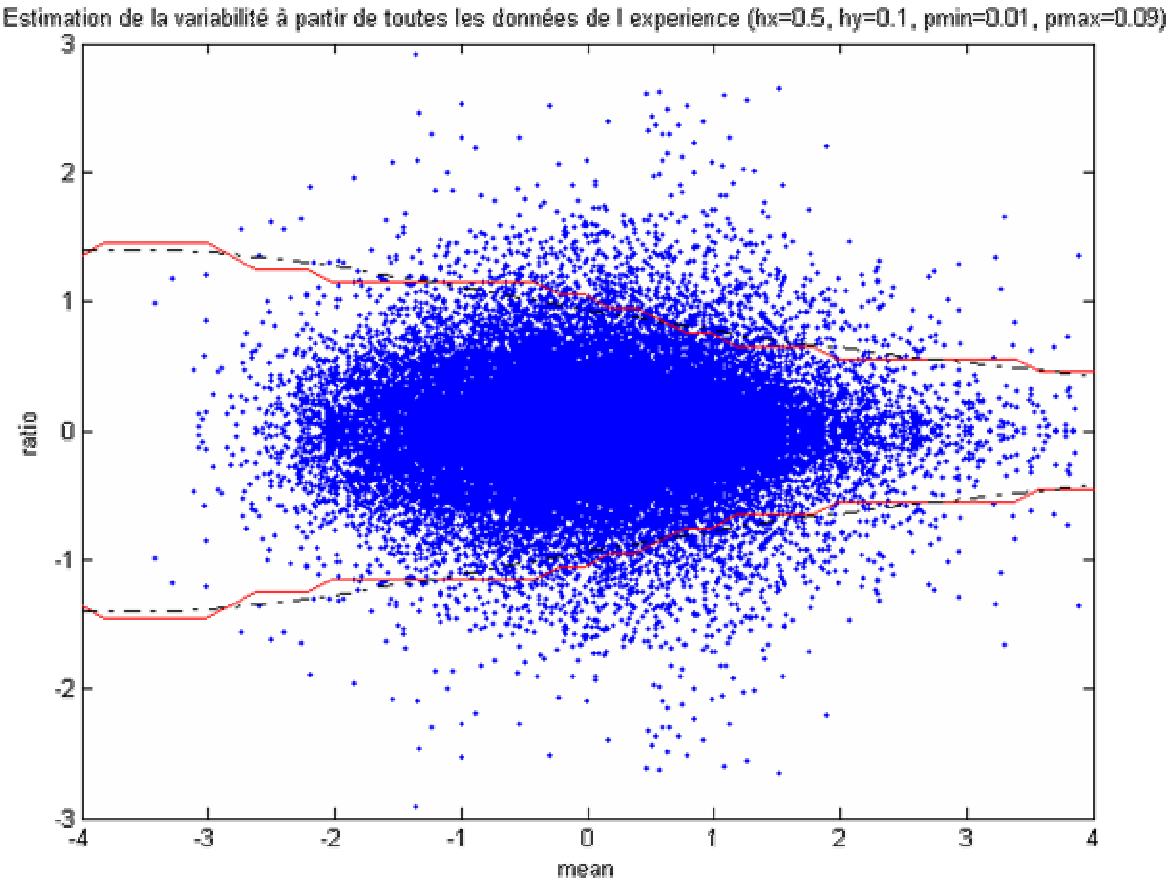


Figure 82: Visualisation de la courbe enveloppe (en rouge) et de son lissage (en pointillé) sur le nuage des points de répétabilité. Ici, la courbe correspond à la limite du risque acceptable par le biologiste. Ce risque est défini en spécifiant un profil tel que celui présenté en Figure 81. Dans ce cas de figure, le biologiste accepte un risque plus élevé pour les taches de forte intensité.

Cependant, l'approche qui consiste, pour chacun des gels de l'expérience, à ne retenir uniquement les taches protéiques dont les ratios sortent de cette courbe enveloppe s'avère trop restrictive. En effet, une tache protéique non retenue comme significative mais proche du seuil défini par le profil, peut s'avérer toutefois intéressante si elle également proche de ce seuil sur plusieurs autres gels de l'expérience. Ainsi, afin d'affiner la recherche des taches protéiques significativement variantes, un indicateur quantitatif de l'intérêt biologique a été mis en place.

Afin que cet indicateur reflète l'intérêt biologique, la probabilité associée à une tache protéique à l'issue de l'étude de variabilité (présentée précédemment) est pondérée de manière à ce que la ligne de niveau correspondant à $p = (p\min + p\max)/2$ définie sur les nouvelles valeurs soit identique à la ligne de niveau associée au profil, sur les p-valeurs initiales. C'est-à-dire que, si l'on fixe un seuil constant $p = (p\min + p\max)/2$ sur les nouvelles valeurs, les taches protéiques considérées comme intéressantes sont les mêmes que celles obtenues en considérant le profil sur les p-valeurs initiales. La modification des p-valeurs est observable sur la Figure 83. Il faut noter que, à partir de cet instant, les valeurs associées aux taches protéiques (on ne parle plus de p-valeurs) perdent leur sens statistique. Le sens statis-

tique peut cependant être retrouvé en annulant la pondération, à condition de connaître l'intensité de la tache protéique et le profil de risque associé.

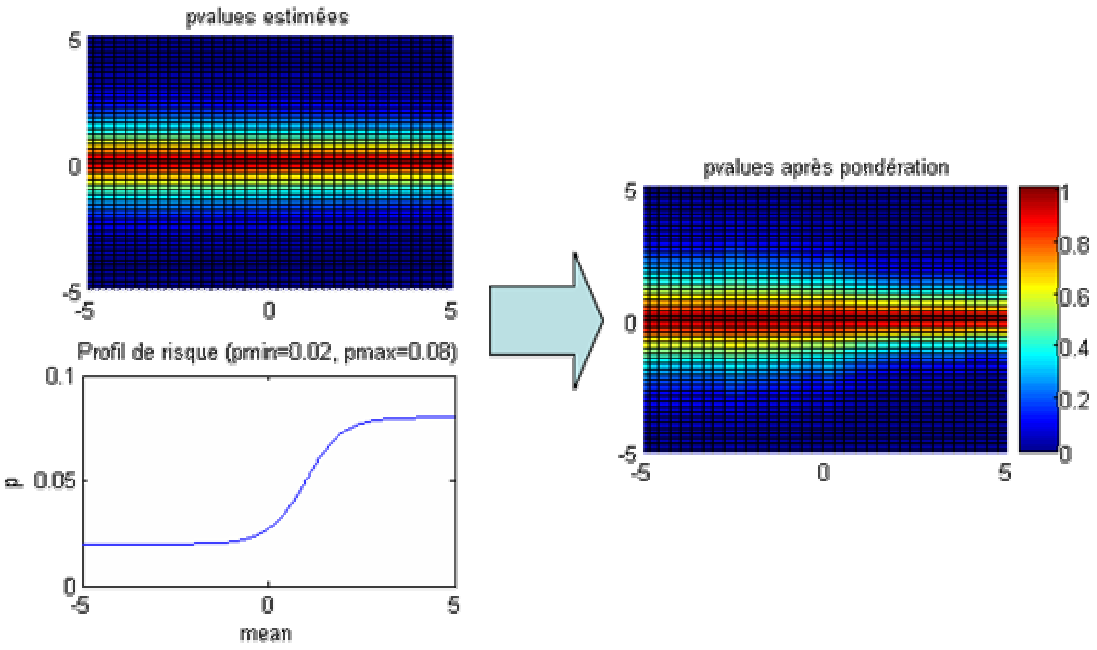


Figure 83 : Les valeurs de probabilité obtenues par l'analyse de la variabilité sont pondérées afin de tenir compte de l'intérêt des biologistes qui est fonction de l'intensité des taches.

Afin de rapprocher l'indicateur de la logique binaire généralement adoptée pour désigner les marqueurs potentiels, il est ensuite nécessaire de considérer le complémentaire à 1 de cette probabilité pondérée. Puis, pour rendre compte de la nature de l'expression, l'indicateur est signé : négativement pour une sous-expression, positivement pour une sur-expression. De cette manière, lorsque l'indicateur d'intérêt biologique est proche de 1 ou de -1, la tache protéique concernée peut être considérée respectivement en sur-expression ou en sous-expression suffisante par rapport à la gamme d'intensité dans laquelle elle se trouve.

Comme indiqué précédemment, cet indicateur n'a pas un sens statistique direct et doit être considéré comme une quantification de ce qui fait l'intérêt d'une tache protéique aux yeux du biologiste. Cet indicateur permet surtout d'établir des classements conduisant les biologistes à s'intéresser en priorité aux taches protéiques les plus susceptibles d'être à la fois statistiquement significative et biologiquement exploitables. L'aspect quantitatif étant conservé à l'issue de l'analyse gel par gel, il est possible de compiler l'ensemble des résultats d'une ou même plusieurs expériences de DIGE. Cet aspect de l'exploitation de l'intérêt biologique sera mis en avant dans l'application présentée en troisième partie de la thèse.

V.3.4 Conclusion: le Workflow IDADIGE

Le workflow IDADIGE, représenté sur la Figure 84 permet de répondre aux spécificités de notre projet. Ce schéma permet de se figurer la succession des différentes méthodes que nous venons de mettre en place. Il comprend un certain nombre de traitements d'images et de données qui ont été développés, implémentés et testés

dans l'environnement Matlab. Cet environnement permet une grande souplesse et une grande réactivité. Le workflow fait également appel à d'autres logiciels pour certaines étapes critiques qui nécessitent une grande ergonomie. Ainsi, l'alignement des images est réalisé grâce au logiciel TT900 (Nonlinear Dynamics) tandis que la détection, la retouche et l'annotation des spots sont réalisées sous ImageMaster 2D (Amersham Biosciences). Enfin, le workflow fait également appel à l'expertise d'un biologiste pour ces mêmes étapes critiques. L'outil proDIGE a donc été développé en utilisant du codage de novo sous Matlab, et fonctionnant en association à des logiciels existants.

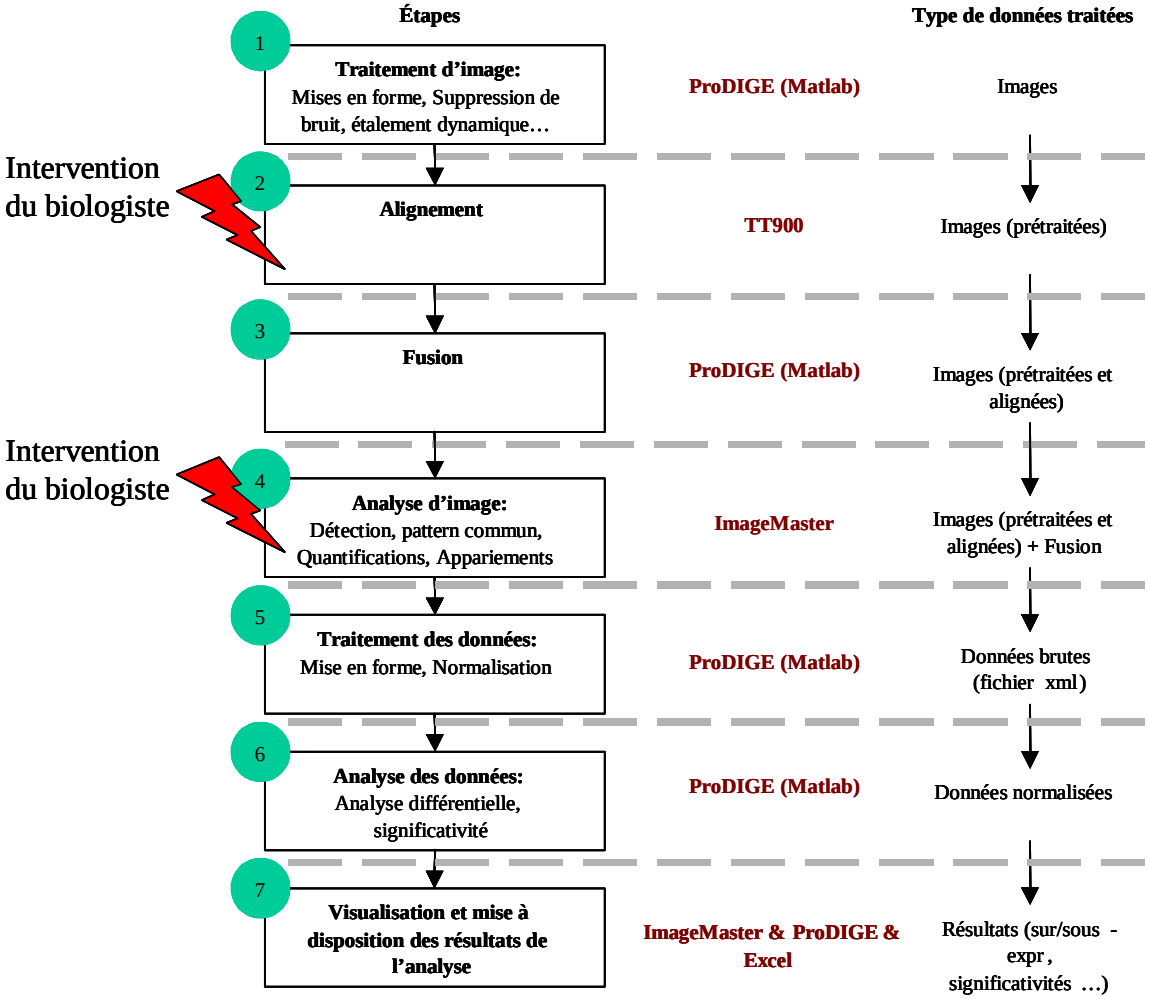


Figure 84 : Représentation du workflow IdaDIGE défini pour notre étude, et comprenant ProDIGE créé sous Matlab

Les logiciels nécessaires au workflow sont:

- Excel,
- ProDIGE (script principal et routines) créé à l'aide de Matlab doté de la toolbox d'analyse d'image,
- TT900 de nonlinear Dynamics
- et ImageMaster 2D Platinium de GE Healthcare.

Le logiciel d'alignement TT900 peut être remplacé par un équivalent Par contre l'usage d'ImageMaster 2D est imposé étant donné que ProDIGE ne sait extraire les données qu'à condition qu'elles soient au format particulier généré par ImageMaster. De même ProDIGE génère des résultats pouvant être importés sous ImageMaster 2D.

VI. Application du workflow IDADIGE pour la découverte de marqueurs potentiels du cancer colorectal

VI.1 Introduction

Le workflow IDADIGE a permis d'exploiter 5 expériences DIGE différentielles ayant pour but la découverte de marqueurs potentiels du cancer colorectal. Ces expériences ne comprennent pas toutes le même nombre de patients étudiés. Ceci s'explique par la disponibilité variable des échantillons biologiques ainsi que par certains aléas expérimentaux qui ont conduit à écarter certains patients pour lesquels les données ont été jugées inexploitables (gels de mauvaise qualité, quantité insuffisante de protéines…). Ces expériences sont celles qui ont été menées en

- novembre 2004 (5 patients étudiés),
- en février 2005 (6 patients étudiés),
- en juin 2005 (5 patients étudiés),
- en janvier 2006 (3 patients étudiés),
- ainsi qu'en mars 2006 (5 patients étudiés).

Par ailleurs d'autres types d'expériences DIGE ont été menés et exploités suivant le workflow IDADIGE. Il s'agit notamment d'expériences ayant pour but d'évaluer la spécificité des marqueurs protéiques potentiels du cancer colorectal. Pour cela, l'étude différentielle était menée entre une classe d'échantillons témoins et une classe d'échantillons associés à des maladies inflammatoires du rectum et du colon. Ces études ont permis d'écarter certaines protéines pour lesquels le changement d'expression n'était pas spécifique au cancer colorectal.

Tout d'abord, afin d'illustrer le worflow IDADIGE, nous présentons le cas de l'exploitation des données de janvier 2006. Ce jeu a été choisi pour la qualité très représentative des images de gels qui lui sont associées. L'utilisateur pourra se référer à l'annexe B pour une description pratique du déroulement des opérations.

Ensuite, nous reviendrons sur l'intérêt du workflow IDADIGE en terme de marqueurs potentiels mis en évidence en élargissant aux résultats obtenus à partir de l'ensemble des expériences DIGE exploitées. Nous discuterons de l'apport aux biologistes dans le contexte de la recherche de marqueurs protéiques.

VI.2 Présentation du jeu de données

L'expérience DIGE, présentée ici, comprend trois patients. Chacun d'eux a fourni un échantillon sain (Muqueuse) et un échantillon pathologique (Tumeur). Le schéma expérimental précédemment décrit et rappelé en Figure 85 a été appliqué. Pour chacun des trois patients, une première co-migration DIGE est réalisée au sein d'un même gel. Cette co-migration DIGE concerne les 3 populations protéiques correspondant aux 2 classes étudiées (tumeur et témoin) et au standard commun (lignée cellulaire de Caco2). Cette co-migration DIGE est répétée, pour chacun des trois patients avec, pour seul changement, l'inversion du marquage en fluorescence pour les populations protéiques tumeur et muqueuse. Il s'agit de la répétition dite en « dye-swap ». La troisième voie (le troisième fluorophore) est réservée à la lignée Caco2.

Comme nous l'avons évoqué précédemment, la lignée Caco2 est essentielle pour l'alignement les images issues de gels différents. Elle permet, par ailleurs, de faciliter le rapprochement inter-expérimental en fournissant un ensemble de motifs de taches protéiques biologiquement identiques et facilement identifiables d'une expérience à l'autre.

Le schéma de la Figure 86 représente les images acquises depuis les gels issus de l'expérience. Ce schéma permet de visualiser la structure des données et le nombre d'images à exploiter : il y a 3 fois 2 gels qui fournissent chacun 3 images, ce qui fait, au total, 18 images à considérer.

Avant d'aller plus loin dans la description de la mise en œuvre du workflow, il est utile d'évoquer la nomenclature choisie pour les différentes dénominations.

- Les gels :
- Préfixe = Année (01 pour 2001)
- Suffixe = Incrémentation
- Les patients :
- Préfixe = Provenance
- Suffixe = Incrémentation
- Les classes et le marquage :
- Préfixe = Classe (M pour Muqueuse et T pour Tumeur)
- Suffixe = Marquage en fluorescence (3 pour Cy3 et 5 pour Cy5)
  La lignée Caco2 étant systématiquement marquée par le même cyanine, on se contente de préciser Cy2.
- Les noms des fichiers des images de gel :
  Leurs noms précisent le gel, le patient et la classe concernée, selon la nomenclature précédemment mise en place. Par ailleurs, au cours des différents traitements (filtrages, mises à niveau des dynamiques, etc.) des suffixes sont ajoutés automatiquement par l'outil proDIGE.

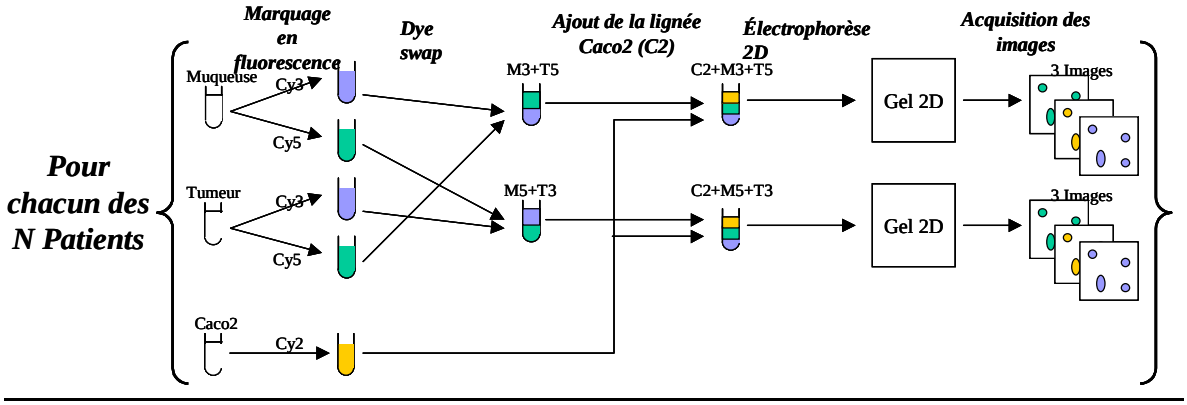


Figure 85: Schéma du procédé expérimental adopté lors des expériences DIGE. Dans le cas de l'étude présentée ici N=3.

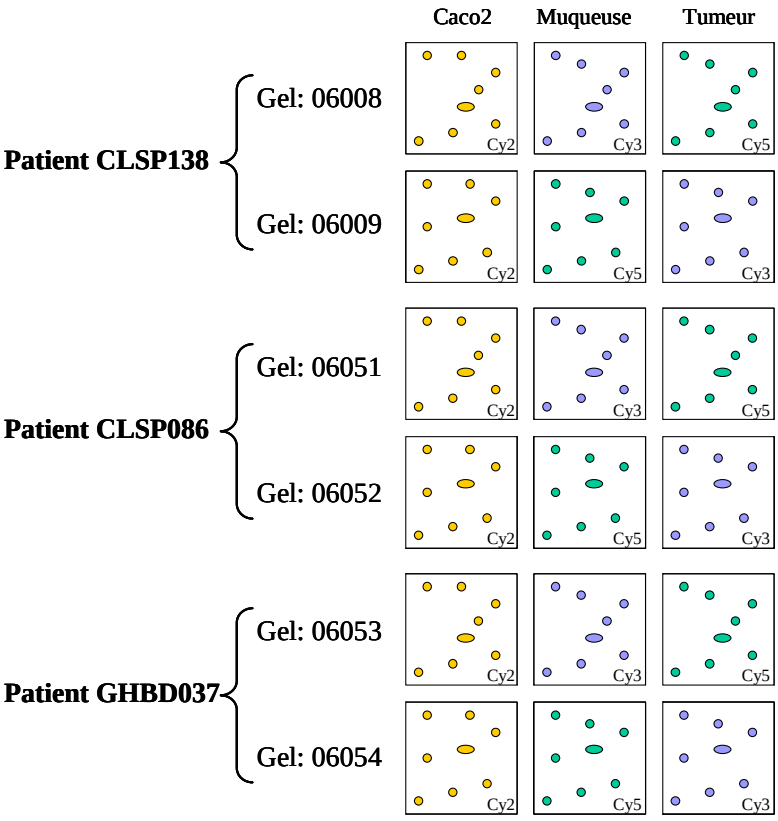


Figure 86: Schéma résumant l'ensemble des données brutes (les images) qui nous servent pour l'illustration de l'application du processus ProDIGE. Chaque patient et chaque gel se voient attribuer un nom unique suivant une nomenclature notifiant notamment son origine (hôpital où a été réalisé le prélèvement) mais que nous ne détaillerons pas ici.

VI.3 Traitement des images

Avant même d'aligner les images il est nécessaire de les mettre en forme et de supprimer les biais éventuels (principalement le bruit. Comme décrit dans la partie 2, nous conservons le fond, même si l'outil ProDIGE permet de le supprimer quand on le souhaite). Ce prétraitement doit être le moins destructif possible, afin de préserver l'information portée par l'image, tout en maximisant le rapport signal sur bruit. Il vise surtout la mise en forme (actions basiques sur les images comme le pivotement ou l'inversion des niveaux de gris) ainsi que l'amélioration de l'aspect visuel (étalement de l'histogramme des niveaux de gris afin d'occuper toute la gamme dynamique). La suppression du bruit est l'action la plus destructive du signal et doit être réservée aux cas extrêmes, heureusement très rares.

VI.3.1 Les données d'entrée

Les données d'entrées sont les images brutes, issues de l'acquisition des gels par le système d'imagerie à fluorescence. Elles sont au format « tif » et présentent des dimensions différentes suivant la taille du gel et le choix de la définition pour l'acquisition :

> 06008 CLSP138 Cy2.tif
> 06008 CLSP138 M3.tif
> 06008 CLSP138 T5.tif
> 06009 CLSP138 Cy2.tif
> 06009 CLSP138 M5.tif
> 06009 CLSP138 T3.tif
> 06051 CLSP086 Cy2.tif
> 06051 CLSP086 M3.tif
> 06051 CLSP086 T5.tif
> 06052 CLSP086 Cy2.tif
> 06052 CLSP086 M5.tif
> 06052 CLSP086 T3.tif
> 06053 GHBD037 Cy2.tif
> 06053 GHBD037 M3.tif
> 06053 GHBD037 T5.tif
> 06054 GHBD037 Cy2.tif
> 06054 GHBD037 M5.tif
> 06054 GHBD037 T3.tif

Les images de l'expérience de janvier 2006, prises pour exemple, présentent des dimensions d'environ 900x900 pixels et une gamme dynamique de 16 bits.

VI.3.2 Description du traitement

Dans la grande majorité des cas, il apparaît que les images acquises en fluorescence sont très peu bruitées. Les poussières, taches de coloration et cassures du gel

sont peu courantes et il est souvent préférable de se passer du filtrage alterné séquentiel proposé par proDIGE. En effet, les quelques artefacts présents sur les images ont pour simple conséquence d'entraîner de fausses détections qu'il faut alors identifier et éliminer. Ces corrections seront effectuées par des retouches manuelles sur le patron de détection, plus tard dans le workflow. Traiter l'ensemble de l'image pour quelques artefacts facilement localisables est donc souvent un mauvais choix car le filtrage alterné séquentiel, bien que peu destructif, altère tout de même légèrement l'image, ce qui signifie une perte d'information.

Les images de janvier 2006 sont peu bruitées et le seul prétraitement qui leur est appliqué consiste en un étalement de la dynamique. En effet, les images de ce jeu de donnée n'occupent pas systématiquement toute la gamme dynamique, c'est-à-dire que les valeurs extrêmes de leurs niveaux de gris ne sont pas 0 et 65535 (dynamique 16 bits). Comme le montrent la

Figure 87 et la Figure 88, l'étirement de la dynamique améliore le contraste de l'image. Ce traitement n'étant pas destructif, il peut être réalisé de manière systématique afin d'améliorer l'aspect visuel des images lorsque celles-ci n'ont pas été acquises dans des conditions optimales.

L'efficacité de l'amélioration visuelle des images par cette technique est cependant limitée. Il suffit, en effet, de deux pixels « aberrants » pour occuper toute la gamme dynamique. La présence des marqueurs moléculaires, visibles sur la gauche des images de gel, ainsi que de quelques artefacts, écrase donc souvent la dynamique. Pour remédier à cette limitation, un autre étalement est possible mais n'est pas proposé lors l'étape de prétraitement, car il est destructif. Il sera présenté par la suite lors de l'étape de fusion des images.

VI.3.3 Les données de sortie

Pour les données de sortie, nous avons choisi de conserver le format tif à toutes les étapes. Un ou plusieurs suffixes, accolés automatiquement au nom du fichier, permettent de connaître les traitements ayant été effectués :

- « -dm » pour l'étalement de l'histogramme sur toute la gamme dynamique,
- « -bs » pour la suppression du fond (pas appliquée ici)
- « -nf » pour le filtrage du bruit (pas appliqué ici).

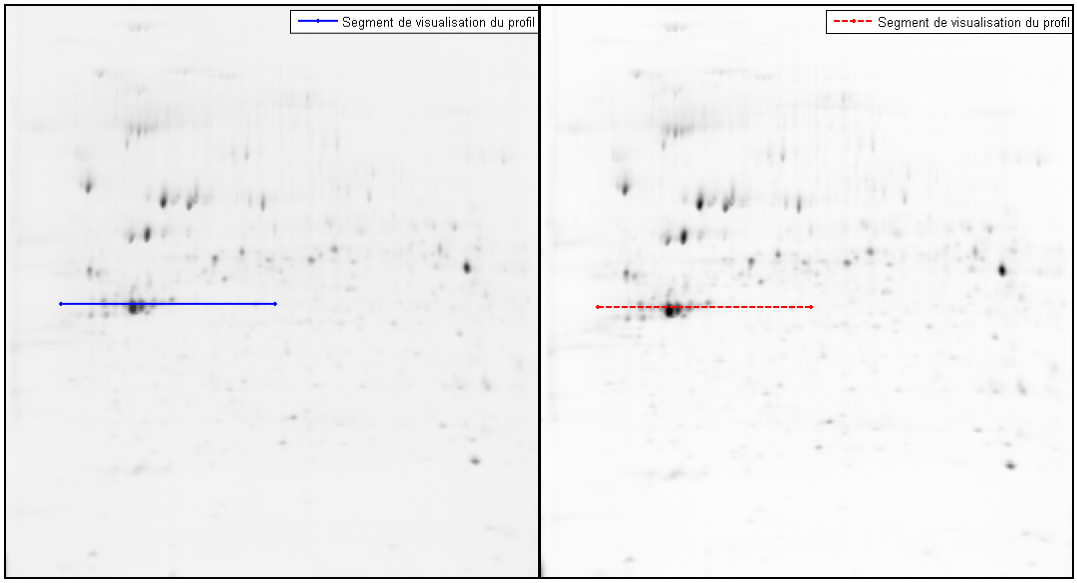


Figure 87: Effet de l'étalement de la gamme dynamique. À gauche, l'image brute associée à l'échantillon Tumeur du patient CLSP138 (image 06009_CLSP138_T3.tif) et à droite la même image après étalement de la dynamique. Après le traitement l'image présente un meilleur contraste permettant de distinguer certaines taches protéiques de faible intensité.

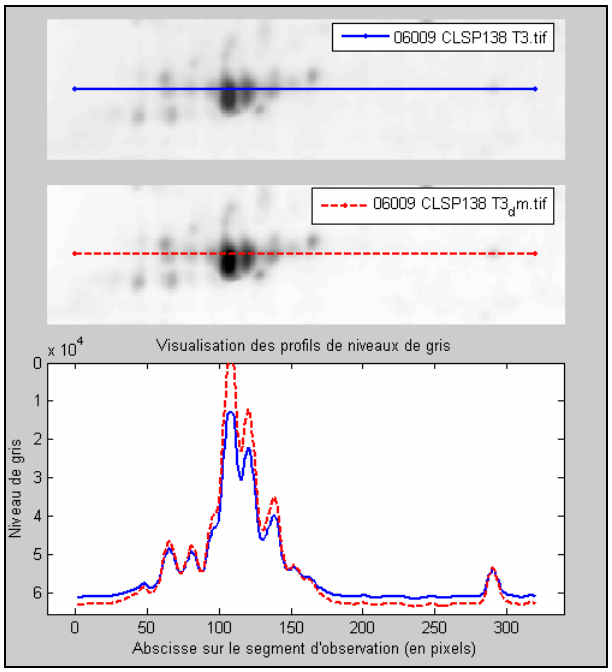


Figure 88: Visualisation de l'effet de l'étalement de la dynamique sur un profil de niveau de gris de l'image associée à l'échantillon Tumeur du patient CLSP138 (image 06009_CLSP138_T3.tif). La localisation du profil est visible sur la Figure 87. Le profil en bleu correspond à l'image brute et celui en pointillés rouges à l'image après l'étalement de la dynamique. La ligne de fond du profil rouge ainsi que le sommet des taches protéiques se rapprochent des valeurs extrêmes de la dynamique 16 bits.

VI.4 Alignement des images

VI.4.1 Les données d'entrée

Les images issues du prétraitement (

Figure 87) ont été mises en forme et leurs histogrammes ont été étalés sur toute la gamme dynamique. Il est maintenant nécessaire d'aligner ces images entre elles afin de pouvoir les interpréter. Les images à aligner sont donc les images au format TIF suivantes :

> 06008 CLSP138 Cy2-dm.tif
> 06008 CLSP138 M3-dm.tif
> 06008 CLSP138 T5-dm.tif
> 06009 CLSP138 Cy2-dm.tif
> 06009 CLSP138 M5-dm.tif
> 06009 CLSP138 T3-dm.tif
> 06051 CLSP086 Cy2-dm.tif
> 06051 CLSP086 M3-dm.tif
> 06051 CLSP086 T5-dm.tif
> 06052 CLSP086 Cy2-dm.tif
> 06052 CLSP086 M5-dm.tif
> 06052 CLSP086 T3-dm.tif
> 06053 GHBD037 Cy2-dm.tif
> 06053 GHBD037 M3-dm.tif
> 06053 GHBD037 T5-dm.tif
> 06054 GHBD037 Cy2-dm.tif
> 06054 GHBD037 M5-dm.tif
> 06054 GHBD037 T3-dm.tif

VI.4.2 Le traitement

Cette étape est réalisée entièrement sous le logiciel TT900 de Nonlinear Dynamics. Elle doit être conduite presque impérativement par l'expert biologiste car elle nécessite une compréhension des mécanismes biologiques qui peuvent induire des déformations du patron des taches protéiques qu'il ne faut pas corriger. Les décisions liées à l'alignement, notamment dans certaines régions fortement déformées des images, relèvent donc de la compétence et de la responsabilité du biologiste.

Cette étape exploite le potentiel lié à l'usage de la technologie DIGE ainsi que la présence d'un standard commun à tous les gels (lignée Caco2) utilisant un des fluorophores (Cy2) dans chaque gel. Sous TT900, le travail consiste en l'alignement des images Cy2 entre elles et par rapport à une image de référence préalablement choisie, le logiciel en déduit l'alignement de toutes les images. Cette image de référence est l'une des images en Cy2. On choisit l'image la moins bruitée qui comprend le plus grand nombre de spots. La Figure 89 présente une capture d'écran du logiciel.

VI.4.3 Les données de sortie

À l'issue de cette étape, les images Cy2, Cy3 et Cy5 alignées correspondant à l'analyse différentielle sont exportées au format TIF. Les noms des fichiers TIF possèdent désormais le suffixe warped indiquant que l'opération d'alignement a bien été opérée :

> 06008 CLSP138 Cy2-dm-warped.tif
> 06008 CLSP138 M3-dm-warped.tif
> 06008 CLSP138 T5-dm-warped.tif
> 06009 CLSP138 Cy2-dm-warped.tif
> 06009 CLSP138 M5-dm-warped.tif
> 06009 CLSP138 T3-dm-warped.tif
> 06051 CLSP086 Cy2-dm-warped.tif
> 06051 CLSP086 M3-dm-warped.tif
> 06051 CLSP086 T5-dm-warped.tif
> 06052 CLSP086 Cy2-dm-warped.tif
> 06052 CLSP086 M5-dm-warped.tif
> 06052 CLSP086 T3-dm-warped.tif
> 06053 GHBD037 Cy2-dm-warped.tif
> 06053 GHBD037 M3-dm-warped.tif
> 06053 GHBD037 T5-dm-warped.tif
> 06054 GHBD037 Cy2-dm-warped.tif
> 06054 GHBD037 M5-dm-warped.tif
> 06054 GHBD037 T3-dm-warped.tif

Il faut noter par ailleurs que les images Cy2 ne sont pas nécessaires pour la suite de l'analyse différentielle. Elles pourront cependant se révéler utiles dans le cadre d'un rapprochement ultérieur avec des études menées sur d'autres jeux de données.

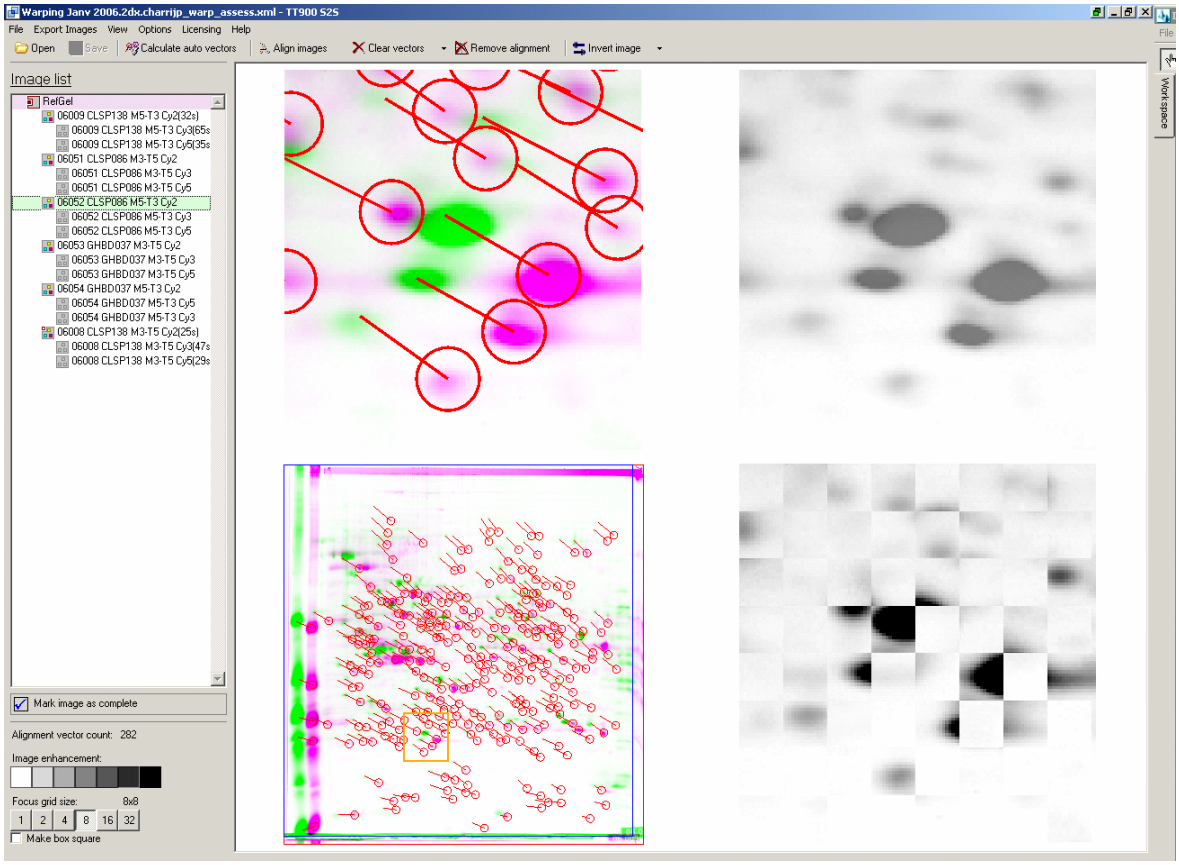


Figure 89: Capture d'écran du logiciel TT900 de Nonlinear Dynamics. Les 4 fenêtres de visualisation proposées par le logiciel, permettent à l'utilisateur de placer les vecteurs (en rouge) de mise en correspondance des taches protéiques des deux images à alignées. Un espace de travail, reprenant la structure DIGE de l'expérience, est présent sur la gauche de l'écran.

VI.5 Fusion des images

Cette étape met à profit l'alignement des images afin de produire une image dont la cartographie des taches protéiques regroupe toutes les protéines des toutes les populations de l'expérience.

VI.5.1 Les données d'entrée

Les données d'entrée sont les images prétraitées et alignées. Pour cette étape du workflow, les images associées à la lignée Caco2 sont écartées. En effet, la fusion a pour objectif de déterminer la cartographie de l'ensemble des protéines concernées par l'analyse différentielle c'est-à-dire des protéines issues des échantillons de tumeur et muqueuse mais pas celles issues de la lignée cellulaire de Caco-2.

> 06008 CLSP138 M3-dm-warped.tif
> 06008 CLSP138 T5-dm-warped.tif
> 06009 CLSP138 M5-dm-warped.tif
> 06009 CLSP138 T3-dm-warped.tif
> 06051 CLSP086 M3-dm-warped.tif
> 06051 CLSP086 T5-dm-warped.tif
> 06052 CLSP086 M5-dm-warped.tif
> 06052 CLSP086 T3-dm-warped.tif
> 06053 GHBD037 M3-dm-warped.tif
> 06053 GHBD037 T5-dm-warped.tif
> 06054 GHBD037 M5-dm-warped.tif
> 06054 GHBD037 T3-dm-warped.tif

VI.5.2 Le traitement

Malgré les prétraitements, des différences significatives subsistent entre les 12 images en termes de dynamique de niveaux de gris et une égalisation est nécessaire afin que la fusion ne favorise pas l'information de certaines images par rapport à d'autres.

L'algorithme d'égalisation proposé dans l'outil proDIGE est décrit en détail dans le paragraphe V.3.3.2.1. L'utilisateur précise la zone d'intérêt de l'image afin d'écarter les effets de bords et les mes marqueurs protéiques. L'utilisateur précise également une approximation du nombre de taches protéiques présentes sur l'image. Le choix de ce nombre se fait de manière pragmatique mais n'a pas un grand impact sur le résultat de l'égalisation s'il est suffisamment élevé pour pouvoir considérer les variations biologiques comme minoritaires et suffisamment faible pour ne pas être amené à considérer les bruits éventuellement présents comme des centroïdes de spots. Ces conditions ne sont pas très contraignantes. Et dans la grande majorité des cas observés dans le cadre du projet NODDICCAP, le choix de N=500 convenait parfaitement. C'est le cas pour les données de juin 2006. L'algorithme quantifie alors les niveaux d'intensité (niveaux de gris) des centroïdes des 500 taches protéiques les plus intenses de chaque image. L'égalisation des niveaux de gris de consiste alors à appli-

quer pour chaque image une transformation linéaire spécifique de manière à rapprocher les distributions des 12 populations les unes des autres.

Maintenant que les images sont mises à niveaux, il est possible d'appliquer l'algorithme de fusion décrit dans le paragraphe V.3.3.2.2.

VI.5.3 Les données de sortie

La fusion des images de notre jeu de données permet l'obtention de l'image présentée sur la Figure 90. Cette image est représentative de la population protéique observée sur l'ensemble des échantillons de l'expérience.

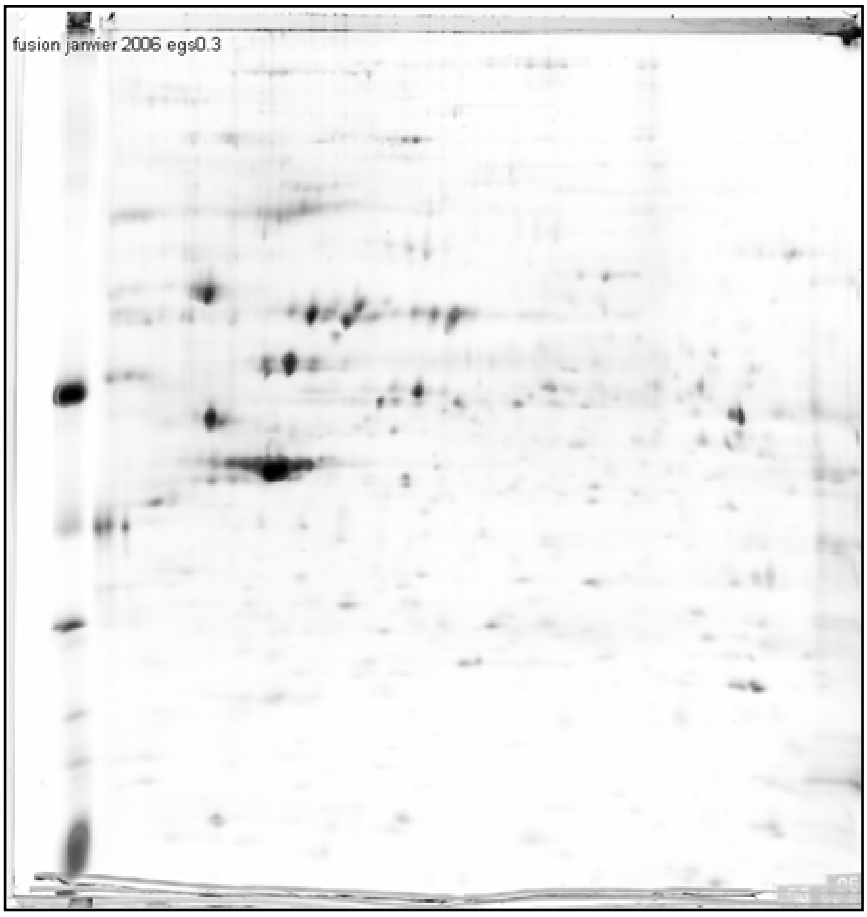


Figure 90: Image obtenue grâce à l'algorithme de fusion appliqué aux images associées au échantillon de tumeur et de muqueuse de l'expérience DIGE de janvier 2006.

De plus, comme l'algorithme utilisé est basé sur une moyenne pondérée, les taches protéiques, présentes sur cette image, conservent des contours et des reliefs harmonieux, sans épaulement ou discontinuité artificiels. Cette qualité de l'image de fusion permettra, par la suite, une détection automatique efficace. Les données de sorties de l'étape de fusion sont constituées d'une part des images égalisées :

> 06008 CLSP138 M3-dm-warped-egs0.tif
> 06008 CLSP138 T5-dm-warped-egs0.tif
> 06009 CLSP138 M5-dm-warped-egs0.tif
> 06009 CLSP138 T3-dm-warped-egs0.tif

> 06051 CLSP086 M3-dm-warped-egs0.tif
> 06051 CLSP086 T5-dm-warped-egs0.tif
> 06052 CLSP086 M5-dm-warped-egs0.tif
> 06052 CLSP086 T3-dm-warped-egs0.tif
> 06053 GHBD037 M3-dm-warped-egs0.tif
> 06053 GHBD037 T5-dm-warped-egs0.tif
> 06054 GHBD037 M5-dm-warped-egs0.tif
> 06054 GHBD037 T3-dm-warped-egs0.tif

et de l'image de fusion:

> Fusion Janvier 2006.tif

VI.6 Analyse des images

L'analyse d'image a pour but de quantifier l'intensité de chacune des taches protéiques présentes sur les images de l'expérience. L'intensité mesurée est le volume de la tache, c'est-à-dire la somme des niveaux de gris des pixels présents sur la surface de la tache. C'est à partir de ces données de volumes que sera faite l'analyse différentielle. Cette étape correspond donc à la traduction des images en informations quantitatives représentatives des différentes populations protéiques. Il s'agit d'une interprétation des images dans le sens où des connaissances a priori (paramètres de forme et nombre de spots) sont utilisées pour transformer un type de données en un autre.

VI.6.1 Les données d'entrée

L'analyse d'image nécessite l'ensemble des images égalisées issues du prétraitement ainsi que l'image de fusion obtenue à l'issue de l'étape précédente :

> 06008 CLSP138 Cy2_dm.tif
> 06008 CLSP138 M3_dm.tif
> 06008 CLSP138 T5_dm.tif
> 06009 CLSP138 Cy2_dm.tif
> 06009 CLSP138 M5_dm.tif
> 06009 CLSP138 T3_dm.tif
> 06051 CLSP086 Cy2_dm.tif
> 06051 CLSP086 M3_dm.tif
> 06051 CLSP086 T5_dm.tif
> 06052 CLSP086 Cy2_dm.tif
> 06052 CLSP086 M5_dm.tif
> 06052 CLSP086 T3_dm.tif
> 06053 GHBD037 Cy2_dm.tif
> 06053 GHBD037 M3_dm.tif
> 06053 GHBD037 T5_dm.tif
> 06054 GHBD037 Cy2_dm.tif
> 06054 GHBD037 M5_dm.tif
> 06054 GHBD037 T3_dm.tif

> Fusion Janvier 2006.tif

VI.6.2 Le traitement

Ce traitement est entièrement mené sous le logiciel ImageMaster 2D. Il consiste en l'application de la méthode du patron commun inter-gels présentée dans le paragraphe V.3.3.3 partie 2. Par cette méthode, les volumes des taches protéiques définies par le patron commun sont quantifiés sur chacune des images de l'expérience. Le traitement se déroule donc en deux étapes : la détermination du pa-

tron commun inter-gels et la quantification, à partir de ce patron, du volume de l'ensemble des taches protéiques.

VI.6.2.1 Détermination du patron commun

Le patron présenté en Figure 91 comprend les contours d'un maximum de taches protéiques. Il a été obtenu en appliquant l'algorithme de détection du logiciel Image Master 2D à l'image de fusion.

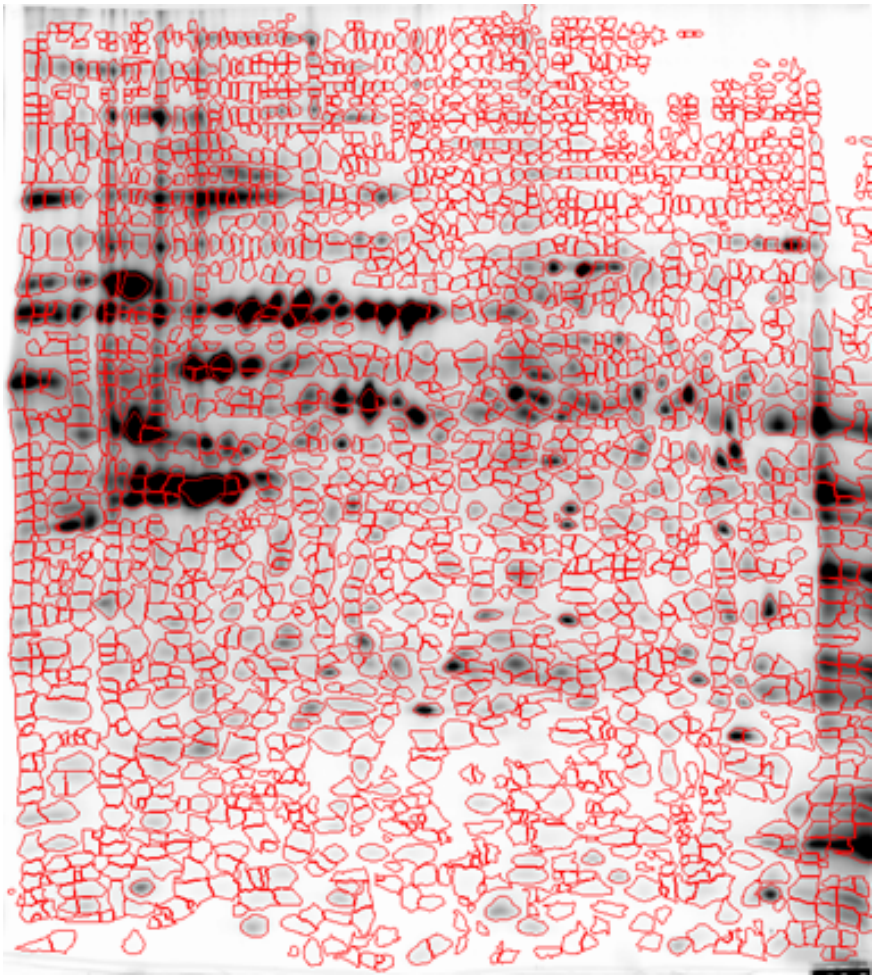


Figure 91 : Patron de détection obtenu grâce à l'utilisation de l'algorithme d'ImageMaster sur l'image de fusion.

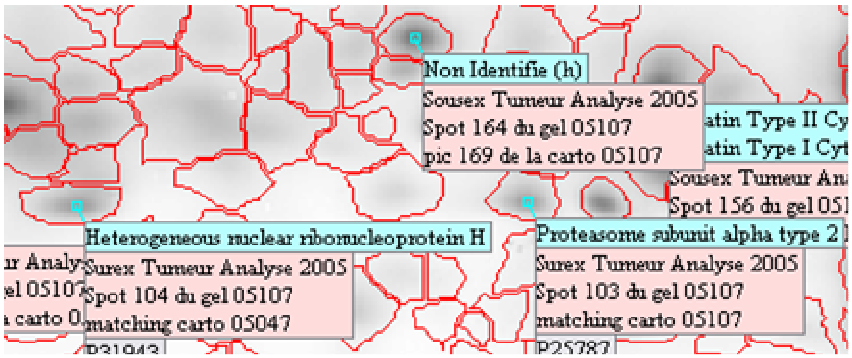


Figure 92 : Détail du patron commun détecté et annoté sur l'image de fusion.

Le patron est contrôlé et corrigé, si nécessaire, par le biologiste. Par ailleurs, comme le montrent la Figure 92 des annotations sont reportées sur ce patron commun. Les annotations sont les informations déjà connues telles que les noms des protéines, leur identifiant, ou bien encore, d'autres caractéristiques attribuées par le biologiste.

VI.6.2.2 Report du patron commun sur l'ensemble des images

Le patron commun est reporté sur l'ensemble des images prétraitées grâce aux fonctionnalités du logiciel ImageMaster. La qualité du calage du patron sur chacune des images dépend principalement de la qualité de l'alignement précédemment effectué dans l'analyse. L'observation de la Figure 93 permet de constater que, dans notre cas d'étude, le patron est parfaitement calé sur les taches protéiques, pour chacune des 12 images concernées par l'analyse différentielle. Quand ce n'est pas le cas il est possible de revenir sur l'alignement des images avec le logiciel TT900. Sur la base de ce patron commun, ImageMaster 2D détermine les quantifications volumiques des taches protéiques sur l'ensemble des images de l'expérience. Ces quantifications sont visualisables au sein du logiciel sous forme de tableaux et sont exportables sous forme de fichiers de données structurées de type XML.

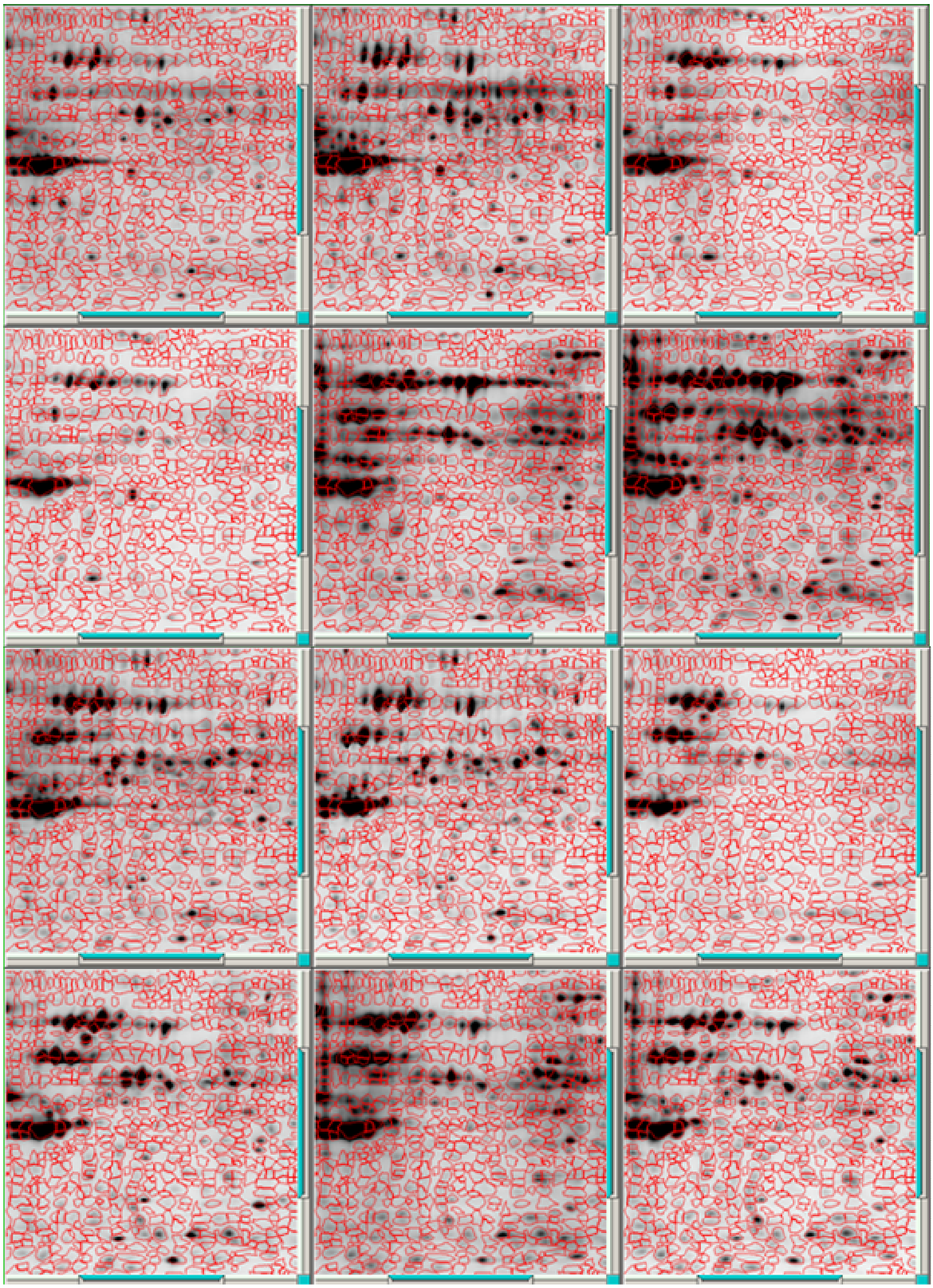


Figure 93: Illustration du report du patron. Détail de chacune des 12 images concernées par l'analyse différentielle. Quelle que soit l'image considérée, le patron est parfaitement calé sur les taches protéiques.

VI.6.3 Les données de sortie

Le fichier exporté à l'issue de l'analyse des images contient de nombreuses données et des informations, parmi lesquelles :

- les identifiants des taches protéiques (attribués par Image Master),
- les quantifications volumiques de chaque tache protéique,
- les coordonnées des centroïdes de chaque tache protéique,
- les contours des taches,
- et les annotations attribuées par le biologiste.

Ces données sont structurées au format XML dont la Figure 94 donne un aperçu.

```
<?xml version="1.0" encoding="iso-8859-1"?>
<Gels>
 <Fileinfo>
 <Gels_Data>
  <Version>1</Version>
  <Gel Id="78959aa6-5175-452d-95d1-77e01b8e78de" Cols="917" Ref.="0"
Rows="964" Class="" Spots="2199" IsDige="0" Caption="a" GelName=
"U:\ImageMaster\Matches\47920af2-3a81-4116-8f10-b8bb3b731b0d\Master_Matc
hSet Janvier 2006.mel" MaxGray="65535.0" MinGray="0" MatchSet=""
MaxValue="65535.0" MinValue="0" PixWidth="5000" Staining="" IsVirtual=
"1" PixHeight="5000" Gray_Slope="1.00000" Annotations="45" Gray_Offset=
"0" IsSynthetic="0" X_Axis_Unit="pI" Y_Axis_Unit="MW" Selected_Spots=
"0" Selected_Annotations="0">
   <Gel_Properties/>
   <Spots>
  </Gel>
  <Gel Id="0dbf6868-102b-4627-96f8-dc648846659b" Cols="917" Ref.="2"
Rows="964" Class="" Spots="2199" IsDige="0" Caption="b" GelName=
"P:\Labo\Manip\CCR\Tissus CCR\Traitement analyses DIGE\Analyse Juin
2006\Images warpées\Janv 2006\Gels ImageMaster\06008 CLSP138 M3-T5
Cy5(29s)_warped.mel" MaxGray="65535.0" MinGray="0" MatchSet="" MaxValue
="65535.0" MinValue="0" PixWidth="5000" Staining="" IsVirtual="0"
PixHeight="5000" Gray_Slope="1.00000" Annotations="45" Gray_Offset="0"
IsSynthetic="0" X_Axis_Unit="pI" Y_Axis_Unit="MW" Selected_Spots="0"
Selected_Annotations="0">
   <Gel_Properties/>
   <Spots>
    <Spot X="381" Y="708" Id="2600" Vol="1923118" Area="4550.00" Flag=
"0" X_Align="0" Y_Align="0" Saliency="0" ChainCode="376 704 5 4 5 4 5
6 6 6 0 0 0 0 0 0 0 0 0 0 0 0 0 7 6 7 0 7 0 0 1 0 0 0 0 0 1 2 2 2 3 2
3 4 4 4 4 4 4 4 4 4 4 4 4 3 4 3 4 4 4 " Intensity="1181.00"
Dige_Slope="0" Real_Val_X="-1.00000" Real_Val_Y="-1.00000" percentVol=
"0.0427296" Dige_Vol_Ratio="0" percentIntensity="0.0563817"
Dige_Vol_Ratio_Std="0"/>
    <Spot X="649" Y="683" Id="2601" Vol="788428" Area="4125.00" Flag=
```

Figure 94: Aperçu du fichier XML exporté depuis ImageMaster 2D et contenant l'ensemble des données nécessaires à l'analyse différentielle.

VI.7 Prétraitement des données volumiques : mise en forme et normalisation

Avant de pouvoir réaliser l'analyse différentielle, les données brutes des différentes images (volumes) doivent subir un prétraitement afin de les mettre en forme et de s'affranchir des biais systématiques qui les affectent. Cette étape du workflow est réalisée à partir des fonctions de ProDIGE implémentées sous Matlab.

VI.7.1 Les données d'entrée

Les informations contenues dans le fichier XML exporté d'ImageMaster à l'issue de l'analyse d'images sont immédiatement extraites sous Matlab afin de former des tableaux de données au format Excel. Les tableaux au format Excel ont le double avantage d'être facilement manipulables sous Matlab et d'être exploitables directement par l'utilisateur. L'analyse différentielle concerne le tableau des données volumiques brutes correspondant au fichier Excel principal issu de l'extraction et dont le nom est spécifié par l'utilisateur (ici « Analyse différentielle janvier 2006.xls »). Le Tableau 7 présente un aperçu du contenu et de la forme de ce fichier principal :

Master_MatchSet	CLSP138_1_M	CLSP138_1_T	CLSP138_2_M	CLSP138_2_T	CLSP086_1_M	CLSP086_1_T	CLSP086_2_M	CLSP086_2_T	GHBD037_1_M	GHBD037_1_T	GHBD037_2_M	GHBD037_2_T
2601	655553	1923118	672545	2663275	1293615	2645550	1023348	6742175	1876310	2610150	3751875	4491725
2602	580165	788428	1052805	1174393	1225100	1868323	761280	3987375	2154985	3178900	2168518	2005833
2603	311298	645205	341420	500858	1583920	1036872	593505	983998	2109553	2443275	1095055	841738
2604	192028	549758	350395	1372865	711178	887155	420050	916448	1710782	1587680	1956705	882480
2605	277670	293785	616998	920805	216776	411850	312835	885998	582960	1284945	1512125	2029060
2606	447138	462155	848553	1331220	728458	614703	738828	842760	775070	1528478	734658	939095
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
4793	98789.3	38829.3	44904.3	120487	141283	131954	807190	176279	110312	100943	519185	128486
4794	120080	163849	97816	195396	171816	145922	2600625	450547	224010	166131	423532	116461
4795	62202	43459.5	63501	127231	141572	67759.5	743772	132275	65589.8	25562	271020	55018.8
4796	85051.5	105511	85940	205218	342888	91624	1102100	124015	91711	100852	800532	118318
4797	98141.3	163881	57870.8	152908	206391	584263	938495	297085	119161	279560	450605	172875
4798	96953.8	207944	113798	131270	153074	67635.8	114811	157546	920415	703095	190054	179565
4799	290480	311393	326050	854080	481868	228419	4287250	861953	412605	341873	2703575	1576140

Tableau 7: Aperçu du tableau des données volumiques brutes

En plus de ce tableau des données volumiques brutes, un fichier Excel supplémentaire est créé, il porte le nom spécifié par l'utilisateur augmenté du suffixe « _rawdata ». Il contient les données brutes déjà contenues dans le fichier principal ainsi que des données telles que les contours des spots qui ne sont pas directement utiles à l'utilisateur, mais qui seront exploités par l'outil proDIGE.

VI.7.2 Application de la normalisation des quantifications volumiques

L'outil proDIGE propose tout d'abord de préciser la structure DIGE des colonnes des différents tableaux de données. Cette étape est indispensable car, par la suite, les méthodes de normalisation et de fouille des données exploitent cette information.

Une fois les données mises en forme, proDIGE permet d'observer les effets des deux principales méthodes de normalisation présentées dans le paragraphe V.3.3.4, sur le nuage des points des graphes MA (ratio volumique des taches protéiques en fonction de leur intensité moyenne), pour chacun des 6 gels de l'expérience. Ces observations sont regroupées sur la Figure 95 et la Figure 96. Comme nous l'évoquions dans le paragraphe V.3.3.4, l'usage de la régression LOESS pour la correction des données est assez risqué notamment pour les points situés dans les intensités extrêmes, qui, du fait de leur rareté influencent trop la régression. Par ailleurs, on constate que, mis à part pour ces valeurs extrêmes d'intensité, les deux régressions sont très proches. Il apparaît donc judicieux d'appliquer la correction LTS. Le résultat de la correction permet de constater le rééquilibrage des nuages de points conformément aux hypothèses d'auto-cohérence des données. Cet effet est particulièrement visible pour le gel n°2 (06009) de la Figure 95.

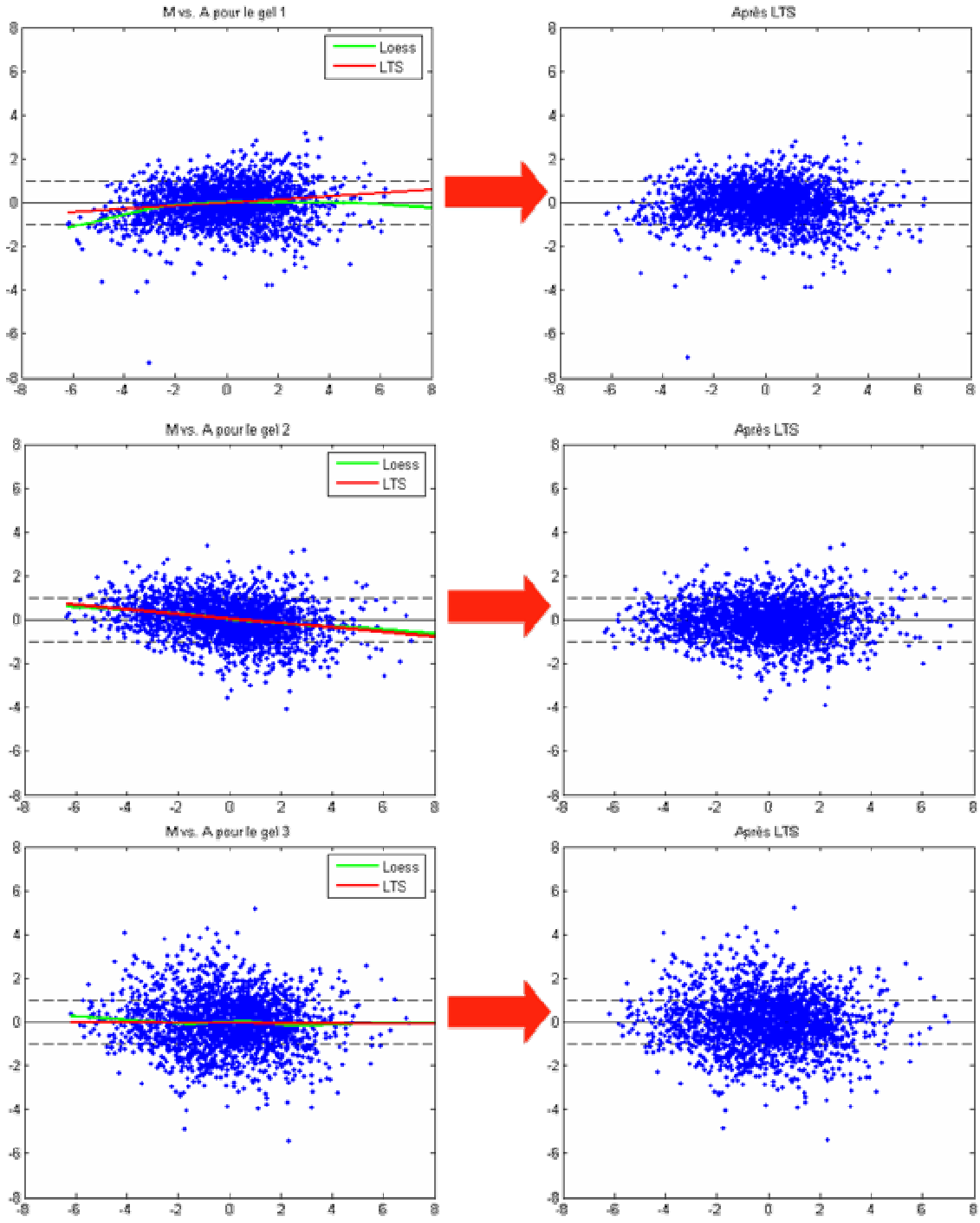


Figure 95: Nuages de points des ratios en fonction de l'intensité, pour chacun des 6 premiers gels de l'expérience de juin 2006. Sur les graphiques de gauches, il s'agit des données brutes à partir desquelles ont été réalisées les régressions LOESS et LTS (visibles en vert et en rouge). Les graphiques de gauche sont le résultat de l'application de la correction par régression LTS.

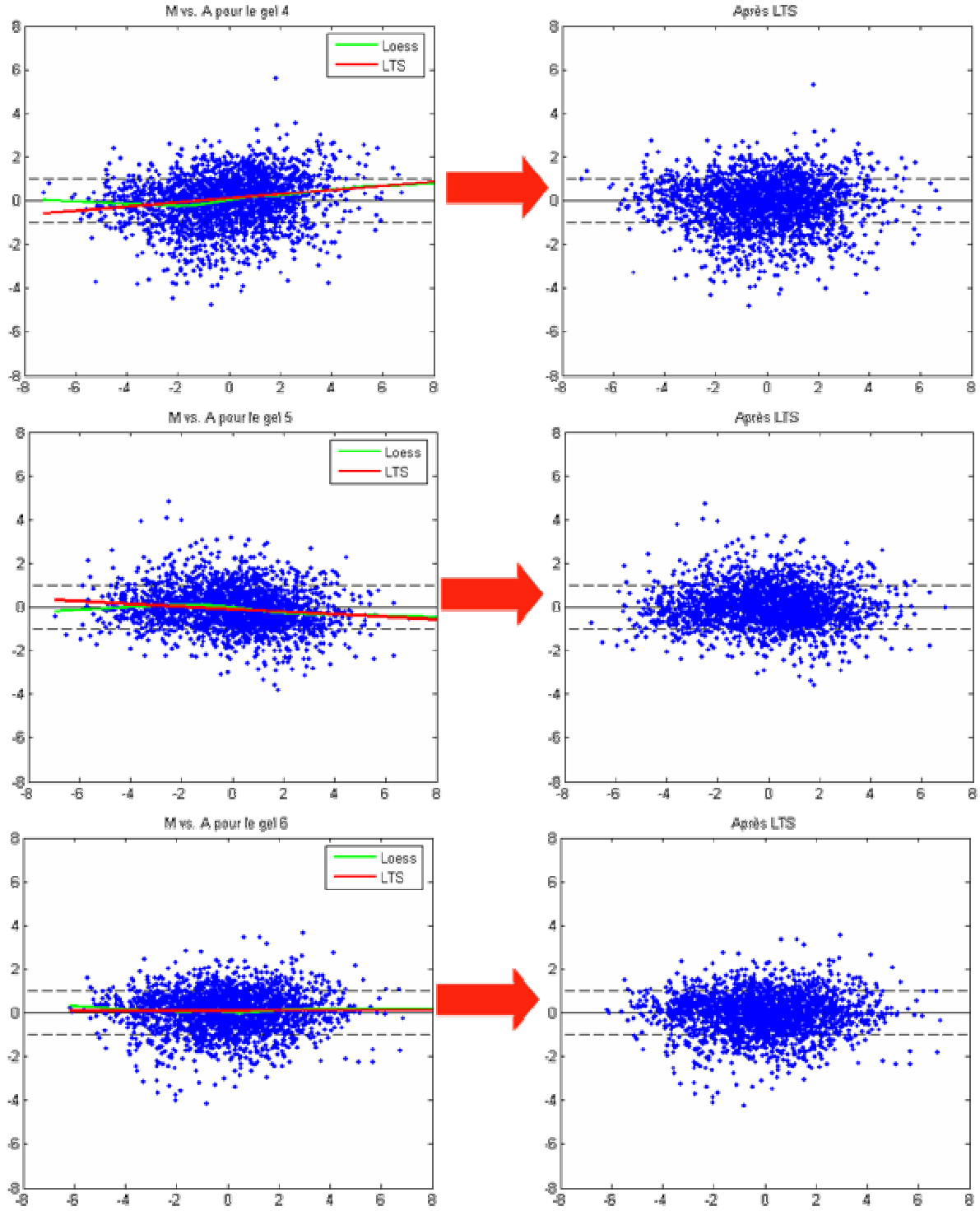


Figure 96 : Nuages de points des ratios en fonction de l'intensité, pour chacun des 6 autres gels de l'expérience de juin 2006. Sur les graphiques de gauches, il s'agit des données brutes à partir desquelles ont été réalisées les régressions LOESS et LTS (visibles en vert et en rouge). Les graphiques de gauche sont le résultat de l'application de la correction par régression LTS.

VI.7.3 Les données de sortie

Les données volumiques obtenues après la normalisation LTS sont sauvegardées sous la forme d'un tableau contenu dans un nouvel onglet du fichier Excel principal des données brutes et dont le Tableau 8 est un aperçu. Cette fois ce sont les valeurs du logarithme base 2 des volumes normalisés qui sont sauvegardées.

Master_MatchSet	CLSP138_1_M	CLSP138_1_T	CLSP138_2_M	CLSP138_2_T	CLSP086_1_M	CLSP086_1_T	CLSP086_2_M	CLSP086_2_T	GHBD037_1_M	GHBD037_1_T	GHBD037_2_M	GHBD037_2_T
2601	0.17961	1.30714	-0.03461	0.98719	0.13794	1.21753	0.13139	2.26075	0.3599	0.4639	0.8775	1.55317
2602	0.00336	0.02074	0.61192	-0.1941	0.05943	0.71571	-0.29541	1.50298	0.55968	0.7483	0.0866	0.3901
2603	-0.89481	-0.26848	-1.01269	-1.42354	0.43003	-0.1338	-0.65458	-0.51574	0.52893	0.36858	-0.89911	-0.86266
2604	-1.59179	-0.49944	-0.97526	0.03118	-0.72519	-0.35878	-1.15328	-0.61834	0.22665	-0.25331	-0.06168	-0.79447
2605	-1.05973	-1.40348	-0.15898	-0.54504	-2.43919	-1.46585	-1.57844	-0.66709	-1.32653	-0.55853	-0.43353	0.40671
2606	-0.37238	-0.74986	0.30076	-0.01326	-0.69055	-0.88808	-0.3386	-0.73927	-0.9156	-0.30814	-1.47497	-0.70476
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
4793	-2.55068	-4.32302	-3.93932	-3.47906	-3.05681	-3.10793	-0.21093	-2.99653	-3.72834	-4.22862	-1.97579	-3.57442
4794	-2.26911	-2.24587	-2.8161	-2.78154	-2.77453	-2.96277	1.47695	-1.64271	-2.70637	-3.50984	-2.26957	-3.71619
4795	-3.21807	-4.1605	-3.43939	-3.40049	-3.05386	-4.06947	-0.32898	-3.41085	-4.47839	-6.21009	-2.91364	-4.79804
4796	-2.76669	-2.88085	-3.00284	-2.71078	-1.77766	-3.63417	0.23834	-3.50388	-3.99476	-4.22992	-1.35108	-3.69336
4797	-2.56017	-2.24559	-3.57333	-3.13528	-2.51002	-0.96135	0.00651	-2.24352	-3.61702	-2.759	-2.18017	-3.1463
4798	-2.57773	-1.90204	-2.59777	-3.3554	-2.94117	-4.07211	-3.02458	-3.15862	-0.66765	-1.42844	-3.42563	-3.09153
4799	-0.99466	-1.3195	-1.07915	-0.65357	-1.28676	-2.31628	2.19814	-0.70678	-1.82517	-2.4687	0.40476	0.04229

Tableau 8 : Aperçu du tableau des données après normalisation. Il s'agit des valeurs logarithmiques base 2 des quantifications normalisées.

VI.8 Fouille des données par analyse différentielle des données volumiques normalisées

La normalisation a permis de s'affranchir des biais systématiques et de rendre ainsi les différentes populations comparables, il s'agit maintenant de considérer les classes de données (témoin et pathologique) et de réaliser l'analyse différentielle. Cette fouille des données doit permettre la découverte des protéines dont les expressions sont significativement différentes d'une classe à l'autre. L'outil ProDIGE permet d'appliquer les deux approches décrites lors de la mise en place du workflow, dans le paragraphe V.3.3.5. Les intérêts spécifiques à chacune de ces deux méthodes seront décrits au moment de s'interroger sur la pertinence des résultats, dans le paragraphe VI.9.

VI.8.1 Données d'entrée

L'analyse différentielle des données concerne les données volumiques normalisées. L'outil ProDIGE fait également appelle aux informations de contours des taches protéiques ainsi qu'aux images dont le contraste a été égalisé lors de l'étape de fusion, afin d'afficher les taches protéiques identifiées comme marqueurs potentiels.

VI.8.2 1ère méthode : fouille des données à partir d'une courbe enveloppe et de l'intérêt biologique

Cette première méthode cherche à évaluer la variabilité expérimentale du logarithme base 2 du ratio des quantifications volumiques des taches protéiques considérées dans notre cas d'étude. Elle permet également de prendre en considération la volonté du biologiste d'accepter un risque plus ou moins grand dans la désignation des différences d'expression selon qu'il s'agit de protéines plus ou moins fortement exprimées. Ceci est rendu possible grâce à la définition par le biologiste d'un profil de risque fonction de l'intensité (logarithme base 2 des volumes normalisés). Le profil est déterminé à partir de 2 valeurs de risque associées aux 2 valeurs d'intensité minimale et maximale du jeu de donnée considéré. Dans notre cas d'étude, les deux valeurs choisies sont *pmin=0.02* pour les basses intensités, et *pmax=0.08* pour les hautes intensités. Le profil de risque associé est donné en Figure 97.

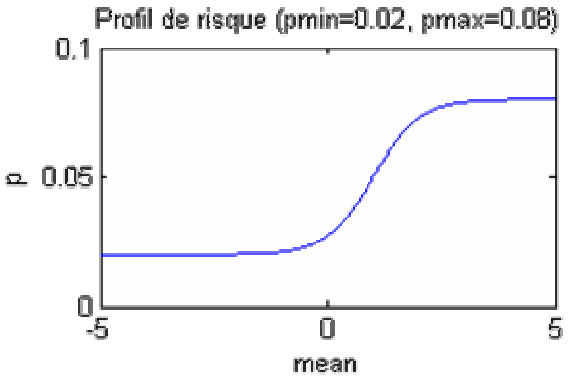


Figure 97: Profil de risque défini par le biologiste pour l'étude du jeu de données de janvier 2006.

L'ensemble des données de répétitions disponibles est compilé sur un même graphe « M vs A » (ratio en fonction de l'intensité) et permet l'estimation d'une courbe enveloppe associée au profil de risque. Cette courbe, qui définit la limite de significativité du ratio en fonction de l'intensité, n'a donc un sens qu'au regard du profil de risque. Le graphe « M vs A » des données de répétition et la courbe enveloppe estimée sont présentés en Figure 98.

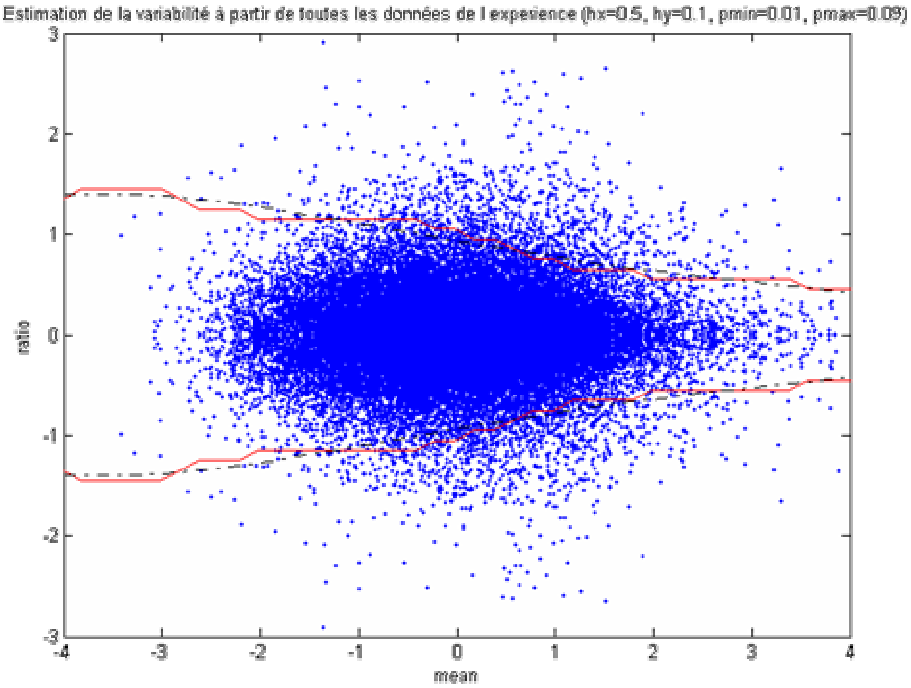


Figure 98: Graphe M vs A associé aux données de répétition de l'expérience de Janvier 2006 et courbe enveloppe (en rouge) définissant la limite de significativité associée au profil de risque défini par le biologiste.

La courbe enveloppe estimée est ensuite utilisée, pour chacun des 6 gels, afin de déterminer les taches protéiques dont l'expression est significative au sens défini par le profil de risque. Comme le montre la Figure 99, la méthode permet de mettre en évidence de 13 à 88 taches protéiques d'intérêt selon le gel considéré (voir Tableau 9).

	sous-expressions	sur-expressions
CLSP138 gel1	33	13
CLSP138 gel2	24	21
CLSP086 gel1	57	46
CLSP086 gel2	88	26
GHBD037 gel1	30	30
GHBD037 gel2	35	13

Tableau 9 : Nombre de taches protéiques sortant de la courbe enveloppe en sous-expression et en sur-expression, pour chacun des 6 gels de janvier 2006.

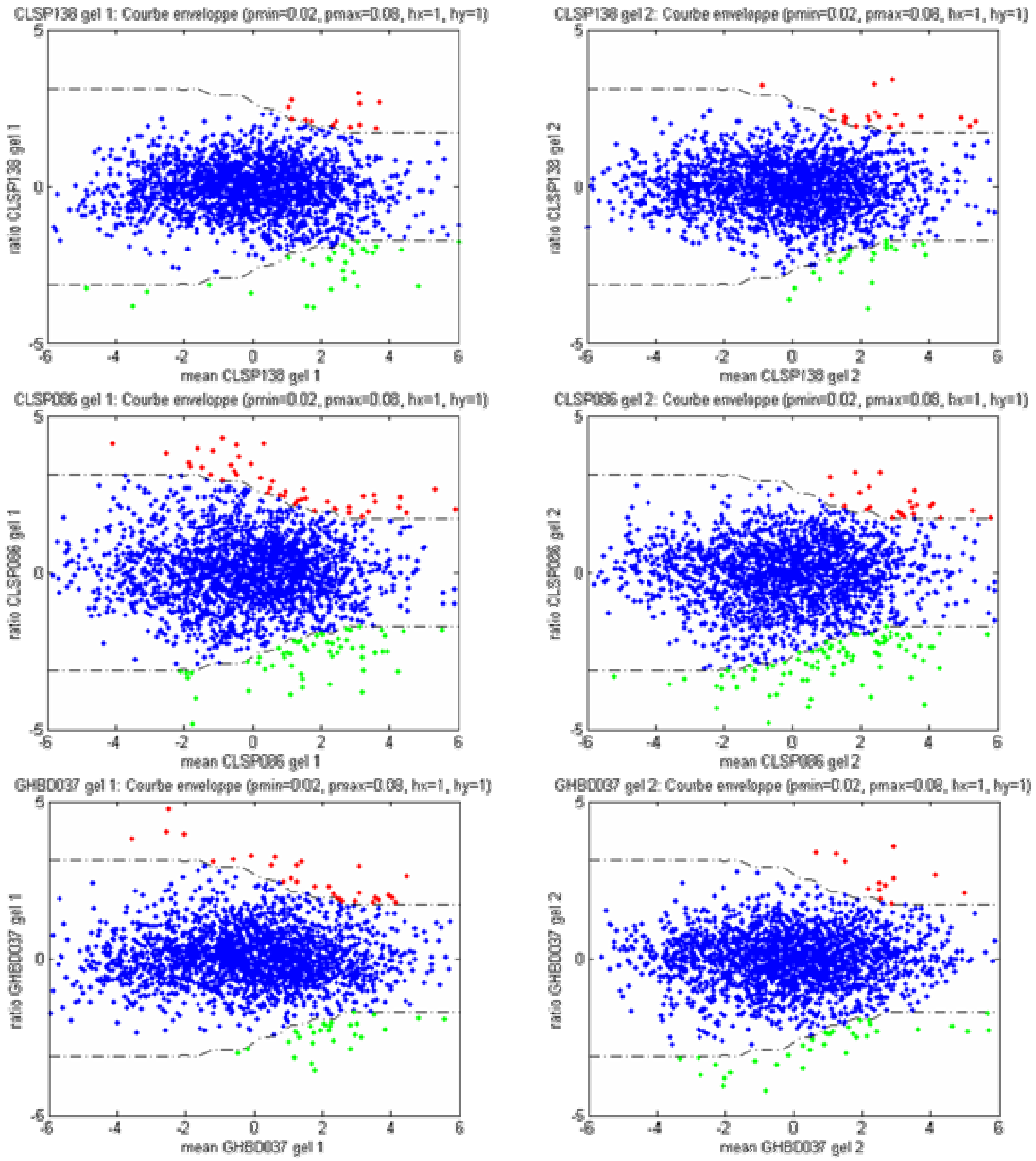


Figure 99 : Application de la courbe enveloppe au graphe « M vs A » de chacun des gels de janvier 2006. Les points en rouge et les points verts correspondent respectivement aux taches protéiques à considérer comme des sur-expressions et comme des sous-expressions significatives pour le biologique.

Afin de pouvoir résumer l'information issue de l'application de la courbe enveloppe, on sépare les taches protéiques en 3 classes. Pour cela, une valeur est attribuée à chacune des taches protéiques de chaque gel selon la position par rapport à la courbe enveloppe sur le graphe « M vs A » associé.

Lorsque la tache protéique se situe en dessous de la courbe enveloppe, il s'agit d'une sous-expression de la protéine dans l'échantillon tumeur par rapport à l'échantillon muqueuse et la valeur attribuée est -1.

Lorsque la tache protéique se situe à l'intérieur de l'enveloppe, la différence d'expression entre l'échantillon tumeur et l'échantillon muqueuse est considérée comme non significative et la valeur attribuée est 0.

Enfin, lorsque la tache se situe au-dessus de la courbe enveloppe, il s'agit d'une sur-expression de la protéine dans l'échantillon tumeur par rapport à l'échantillon muqueuse et la valeur attribuée est +1.
Un score est obtenu, en sommant, pour chaque tache protéique, les valeurs ainsi prises sur chacun des gels. Ce score permet d'isoler les taches protéiques dont la différence d'expression est la plus reproductible. L'expérience de janvier 2006 comportant 6 gels, un score supérieur ou égal à 3, en valeur absolue, signifie que la tache protéique concernée est jugée significative sur au moins la moitié des gels. Le Tableau 10 présente les identifiants des taches protéiques dont le score est au moins égal à 3. Ces taches protéiques peuvent alors être considérées comme des marqueurs potentiels de la pathologie. Cependant, le croisement des informations réduit considérablement le nombre de marqueurs potentiels à considérer. Dans notre cas de figure, le biologiste ne peut pas se contenter des 12 taches protéiques en sous-expressions et des 7 en sur-expressions déclarées significatives. La solution d'abaisser le seuil à 2 n'est pas satisfaisante car les taches protéiques retenues sont alors trop nombreuses et il n'existe aucune hiérarchie entre elles permettant de se limiter aux plus prometteuses.

	Sous-expressions								Sur-expressions							
Identifiant	CLSP138 gel1	CLSP138 gel2	CLSP086 gel1	CLSP086 gel2	GHBD037 gel1	GHBD037 gel2	**Score**	**Identifiant**	CLSP138 gel1	CLSP138 gel2	CLSP086 gel1	CLSP086 gel2	GHBD037 gel1	GHBD037 gel2	**Score**
3804	-1	-1	0	-1	-1	-1	**-5**	**2992**	1	1	0	1	1	1	**5**
2651	-1	-1	-1	-1	0	0	**-4**	**3789**	1	1	1	1	1	0	**5**
3197	-1	-1	-1	-1	0	0	**-4**	**3830**	1	1	0	0	1	1	**4**
3505	-1	-1	-1	-1	0	0	**-4**	**2622**	0	1	1	1	0	0	**3**
3618	-1	-1	-1	0	-1	0	**-4**	**2633**	0	0	1	1	0	1	**3**
4005	-1	0	0	-1	-1	-1	**-4**	**2995**	0	1	0	0	1	1	**3**
3148	0	0	-1	0	-1	-1	**-3**	**3012**	0	1	1	1	0	0	**3**
3180	0	-1	-1	-1	0	0	**-3**								
3359	-1	0	-1	-1	0	0	**-3**								
3401	-1	0	-1	-1	0	0	**-3**								
3674	-1	0	0	0	-1	-1	**-3**								
3806	-1	-1	0	0	0	-1	**-3**								

Tableau 10: Récapitulatif des taches protéiques dont les différences d'expression biologique (-1 pour les sous-expressions, +1 pour les sur-expressions) sont les plus reproductives. Le score est la somme des valeurs de la nature de l'expression différentielle.

Comme nous l'avons vu au paragraphe V.3.3.5.2b), la mise en place du paramètre d'intérêt biologique permet d'affiner la recherche des marqueurs potentiels. Il ne s'agit plus de considérer un seuil d'intérêt mais plutôt une information quantitative reflétant l'intérêt à apporter à chacune des taches protéiques.
Le croisement des informations entre chacun des gels de l'expérience peut alors être réalisé de manière plus fine à partir du score obtenu en sommant, pour chaque tache protéique, les 6 valeurs d'intérêt biologique évaluées sur les 6 gels de l'expérience. L'intérêt biologique prend des valeurs dans l'intervalle [-1,1]. Les taches protéiques dont la valeur de l'intérêt biologique est proche de -1 sont en sous-expression tandis que celles dont la valeur est proche de +1 sont en sur-expression Cette nouvelle approche permet d'établir une hiérarchie des taches protéiques qui tient compte, à la fois de l'intérêt biologique (c'est-à-dire de la valeur du ratio et de l'intensité), et de la reproductibilité sur les différents gels de l'expérience. Le Tableau 11, présente les

identifiants des taches protéiques dont la valeur absolue du score dépasse 4. Le choix de la valeur 4 est arbitraire. En effet, comme nous l'avions remarqué lors de sa définition, la valeur d'intérêt biologique n'est pas directement interprétable en terme de statistique. L'intérêt du paramètre d'intérêt biologique réside dans la constitution d'un classement des taches protéiques selon l'intérêt que le biologiste doit leur porter.

	Sous-expressions								Sur-expressions						
Identifiant	CLSP138 gel1	CLSP138 gel2	CLSP086 gel1	CLSP086 gel2	GHBD037 gel1	GHBD037 gel2	Score	Identifiant	CLSP138 gel1	CLSP138 gel2	CLSP086 gel1	CLSP086 gel2	GHBD037 gel1	GHBD037 gel2	Score
3505	-0.9	-0.9	-1.0	-1.0	-0.9	-0.8	**-5.5**	**2992**	1.0	1.0	0.9	0.9	0.9	1.0	**5.7**
2651	-1.0	-1.0	-1.0	-1.0	-0.6	-0.8	**-5.4**	**3789**	1.0	1.0	1.0	0.9	0.9	0.8	**5.6**
3197	-0.9	-0.9	-1.0	-1.0	-0.8	-0.8	**-5.4**	**2633**	0.9	0.9	1.0	1.0	0.8	1.0	**5.6**
3618	-0.9	-0.9	-0.9	-0.8	-0.9	-0.8	**-5.3**	**2622**	0.8	1.0	1.0	0.9	0.7	0.8	**5.2**
3180	-0.8	-0.9	-0.9	-0.9	-0.8	-0.8	**-5.2**	**2621**	0.8	0.7	0.9	0.8	0.9	0.9	**5.0**
4005	-0.9	-0.8	-0.7	-0.9	-0.9	-1.0	**-5.2**	**2995**	0.8	1.0	0.6	0.6	0.9	0.9	**4.8**
3807	-0.9	-0.9	-0.4	-0.6	-0.7	-0.9	**-4.4**	**3822**	0.6	0.8	0.9	0.9	0.8	0.7	**4.8**
4676	-0.7	-0.4	-0.8	-0.6	-1.0	-1.0	**-4.4**	**3811**	0.7	0.6	0.9	1.0	0.8	0.7	**4.6**
3401	-1.0	-0.4	-1.0	-1.0	-0.8	-0.2	**-4.4**	**2993**	0.8	0.8	0.6	0.7	0.8	0.8	**4.6**
3095	-0.6	-0.7	-0.9	-0.6	-0.8	-0.8	**-4.4**	**2619**	0.9	0.8	0.9	0.9	0.2	0.8	**4.5**
3804	-0.9	-1.0	0.3	-0.9	-0.9	-0.9	**-4.4**	**3707**	0.7	0.8	0.6	0.9	0.7	0.8	**4.5**
3619	-0.7	-0.8	-0.6	-0.9	-0.7	-0.7	**-4.4**	**3834**	0.9	1.0	0.5	0.6	0.7	0.6	**4.3**
3417	-0.8	-0.9	-0.7	-0.8	-0.6	-0.5	**-4.4**	**3045**	0.6	0.6	0.6	0.8	0.8	0.9	**4.2**
3148	0.0	-0.9	-0.9	-0.7	-0.9	-0.9	**-4.3**	**3050**	0.7	0.5	0.8	0.4	0.8	0.8	**4.2**
2751	-0.8	-0.5	-0.9	-0.9	-0.6	-0.4	**-4.1**	**3267**	0.5	0.7	0.9	1.0	0.4	0.6	**4.1**
2614	-0.5	-0.7	-0.7	-0.9	-0.5	-0.8	**-4.1**	**3814**	0.8	0.1	1.0	0.9	0.9	0.4	**4.1**
3503	-0.6	-0.6	-0.9	-0.9	-0.7	-0.3	**-4.1**	**3462**	0.3	0.1	0.8	0.8	1.0	1.0	**4.0**

Tableau 11: Récapitulatif, des taches protéiques dont les différences d'expression biologiques sont les plus reproductives. Le score attribué à chaque tache protéique est la somme des valeurs de l'intérêt biologique estimées sur les 6 gels de l'expérience.

L'observation du Tableau 10 et du Tableau 11 permet d'apprécier l'apport d'une compilation inter-gels des valeurs d'intérêt biologique par rapport à la compilation des informations par classe (-1, 0 ou 1) issue de l'application directe de la courbe enveloppe. Le score calculé à partir des valeurs de l'intérêt biologique permet un classement beaucoup plus fin et la prise en compte de taches protéiques ne sortant pas forcément de la courbe enveloppe pour plus de 3 gels mais dont les valeurs de l'intérêt biologique sont tout de même élevées et peu variables.

Afin de cerner les différences et les convergences entre les 2 méthodes de compilation, il est intéressant de situer les taches protéiques retenues du Tableau 10 dans le classement établi grâce aux scores d'intérêt biologique. Pour cela, les taches protéiques retenues du Tableau 10 sont situées parmi les 30 premières taches protéiques les plus significatives en sur-expression et en sous-expression au sens du score d'intérêt biologique. Le Tableau 12, présente le résultat de ce rapprochement.

Sousex				Surex		
Classements des sous-expressions				Classements des sur-expressions		
Identifiants	Classement selon le score d'intérêt biologique	Classement selon le score courbe enveloppe		Identifiants	Classement selon le score d'intérêt biologique	Classement selon le score courbe enveloppe
3505	1	4		2992	1	1
2651	2	2		3789	2	2
3197	3	3		2633	3	5
3618	4	5		2622	4	4
3180	5	8		2995	6	6
4005	6	6		2621	5	Absent
3401	9	10		3822	7	Absent
3804	11	1		3811	8	Absent
3148	14	7		2993	9	Absent
3807	7	Absent		2619	10	Absent
4676	8	Absent		3707	11	Absent
3095	10	Absent		3834	12	Absent
3619	12	Absent		3045	13	Absent
3417	13	Absent		3050	14	Absent
2751	15	Absent		3267	15	Absent
2614	16	Absent		3814	16	Absent
3503	17	Absent		3462	17	Absent
4006	18	Absent		3487	18	Absent
3499	19	Absent		3486	19	Absent
2737	20	Absent		2982	20	Absent
3820	21	Absent		3327	21	Absent
4670	22	Absent		3422	22	Absent
2731	23	Absent		3406	23	Absent
4410	24	Absent		3361	24	Absent
2965	25	Absent		2618	25	Absent
4683	26	Absent		3496	26	Absent
4679	27	Absent		3064	27	Absent
3263	28	Absent		3055	28	Absent
4414	29	Absent		3685	29	Absent
3724	30	Absent		3268	30	Absent
3359	Absent	9		3830	Absent	3
3674	Absent	11		3012	Absent	7
3806	Absent	12				

Tableau 12

Sur les 12 taches en sous-expression retenues par la compilation des informations de la courbe enveloppe, 9 se retrouvent parmi les 14 premières selon le score de l'intérêt biologique. En sur-expression, 5 taches se retrouvent parmi les 6 premières. Les taches singulières sont celles retenues par la compilation des informations de la courbe enveloppe et qui ne sont pas présentes parmi les 30 premières taches protéiques les plus significatives au sens du score d'intérêt biologique. Ces taches sont au nombre de 3 en sous-expression et au nombre de 2 en sur-expression. Les valeurs de l'intérêt biologique ainsi que le classement de chacune de ces taches sont présentés dans le Tableau 13. On peut voir que les 3 spots 2992, 3789, 2633 et 2622 semblent intéressants car ils sont très bien classés par les 2 méthodes. Les taches protéiques 3674,

3806 et 3830 désignées comme marqueur potentiel après compilation des informations apportées par la courbe enveloppe ne sont pas présente présentes parmi les 30 premières dans le classement selon le score d'intérêt biologique, mais elle sont tout de même bien classées (respectivement 33ème, 34ème et 39ème). Concernant les taches protéiques 3359 et 3012, les grands écarts de classement sont dus aux valeurs prises par l'intérêt biologique sur certains gels. Par exemple, dans le cas de la tache 3012, les valeurs de l'intérêt biologique évaluées sur les 2 gels du patient GHBD037 correspondent à une sous-expression alors que sur les 4 autres gels, les valeurs semblait indiquer la présence d'une sur-expression assez nette. Le fait de sommer les 6 valeurs permet de tenir compte de cette contradiction en éloignant la protéine dans le classement. Ce n'est pas le cas lors de la compilation des valeurs de classes associées à la courbe enveloppe, car la tache protéique se situe à l'intérieur de la courbe enveloppe et se voit attribuer la valeur 0 pour chacun des 2 gels du patient GHBD037.

Sousex								
	Intérêt biologique							
Identifiant	CLSP138 gel1	CLSP138 gel2	CLSP086 gel1	CLSP086 gel2	GHBD037 gel1	GHBD037 gel2	**Score**	Classement
3359	-1	-0.2	-1	-1	-0.4	0.28	**-3.3**	47
3674	-0.9	-0.7	0.08	-0.2	-0.9	-0.9	**-3.5**	33
3806	-0.8	-0.9	0.18	-0.2	-0.9	-0.9	**-3.5**	34

Surex								
	Intérêt biologique							
Identifiant	CLSP138 gel1	CLSP138 gel2	CLSP086 gel1	CLSP086 gel2	GHBD037 gel1	GHBD037 gel2	**Score**	Classement
3012	0.89	0.96	0.96	0.96	-0.6	-0.6	**2.6**	122
3830	0.96	0.99	-0.6	0.21	0.97	0.98	**3.5**	39

Tableau 13 : Détails des valeurs de l'intérêt biologique pour les taches protéiques retenues parmi les plus intéressantes selon l'approche « courbe enveloppe » mais non présentes parmi les 30 plus intéressantes au sens du score d'intérêt biologique.

À la lumière du Tableau 12 et du Tableau 13 il est clairement préférable d'utiliser le score basé sur la somme des intérêts biologiques plutôt que de comptabiliser le nombre de gels pour lesquels la tache considérée est au-dessus ou au-dessous de la courbe enveloppe. Cela correspond à la volonté de conserver l'aspect quantitatif de l'information le plus loin possible dans l'analyse. Cependant, l'application de la courbe enveloppe reste intéressante lorsque l'on s'intéresse à chacun des gels séparément et que l'on désire des résultats statistiquement interprétables.

VI.8.3 2ème méthode : fouille des données par t-test régulé

La seconde méthode proposée par l'outil ProDIGE et appliquée à notre cas d'étude, est celle définie dans le paragraphe 1.V.3.3.5.1. Elle consiste en un test de Student avec régulation de la variance. Cette méthode permet, d'une part, de travailler sur des populations restreintes (dans notre cas nous n'avons que 6 gels) et, d'autre part, d'intégrer dans l'estimation de la significativité, la dépendance existante entre la variabilité de l'expression et le niveau d'intensité de l'expression. L'application de cette méthode nécessite en premier lieu de définir deux paramètres.

Le premier est le nombre de pseudo observations à considérer (voir détails dans 1.V.3.3.5.1). Sachant que cette expérience est composée de 6 gels, c'est-à-dire de 6 observations réelles, on choisit de considérer 4 pseudo observations afin d'atteindre un total de 10. De cette manière, la variance calculée est plus robuste et les résultats pourront être rapprochés de manière plus légitime de ceux d'une expérience comprenant 10 observations réelles.

Le deuxième paramètre sert à définir la taille de la fenêtre pour l'estimation de la variance de fond (voir détails dans 1.V.3.3.5.1). Plus ce nombre est grand plus on néglige une éventuelle dépendance de la variance en fonction de l'intensité qui est la plupart du temps bien réelle. Plus ce nombre est faible moins l'estimation est robuste (l'estimation est alors bruitée). L'usage montre qu'un compromis est possible en prenant ~10% du nombre total de taches protéiques. Notre jeu de données comprend 2119 taches protéiques, ce qui nous amène à définir une fenêtre de 200 points.

La Figure 100 présente la variance de fond ainsi que la variance régulée calculées à partir de notre jeu de données. La dépendance de la variabilité du ratio en fonction de l'intensité est faible : pour les faibles expressions les ratios sont légèrement plus variables que pour les fortes expressions.

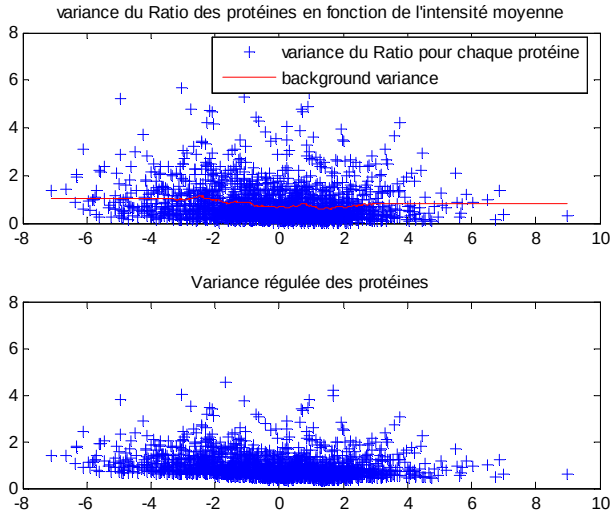


Figure 100 : Visualisation de la variance du ratio des quantifications volumiques normalisées en fonction de l'intensité, avant (en haut) et après (en bas) correction. Cette correction est basée sur la variance de fond estimée (en rouge)

Les valeurs du t-test sont calculées à partir des valeurs de la variance régulées. Il s'agit alors de fixer un seuil de significativité. Ce paramètre est à fixé en accord avec le biologiste, il représente le risque de se tromper en affirmant qu'une protéine (associée à l'une des taches protéiques considérées) est réellement exprimée différentiellement. Les taches protéiques dont la p-value estimée est inférieure ou égale à ce seuil seront reconnus comme différentiellement exprimés. Étant donnée la grande variabilité biologique et expérimentale (reconnue comme telle dans la littérature) des expériences DIGE, il faut surtout considérer ce paramètre comme un seuil fixant le nombre de marqueurs potentiels que le biologique désire étudier. Dans le cas de notre étude différentielle nous considérons le seuil de 0.05.

La Figure 101 présente une des images du jeu de données de juin 2006 (choisie arbitrairement) sur laquelle l'outil proDIGE matérialise la position des taches protéiques retenues comme marqueur potentiel, ainsi que leur identifiant.

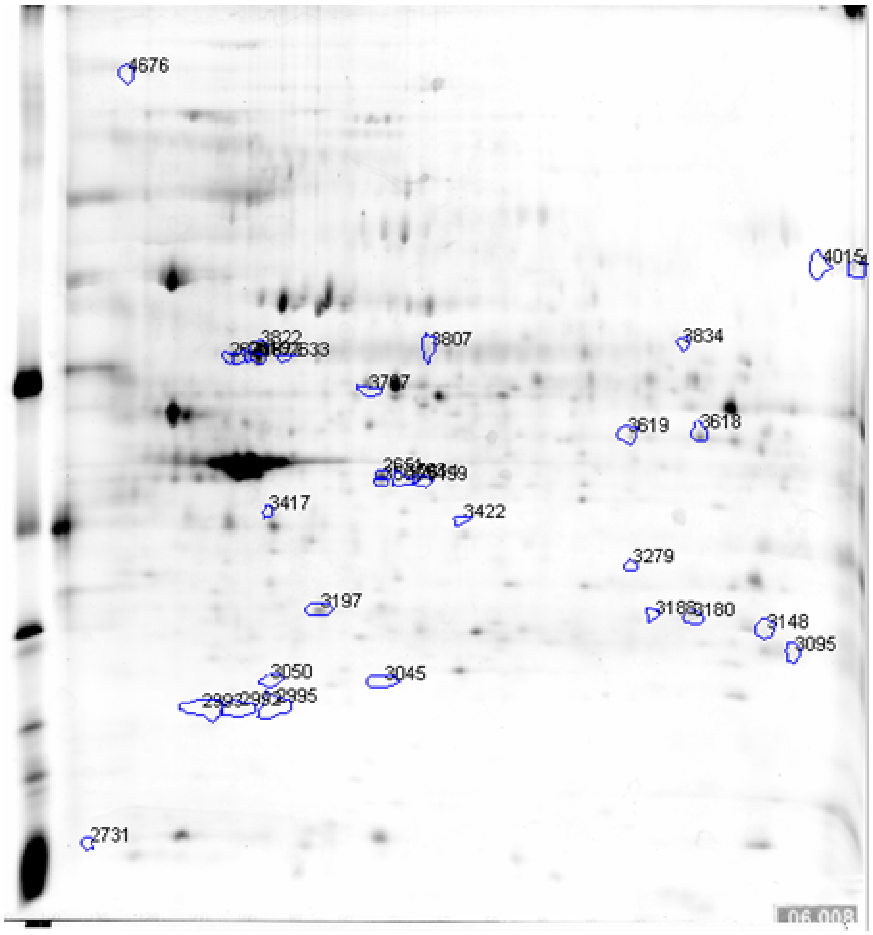


Figure 101: Visualisation de la localisation des taches protéiques retenues comme marqueur potentiel sur une des images de l'expérience de juin 2006. (Afin d'améliorer la visualisation, la dynamique de l'image a été étirée)

Pour chacune des taches protéiques retenues, l'outil proDIGE affiche une fiche récapitulative comprenant :

- son identifiant (défini par le logiciel ImageMaster lors de la détection des taches protéiques sur l'image de fusion),
- la nature de son expression (sur expression ou sous expression dans la classe Tumeur),
- la p-value calculée,
- ainsi que les informations éventuellement annotées par le biologiste lors de l'étape d'analyse d'image. (nom de la protéine, numéro d'accession à une base de données protéique, etc.)

Les figures suivantes montrent l'exemple des fiches de 2 taches protéiques retenues. La Figure 102 présente le marqueur potentiel identifié par le numéro 3180. Il s'agit d'une sous-expression dans la classe Tumeur. À l'inverse, la Figure 103 présente le marqueur potentiel 2992 sur-exprimé dans la classe tumeur. Les commentaires disponibles concernant ces deux protéines permettent de constater que cette étude différentielle sur le jeu de données de juin 2006 confirment les conclusions déjà apportées lors d'une précédente analyse, menée en 2005.

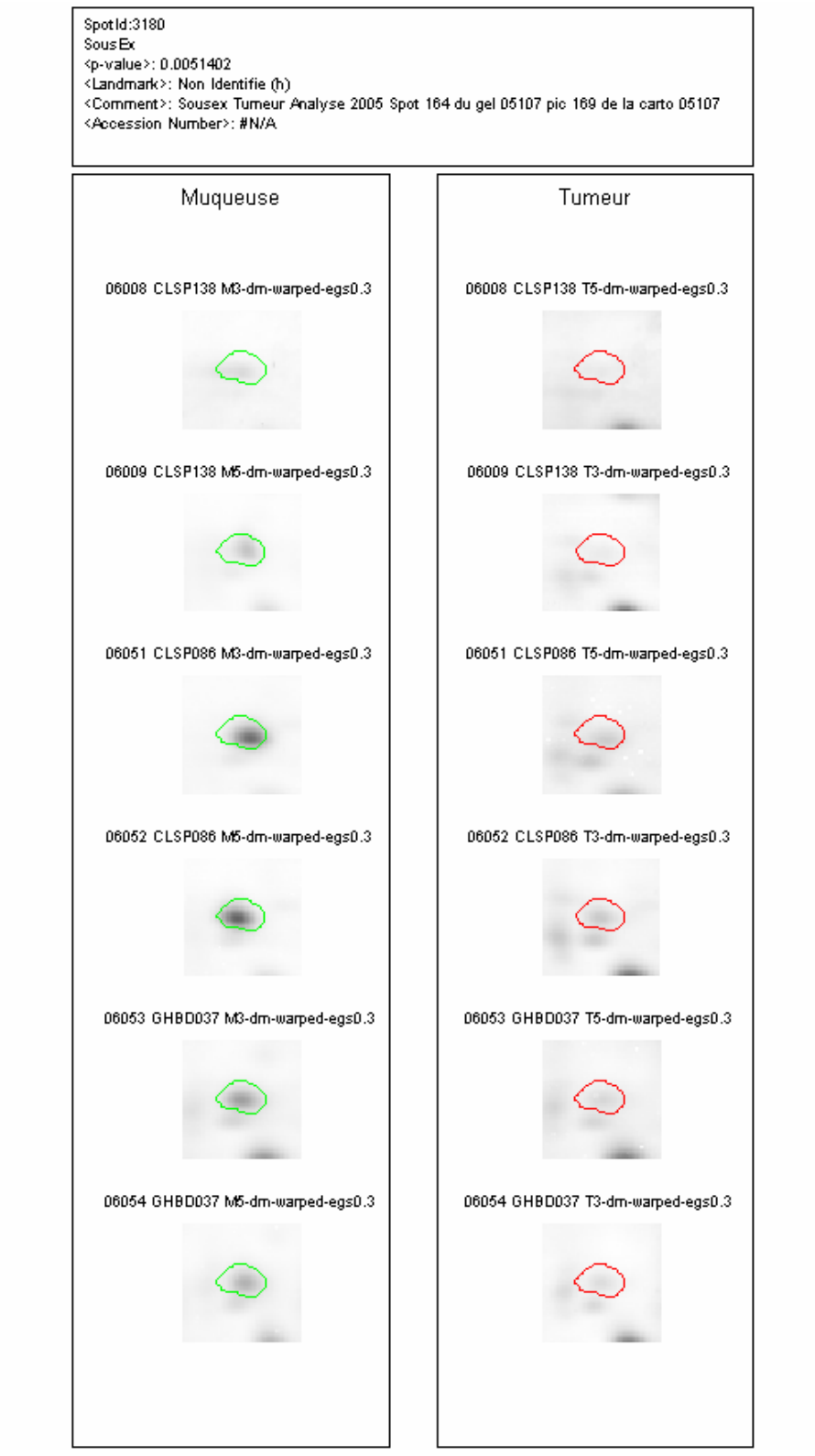


Figure 102: Fiche récapitulative du marqueur potentiel 3180. Cette protéine est sous-exprimée dans la classe Tumeur.

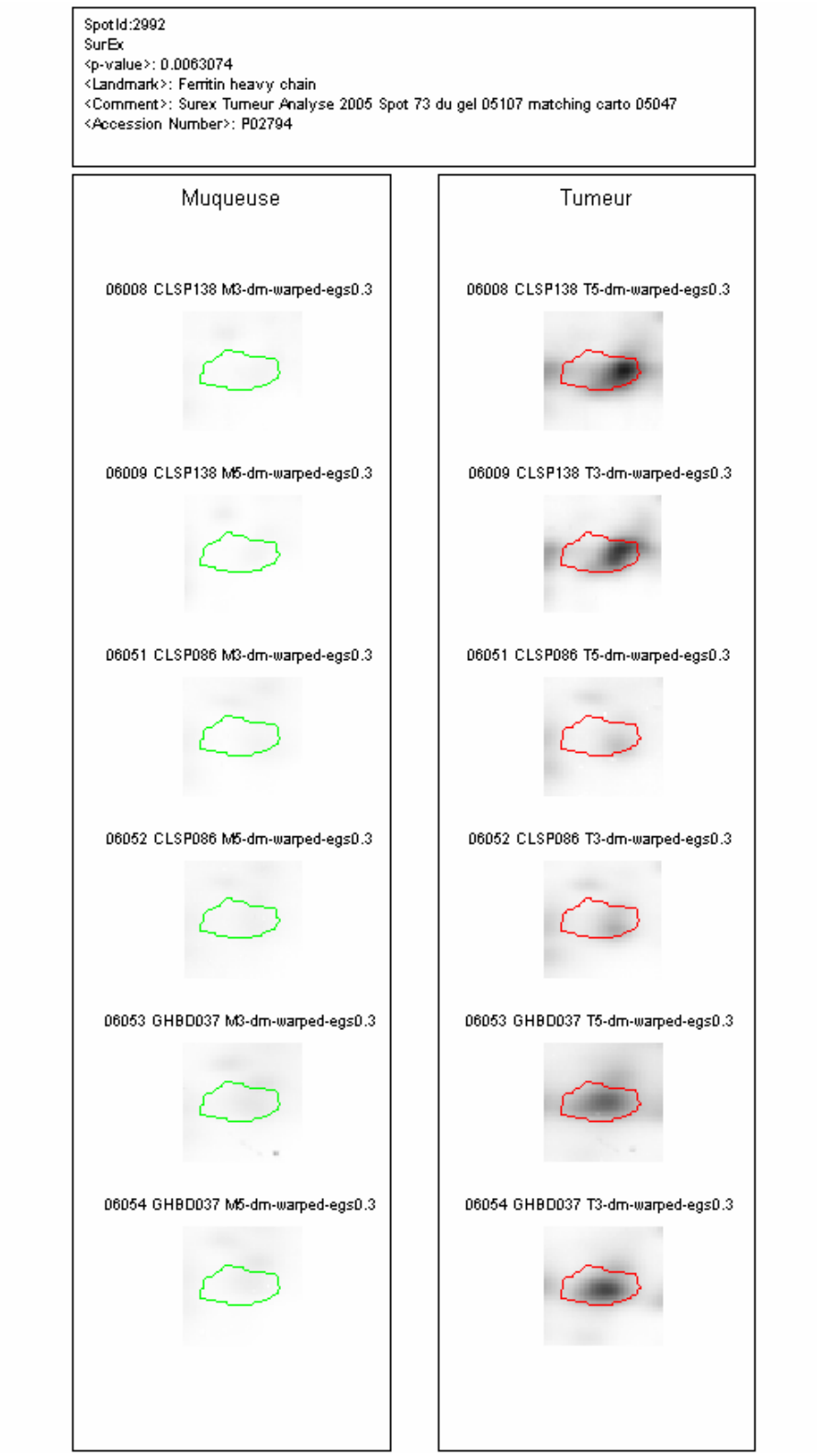


Figure 103: Fiche récapitulative du marqueur potentiel 2992. Cette protéine est sur-exprimée dans la classe tumeur.

Par ailleurs, la plupart des taches protéiques identifiées comme marqueurs protéiques par cette première méthode, le sera également par la méthode du t-test régulé présentée dans le paragraphe suivant. C'est notamment le cas pour la tache protéique 3180,

en sous-expression et la tache 2992, en sur-expression. Ces deux taches protéiques sont présentes dans le Tableau 10 et dans le Tableau 11 et apparaissent également comme résultats de l'approche par t-test régulé (se référer à la Figure 102 et à la Figure 103).

VI.8.4 Les données de sortie

La nature et la forme des données de sortie selon la méthode de fouille des données employée.

VI.8.4.1 Méthode de la courbe enveloppe et du score d'intérêt biologique :

L'application par l'outil ProDIGE de la méthode de la courbe enveloppe et de l'intérêt biologique associé conduit à l'ajout de deux nouveaux feuillets au fichier Excel principal. Le contenu de ces feuillets est résumé par le Tableau 14. Le premier, nommé « Exp diff pmin=0.02 pmax=0.08 », associe à chacune des taches protéiques et pour chacun des 6 gels de l'expérience, la valeur relative à sa classe (-1 pour une sous-expression, +1 pour une sur-expression, et 0 dans les autres cas) ainsi que le score correspondant à la somme de ces valeurs. De la même manière, le deuxième feuillet nommé « int.bio.pmin=0.02 pmax=0.08 », récapitule l'ensemble des valeurs de l'intérêt biologique et du score associé.

feuillet: "Exp diff pmin=0.02 pmax=0.08"							
Identifiant	CLSP138_gel1	CLSP138_gel2	CLSP086_gel1	CLSP086_gel2	GHBD037_gel1	GHBD037_gel2	Score
3804	-1	-1	0	-1	-1	-1	-5
2651	-1	-1	-1	-1	0	0	-4
3197	-1	-1	-1	-1	0	0	-4
3505	-1	-1	-1	-1	0	0	-4
3618	-1	-1	-1	0	-1	0	-4
4005	-1	0	0	-1	-1	-1	-4
3148	0	0	-1	0	-1	-1	-3
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
2622	0	1	1	1	0	0	3
2633	0	0	1	1	0	1	3
2995	0	1	0	0	1	1	3
3012	0	1	1	1	0	0	3
3830	1	1	0	0	1	1	4
2992	1	1	0	1	1	1	5
3789	1	1	1	1	1	0	5

feuillet: "int.bio.pmin=0.02 pmax=0.08"							
Identifiant	CLSP138_gel1	CLSP138_gel2	CLSP086_gel1	CLSP086_gel2	GHBD037_gel1	GHBD037_gel2	Score
3505	-0.93	-0.89	-0.99	-1.00	-0.89	-0.83	-5.52
2651	-0.99	-1.00	-0.99	-0.98	-0.65	-0.79	-5.39
3197	-0.89	-0.86	-0.99	-0.99	-0.84	-0.82	-5.39
3618	-0.92	-0.90	-0.95	-0.81	-0.95	-0.80	-5.33
3180	-0.81	-0.91	-0.92	-0.93	-0.85	-0.83	-5.25
4005	-0.93	-0.75	-0.74	-0.90	-0.94	-0.96	-5.22
3807	-0.90	-0.90	-0.44	-0.58	-0.66	-0.94	-4.42
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
3822	0.58	0.85	0.95	0.92	0.77	0.70	4.76
2995	0.83	0.96	0.60	0.59	0.92	0.95	4.85
2621	0.85	0.70	0.93	0.80	0.87	0.88	5.03
2622	0.83	0.95	0.99	0.94	0.70	0.84	5.25
2633	0.90	0.90	0.97	0.99	0.84	0.96	5.56
3789	0.98	0.97	0.97	0.90	0.94	0.79	5.56
2992	0.99	0.99	0.92	0.94	0.93	0.97	5.74

Tableau 14

VI.8.4.2 Méthode du t-test régulé

L'application par l'outil ProDIGE de la fouille des données par t-test régulé conduit à l'ajout d'un feuillet « t test régulé » au fichier Excel principal. Sur ce feuillet, dont un aperçu est donné sur le

Tableau 15, figurent toutes les informations relatives aux résultats de l'analyse :

- les identifiants,
- les volumes moyens,
- les ratios moyens,
- les variances régulées,
- les t-test associés,
- la nature de l'expression (+1 pour sur-expression, -1 pour sous-expression),
- les p-valeurs,
- les ratios
- ainsi que les différentes annotations (landmark, comment, Accession Number…).

Identifiant	Volume moyen	Ratio moyen	Var Régulée	t-test Régulé	Sign	p	P1_ratio	P2_ratio	P3_ratio	P4_ratio	P5_ratio	P6_ratio	Landmark	Comment	Accession Num
3180	2	-2	0.4	-3.3	-1	0.01	-1.7	-2.2	-2.2	-2.2	-1.8	-1.8	Non Identifie (h)	Sousex Tumeur	#N/A
2651	1.3	-2.9	0.9	-3	-1	0.01	-3.8	-3.9	-3.4	-3	-1.6	-1.8	Actin Cytoplasm	Sousex Tumeur	P60709
3618	2.8	-2	0.5	-2.9	-1	0.01	-2.2	-1.8	-2.4	-1.6	-2.3	-1.6	#N/A	#N/A	#N/A
3505	1.4	-2.7	0.9	-2.9	-1	0.01	-2.3	-2	-3.9	-4	-2.4	-1.8	#N/A	#N/A	#N/A
3417	-0.5	-1.9	0.5	-2.9	-1	0.01	-2.2	-2.2	-2	-2.3	-1.5	-1.3	#N/A	#N/A	#N/A
3197	2.8	-2.2	0.6	-2.7	-1	0.01	-2	-1.8	-3.1	-2.8	-1.7	-1.6	Actin cytoplasm	Sousex Tumeur	P60709
4005	1.9	-2.1	0.7	-2.6	-1	0.02	-2.1	-1.4	-1.4	-2	-2.4	-3.5	#N/A	#N/A	#N/A
3095	1.1	-1.6	0.5	-2.3	-1	0.02	-1.2	-1.3	-1.9	-1.1	-2	-2.2	#N/A	#N/A	#N/A
2614	-2.6	-2	0.8	-2.3	-1	0.03	-1.4	-1.8	-2	-3.1	-1.2	-2.5	Non Identifie (k)	Sousex Tumeur	#N/A
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
3834	1	1.5	0.6	2	1	0.04	2.1	2.2	1	1	1.2	1.1	#N/A	#N/A	#N/A
3045	1.5	1.2	0.4	2	1	0.04	0.9	1	0.9	1.2	1.6	1.5	#N/A	#N/A	#N/A
2993	2.9	1.3	0.4	2.1	1	0.03	1.5	1.3	1	1.2	1.5	1.5	#N/A	#N/A	#N/A
2995	2.2	1.6	0.6	2.2	1	0.03	1.5	2.4	0.9	1	1.9	2.2	#N/A	#N/A	#N/A
2622	5.3	1.8	0.6	2.3	1	0.02	1.4	2.1	2.7	2	1	1.4	#N/A	#N/A	#N/A
2621	2.9	1.6	0.4	2.4	1	0.02	1.6	1.1	2.1	1.5	1.6	1.6	Splicing factor 3	Surex Tumeur A	Q12874
3789	3.8	2.1	0.5	2.8	1	0.01	2.7	2.2	2.5	1.9	1.9	1.2	#N/A	#N/A	#N/A
2633	2.5	2.2	0.6	2.9	1	0.01	1.8	1.8	2.4	3.2	1.5	2.4	#N/A	#N/A	#N/A
2992	2.7	2.5	0.6	3.2	1	0.01	3	3.4	1.9	2	2.1	2.5	Ferritin heavy ch	Surex Tumeur A	P02794

Tableau 15: Aperçu du tableau résultat de l'analyse différentielle par t-test régulé.

Par ailleurs, un fichier texte (.txt) est créé qui permet l'importation, sous ImageMaster 2D platinium, des annotations désignant les marqueurs potentiels.

VI.9 Pertinence des résultats

Afin de constater la pertinence des résultats issus de l'application du workflow IDADIGE, il faut s'assurer, dans un premier temps, que les différences d'intensités existantes des taches protéiques de la classe muqueuse et celles de la classe muqueuse, sont bien mises en évidence. Dans un deuxième temps, il faut vérifier la pertinence biologique des taches protéiques retenues comme marqueurs potentiels.

VI.9.1 Pertinence statistique

En ce qui concerne la mise en évidence des différences d'intensité interclasse, la pertinence des résultats est dépendante de la signification apportée à la notion de « différence ». Les deux méthodes proposées (courbe enveloppe et test de Student régulé V.3.3.5) pour la recherche des marqueurs potentiels se distinguent sur l'interprétation de cette notion de différence. La première méthode proposée cherche à quantifier la variabilité expérimentale du ratio à partir de l'ensemble des données de l'expérience. Cette variabilité, bien que sur évaluée, permet d'attribuer une valeur de significativité pour chaque tache protéique de chaque gel. Cette information sur la significativité est pondérée, en fonction de l'intensité des taches, selon un profil défini par le biologiste. Pour cette première méthode, la notion de différence d'intensité concerne donc chaque gel indépendamment. La seconde méthode présentée s'appuie sur un test de Student régulé et appliqué pour chaque tache protéique, aux N valeurs du ratio évalué sur les N gels de l'expérience. Il en résulte que la différence d'intensité mise en évidence pour cette méthode, tient compte de la répétition sur l'ensemble des gels (et donc des patients) de l'expérience.

La différence entre ces deux approches de fouilles des données se traduit dans les résultats, par la constitution de listes de marqueurs potentiels distinctes. Pour s'en rendre compte, la liste des 30 taches protéiques les plus significatives au sens de l'intérêt biologique est rapprochée de celle des 30 taches les plus significatives au sens du t-test régulé, en sous-expression ainsi qu'en sur-expression. Le Tableau 16 présente le résultat de cette comparaison. Il ressort de cette étude que sur les 30 taches désignées selon le critère de l'intérêt biologique comme des sous-expressions, 21 le sont également par le t-test régulé. En sur-expression, le total de taches communes est de 25 sur 30. Pour comprendre les écarts entre ces listes, il faut revenir à la différence entre les deux approches.

La première méthode basée sur la l'étude de la variabilité expérimentale du ratio, peut conduire à la mise à l'écart de certaines taches pourtant visiblement variantes entre les deux images tumeur et muqueuse d'un même patient. C'est le cas pour les taches protéiques pour lesquelles le ratio est trop affecté par la variabilité expérimentale estimée pour le niveau d'intensité la concernant.

La seconde méthode ne permet pas non plus de retenir systématiquement les taches protéiques dont l'expression différentielle est bien visible sur deux images tumeur et muqueuse d'un même patient. En effet, pour certaines taches protéiques,

cette expression différentielle est trop variable d'un gel à l'autre et le ratio présente donc également une variabilité trop importante, faisant de la protéine associée un marqueur non fiable.

On s'aperçoit donc que le simple fait qu'une tache protéique présente une différence d'expression visiblement évidente, ne lui assure pas d'être retenue par l'une ou par l'autre des deux approches de fouille des données, et cela pour 2 raisons différentes :

- la variabilité intrinsèque du ratio dans la gamme d'intensité concernée et
- la variabilité du ratio au regard des répétitions sur les différents gels de l'expérience.

Classement des sous-expressions				Classement des sur-expressions		
Identifiant	Classement selon le score d'intérêt biologique	Classement selon le t-test régulé		Identifiant	Classement selon le score d'intérêt biologique	Classement selon le t-test régulé
3505	1	4		2992	1	1
2651	2	2		3789	2	3
3197	3	6		2633	3	2
3618	4	3		2622	4	5
3180	5	1		2621	5	4
4005	6	7		2995	6	6
3807	7	11		3822	7	11
4676	8	14		3811	8	19
3095	10	8		2993	9	7
3804	11	24		2619	10	13
3619	12	10		3707	11	12
3417	13	5		3834	12	9
3148	14	18		3045	13	8
2751	15	22		3050	14	14
2614	16	9		3267	15	26
3503	17	15		3487	18	30
3499	19	17		3486	19	17
2737	20	19		2982	20	16
3820	21	21		3327	21	18
2731	23	12		3422	22	10
4683	26	29		3406	23	23
3401	9	Absent		3361	24	20
4006	18	Absent		3496	26	27
4670	22	Absent		3055	28	22
4410	24	Absent		3268	30	25
2965	25	Absent		3814	16	Absent
4679	27	Absent		3462	17	Absent
3263	28	Absent		2618	25	Absent
4414	29	Absent		3064	27	Absent
3724	30	Absent		3685	29	Absent
3188	Absent	13		3279	Absent	15
4015	Absent	16		2985	Absent	21
3172	Absent	20		4100	Absent	24
4356	Absent	23		3282	Absent	28
3211	Absent	25		3073	Absent	29
3818	Absent	26				
3088	Absent	27				
3710	Absent	28				
3382	Absent	30				

Tableau 16

Par ailleurs, il est important de constater que les différences entre les deux listes concernent les taches les moins bien classées et que se sont les mêmes taches qui se retrouvent en tête des deux classements. Ceci s'explique par le fait que les taches de tête correspondent à des cas triviaux de différences d'expressions flagrantes, répétées sur tous les gels et qui se ressortent de la fouille quelle que soit la méthode employée. On pourra voir plus loin la confirmation de la surexpression biologique pour la protéine correspondant à l'identifiant 2633, surexprimée et très bien classée selon les 2 méthodes.

VI.9.2 Pertinence biologique

La pertinence des différences d'intensité retenues comme significatives est assurée par la nature même des méthodes employées (test de Student régulé et étude de la variance). Cependant, cette pertinence doit être également regardée dans le contexte biologique de la recherche de marqueurs tumoraux. Ce contexte nous a notamment amené à considérer davantage les différences d'expression situées dans les fortes intensités. En effet, les protéines fortement exprimées ont davantage de chances d'être relarguées par les cellules dans la circulation et donc de présenter un intérêt dans la conception d'un test de diagnostic. Pour ce faire, un profil de risque défini par le biologiste permet de contrôler le risque pris sur la désignation des variants significatifs en fonction de leur intensité. Ce contrôle du risque constitue un apport majeur pour la pertinence biologique des résultats. Par ailleurs, la pertinence biologique lors de la découverte de biomarqueurs est en générale également évaluée au regard de ses performances « cliniques » : sensibilité, spécificité, valeurs prédictives positive et négative. En pratique ce genre d'étude des performances est surtout utile pour évaluer la présence du marqueur dans l'échantillon d'intérêt pour le test, à savoir les fluides biologiques dans notre cas. En effet, les phases de recherche puis l'analyse d'image décrites dans cette thèse ont été faites à partir de tissu, mais le projet NODDICCAP vise à doser à la fin des biomarqueurs dans des fluides. La sensibilité et la spécificité des marqueurs pour une application clinique donnée seront évaluées lors des dosages Elisa des marqueurs dans des sérums. Pour le travail décrit ici, nous cherchons bien évidemment à identifier les marqueurs les plus significatifs possible, mais il n'est pas question de faire une étude de sensibilité/spécificité, car les tissus ne seront pas l'échantillon qui servira pour le test clinique. Dans la phase recherche du projet NODDICCAP, les marqueurs protéiques potentiels retenus à l'issu workflow DIGE sont considérés comme des pistes devant être explorées et confirmées par d'autres voies technologiques. Par ailleurs, une expérience DIGE ne comprend au maximum qu'une dizaine de patients et ne permettrait pas d'obtenir une puissance statistique suffisante pour estimer la sensibilité. La performance peut tout de même être améliorée en croisant les résultats issus de différentes expériences DIGE. Malheureusement, la mise en correspondance des taches protéiques entre des images provenant de différentes expériences est une opération souvent délicate du fait de leur qualité très inégale. Afin d'opérer ces croisements nous avons tout d'abord écarté les expériences les plus singulières de par leur qualité (déformation importante des images, faible nombre de taches protéiques, contamination protéiques, etc..). Ensuite, après la mise en correspondance d'un maximum de taches protéiques entre les différentes expériences considérées, nous nous sommes intéressés aux valeurs de l'intérêt biologique défini dans le V.3.3.5.2b). Comme le montre la Figure 104, l'observation de la valeur prise par cet indicateur, pour un maximum de taches protéiques et sur un ensemble important de patients, a permis d'isoler les marqueurs protéiques les plus sensibles et les plus fiables.

Le regroupement de différentes expériences nous permet de classer les protéines en fonction de leur expression préférentielle dans les tumeurs ou dans les muqueuses. Cela permet également de classer les tissus. Pour avoir une première idée

sur la spécificité des marqueurs surexprimés dans les tumeurs, nous avons également étudié quelques tissus issus de patients atteints de maladies inflammatoires digestives : cela nous a permis d'identifier les protéines exprimées dans ces tissus. Ces données ont pu être croisées avec les résultats issus de nos analyses différentielles. Ceci a permis d'écarter certaines protéines désignées comme marqueurs potentiels mais finalement non spécifiques du cancer colorectal.

Il est important de souligner que les deux méthodes statistiques proposées ne permettent pas une quantification de la significativité réelle des différences d'expression entre les classes. Elles permettent seulement d'établir un ordre de significativité entre tous les ratios d'expression observés. Ce constat a d'ailleurs été souligné par Kultina et al dans une étude [34] cherchant à optimiser la normalisation et l'évaluation des différences d'expression de données DIGE. Kultina et al concluent en effet cette étude en précisant que, du fait de la nature même des données DIGE, les p-valeurs obtenues par des tests de significativité ne doivent être utilisées que pour réaliser des classements d'intérêt.

Au cours de cette thèse, les deux méthodes ont été appliquées à des jeux de données réelles et donc les différences d'expression ne sont pas connues a priori. Il a donc été délicat de quantifier leur réelle efficacité. Il a été envisagé d'employer un jeu d'images de gels d'électrophorèse artificielles générées à partir d'un modèle et de quantifications protéiques connues [48]. Cependant, la validation du modèle est extrêmement difficile à réaliser et finalement nos méthodes statistiques ont trouvé leur légitimité au travers de leur exploitation. Elles ont permis la découverte de marqueurs en accord avec la théorie biologique liée à la pathologie. En effet, les croisements des résultats entre les différentes expériences de DIGE mais également avec les autres approches protéomiques engagées dans le projet NODDICCAP, ainsi qu'avec les données de la littérature sont très cohérents.

Une autre approche pour la fouille des données aurait pu être de chercher une signature permettant de classer les tissus, à l'image de ce qui est fait en génomique à partir des données de puces ADN. Cette approche multivariée par analyse discriminante aurait permis de prédire la modalité (pathologique ou sain) d'un patient à partir des intensités des taches protéiques (variables explicatives). Pour une telle approche il faut tenir compte des données manquantes. En effet, dans une technologie de type puce chaque emplacement a une identité connue et il y a peu de données manquantes à part celles dues aux problèmes expérimentaux. L'utilisation du patron commun lors d'une expérience DIGE permet de se situer dans ce cadre de jeu de données complet. Malheureusement, une simple expérience DIGE ne permet d'étudier qu'un faible nombre de patients et la compilation de jeux de données issus de différentes expériences est difficile. Cette compilation entraîne à nouveaux un grand nombre de données manquantes. Il s'agit alors soit d'en tenir compte soit de les écarter de l'analyse. La Figure 104 présente le croisement de plusieurs jeux de données DIGE, les données manquantes ayant été écartées (ce qui réduit grandement le nombre de protéines prises en compte). La figure fait apparaître les données ordonnées à l'aide d'une classification hiérarchique réalisée grâce au logiciel Spotfire (TIBCO Software Inc).

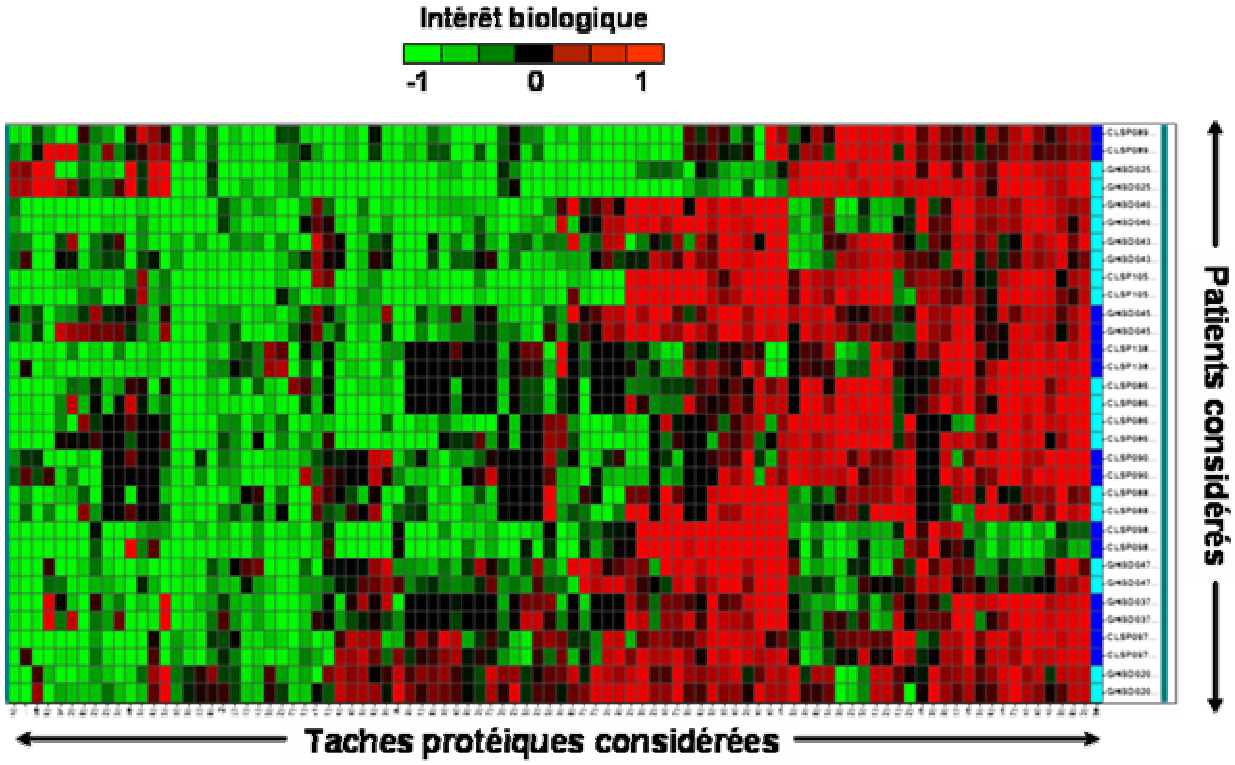


Figure 104: Croisement des données : représentation des valeurs de l'intérêt biologique estimées sur un ensemble de gels provenant de différences expérience DIGE. Seules les valeurs associées aux taches protéiques les plus intéressantes sont représentées. À gauche en vert , les colonnes correspondent à des marqueurs protéiques potentiels en sous-expression et à droite en rouge à des marqueurs en sur-expression (tumeur / muqueuse).

Théoriquement performante et innovante dans le domaine de l'analyse multivariée, cette approche n'est cependant pas adaptée aux données DIGE appliquées à la mise en place d'un test de dépistage du cancer colorectal. En effet, *in fine*, le test de dépistage ne pourra pas être fait à partir d'un échantillon de tissu. Il doit pouvoir s'effectuer à partir d'un fluide biologique (sérum ou des selles). Or toutes les protéines contenues dans le tissu ne se retrouvent pas forcément dans les fluides biologiques, ce qui implique que toute signature multivariée déterminée à partir du tissu est inexploitable.

La meilleure validation des marqueurs potentiels mis en évidence à l'issue de l'application du workflow IDADIGE et du croisement inter expériences, sera de ce fait le test de la quantité de ces protéines dans un fluide biologique. Cela sera seulement possible lorsque les anticorps seront disponibles pour doser les différentes protéines dans les fluides.

VI.9.3 Validation biologique à travers l'exemple d'un marqueur potentiel

L'étape de validation a pu être réalisée pour certains marqueurs potentiels. C'est le cas pour la tache protéique dont l'identifiant est le 2633, classée parmi les plus intéressantes selon les deux méthodes de fouille des données appliquées dans le

cadre de l'étude DIGE de janvier 2006. La Figure 105 présente une visualisation de cette tache pour deux patients de l'expérience, tandis que le Tableau 17 récapitule les données qui lui sont relatives : les ratios normalisés ainsi que les différents paramètres évalués lors de l'application des deux approches de fouille des données. Quelle que soit l'approche employée, la tache protéique ressort en sur-expression significative avec notamment des valeurs d'intérêt biologique proche de 1 pour chacun des 6 gels et une p-valeur, p=0.01, associée à l'observation du t-test régulé, inférieure au seuil fixé par le biologiste. Ces informations justifient donc une validation biologique pour la protéine associée à cette tache protéique.

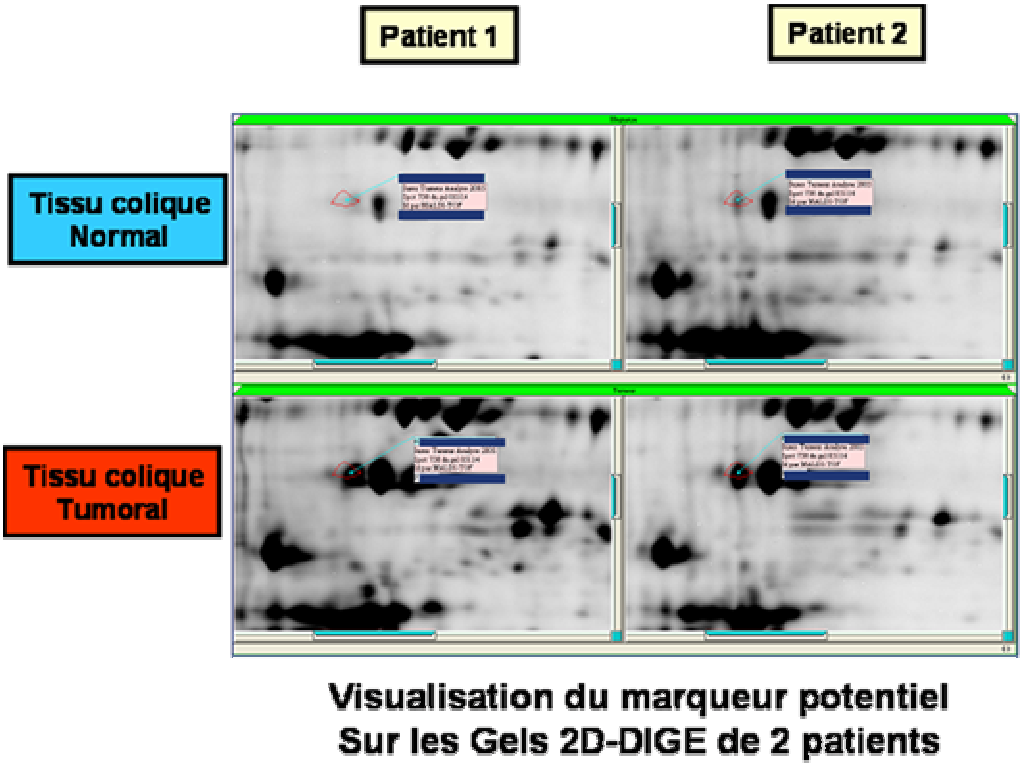


Figure 105 : Visualisation de la tache 3633 , en sur-expression sur le tissu tumoral par rapport au tissu normal, pour deux patients des 6 de l'expérience.

	CLSP138_gel1	CLSP138_gel2	CLSP086_gel1	CLSP086_gel2	GHBD037_gel1	GHBD037_gel2	Score	Classement
Ratio volumique (log2)	1,79	1,76	2,43	3,22	1,48	2,35		
test "courbe enveloppe"	0	0	1	1	0	1	3	4
Intérêt biologique	0,90	0,90	0,97	0,99	0,84	0,96	5,56	3
	Volume moyen	Ratio moyen	Variance Régulée	t-test Régulé	p-valeur			Classement
t-test régulé	2,54	2,17	0,57	2,87	0,01			2

Tableau 17 : Résumé des données relatives à la tache protéique 2633 : les ratios volumiques normalisés (log2) ainsi que les différents paramètres et statistiques calculés lors de l'application des 2 approches de fouille des données du workflow IDADIGE.

La Figure 105 et la Figure 107 illustrent une première étape de validation biologique. Cette étape est réalisée par immunohistochimie qui est un processus de détection d'antigènes dans les tissus au moyen d'anticorps. Lors de l'expérience d'immunohistochimie, la protéine 2633 est retrouvée sur-exprimée de manière systématique sur une vingtaine de patients. Nous confirmons donc bien que la protéine est surexprimée dans les tissus, et qu'il ne s'agit pas d'une observation seulement vue par la technologie DIGE, ni due uniquement à la variabilité expérimentale. L'expression intense observée en immunohistochimie fait espérer un relargage de la protéine dans les fluides biologiques.

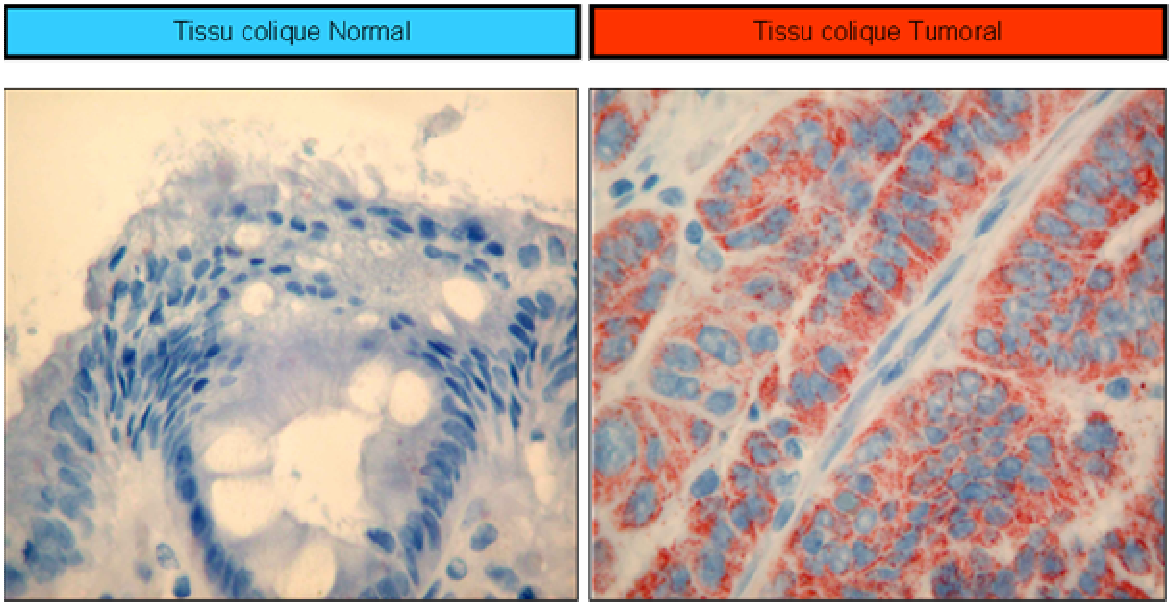


Figure 106 : Visualisation du marquage en immunohistochimie de la protéine 2633 : le marquage (coloration rouge) dénote la présence de la protéine dans la tumeur. À l'inverse la muqueuse n'est pas marquée, ce qui confirme l'intérêt biologique de la protéine.

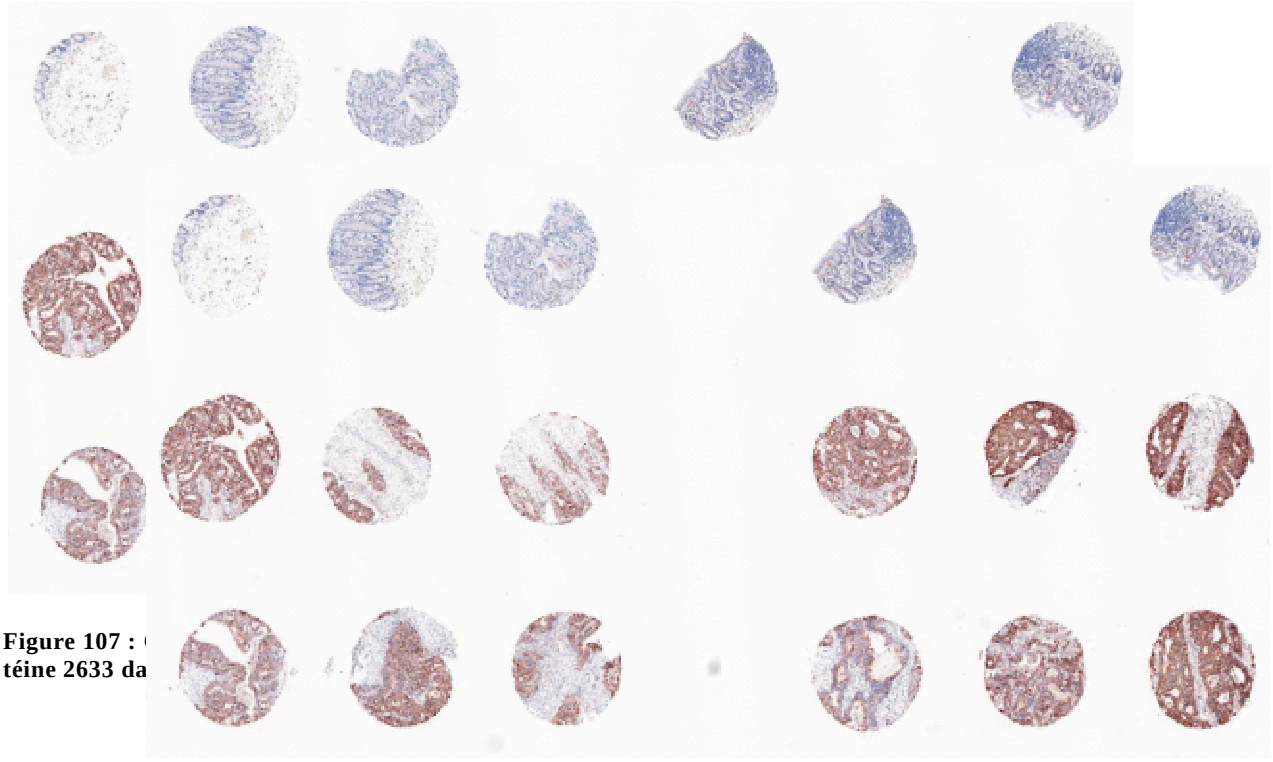


Figure 107 : (
téine 2633 da

Une deuxième étape concerne la validation dans un fluide biologique. Une première évaluation a été réalisée à partir de 94 patients atteints de cancer colorectal et de 127 sujets « sains », à l'aide d'un dosage Elisa dans le sérum.

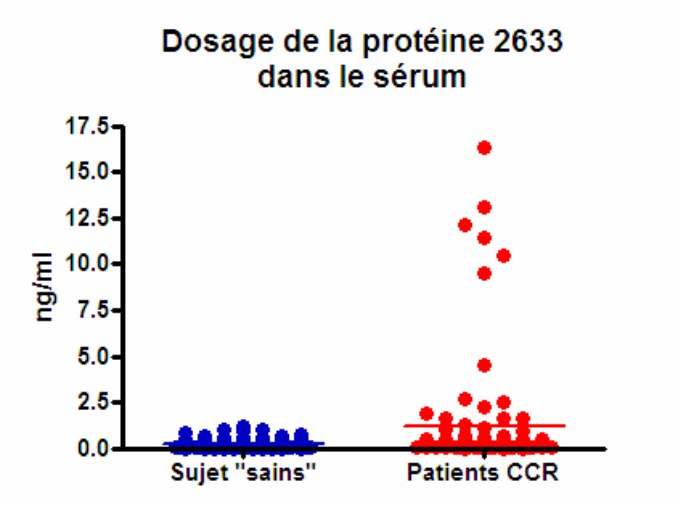


Figure 108 : Visualisation de l'ensemble des dosages Elisa selon que le sujet soit sain ou atteint de cancer colorectal.

L'ensemble des dosages effectués chez des sujets sains et pathologiques a permis d'établir la Figure 108. Cette figure suggère l'existence d'une sur-expression de la protéine dans le sérum des patients atteints de cancer colorectal. La p-valeur p=0.01 associée à l'observation du t-test lors de la comparaison des moyennes des deux classes confirme l'impression laissée par la figure puisqu'elle permet d'affirmer que les moyennes sont significativement ($P<0.05$) différentes.

VII. CONCLUSION

Le workflow IDADIGE a été mis en place de façon à répondre à des attentes bien précises liées à un contexte particulier. Nous avons notamment dû tenir compte de la plateforme protéique de bioMérieux qui comprend un parc de logiciels et d'outils technologiques particuliers. Nous avons également dû tenir compte du cadre de recherche du projet NODDICCAP qui nous impose l'analyse différentielle pour la découverte de marqueurs potentiels du cancer colorectal. Il a donc fallu composer avec les outils logiciels et technologiques disponibles, en identifiant tout d'abord les différentes étapes obligées depuis le prétraitement des images brutes jusqu'à l'interprétations des résultats de l'analyse différentielle. Parmi ces étapes, il a également fallu repérer celles qui représentaient les plus grandes limitations dans la qualité du traitement et de l'interprétation des données. La qualité des images DIGE, très peu bruitées, ajoutées aux algorithmes performants de traitement d'images des logiciels dédiés tels que ImageMaster 2D limitent fortement l'intérêt de porter l'effort de développement sur les problématiques de filtrage des bruits, de segmentation, et alignement des images. La problématique de la mise en correspondance des taches protéiques entre les images de gels différents représentait un des derniers goulots d'étranglement de l'interprétation des images DIGE jusqu'à l'arrivée récente des stratégies de détection des taches par patron commun.

Le travail a donc en cours de thèse davantage porté sur des aspects stratégiques comme le choix du schéma expérimental, car les autres problèmes avaient été spontanément résolus par l'apparition de solutions commerciales. En effet, aujourd'hui, les logiciels dédiés à l'exploitation de la DIGE proposent souvent un panel de solutions pour chaque étape d'un workflow que l'on peut qualifier de classique. Ce workflow classique se caractérise notamment par la présence d'un standard interne parmi les trois populations protéiques chargées sur chaque gel d'une expérience. Ce standard interne permet une première normalisation des quantifications volumiques des taches protéiques. Cependant, dans notre cas d'analyse différentielle, nous travaillons sur les données appariées puisque chaque patient fourni un échantillon sain et un échantillon tumoral. Cette particularité rend superflue l'utilisation du standard, car une normalisation équivalente est assurée par la considération du ratio des données appariées. Le troisième fluorophore a donc pu être utilisé pour le marquage d'une lignée commerciale de Caco2. L'utilisation de cette population toujours identique, a permis d'améliorer le rapprochement et l'alignement d'images de gels issus d'expériences différentes (une telle approche n'a jamais été utilisée à notre connaissance).

Par ailleurs, au moment du développement d'IDADIGE, l'approche consistant à utiliser un seul et même patron de détection était innovante. Aujourd'hui, l'évolution de la plupart des logiciels dédiés a abouti à l'emploi de méthodes similaires. Cependant, la méthode de fusion des images développée dans IDADIGE se démarque de celles proposées par ces logiciels. Elle permet une juste représentation d'un maximum de taches protéiques sur l'image de fusion grâce notamment à une égalisation intelligente des niveaux de gris.

Enfin, le workflow IDADIGE se distingue par l'effort d'innovation qui a été fait au niveau du traitement et de l'exploitation des quantifications volumique des taches. Cet effort se traduit par le transfert et l'adaptation de méthodes de normalisation et de fouilles des données depuis le domaine voisin des puces à ADN. Nous avons également mis en place une méthode sur mesure qui a consisté à établir la notion d'intérêt biologique, tenant compte à la fois de la significativité statistique et du contexte biologique (défini par l'expert) et mesuré par un paramètre ad hoc.

Il aurait été possible d'adapter notre schéma expérimental et notre contexte à ce workflow classique, mais c'est donc l'inverse qui a été choisi. L'exigence de l'optimisation du workflow démontre donc la nécessité du développement du workflow IDADIGE.

VIII. Bibliographie

[1] Alban A, David SO, Bjorkesten L, Andersson C, Sloge E, Lewis S, Currie I, "A novel experimental design for comparative two-dimensional gel analysis: Two-dimensional difference gel electrophoresis incorporating a pooled internal standard", Proteomics, Vol. 3, 1, pp. 36-44, (2003)

[2] Almeida SA, Stanislaus R, Krug E, Arthur JM, "Normalization and analysis of residual variation in two-dimensional gel electrophoresis for quantitative differential proteomics", Proteomics, Vol. 5, pp. 1242-1249, (2005)

[3] Amersham Biosciences, "Fluorescence Imaging: principles and methods", Handbook, Ref. 63-0035-28, (2000)

[4] Anderson NL, Anderson NG, "The Human Plasma Proteome", Molecular & Cellular Proteomics, Vol. 1, 11, pp. 845-867, (2002)

[5] Appel RD, Vargas JR, Palagi P M, "Melanie II - a third-generation software package for analysis of two-dimensional electrophoresis images: II. Algorithms", Electrophoresis, Vol. 18, pp. 2735-2748, (1997)

[6] Appel RD, Hochstrasser DF, Funk M, Vargas JR, Pellegrini C, Muller AF, Scherrer J-R, "The MELANIE project: From a biopsy to automatic protein map interpretation by computer", Electrophoresis, Vol. 12, pp. 722-735, (1991)

[7] Baldi P, Long A D, "A Bayesian framework for the analysis of microarray expression data: regularized t -test and statistical inferences of gene changes", Bioinformatics, Vol. 17, 6, pp. 509-519, (2001)

[8] Belle WV, Sjøholt G, Ånensen N, Høgda K-A, Gjertsen BT, "Adaptive contrast enhancement of two-dimensional electrophoretic protein gel images facilitates visualization, orientation and alignment", Electrophoresis, Vol. 20, pp. 4086-4095, (2006)

[9] Bettens E, Scheunders P, Sijbers J, Van Dyck D and Moens L, "Automatic segmentation and modelling of two-dimensional electrophoresis-gels", Proceedings ICIP'96, Vol. 2, IEEE International Conference on Image Processing, pp. 665-668, (1996)

[10] Beucher S, Lantuejoul C, "Use of watersheds in contour detection", In International Workshop on Image Processing, Real-Time Edge and Motion Detection/Estimation, Rennes, France, (1979)

[11] Bolstad BM, Irizarry RA, Astrand M, Speed TP, "A comparison of normalization methods for high density oligonucleotide array data based on variance and bias", Bioinformatics, Vol. 19, 2, pp.185-193, (2003)

[12] Candes EJ, Donoho DL, "Ridgelets: a key to higher-dimensional intermittency Ridgelets: a key to higher-dimensional intermittency?", Phil. Trans. R. Soc. Lond. A., Vol. 357,1760, pp. 2495-2509, (1999)

[13] Chui H, Anand R, "A new algorithm for non-rigid point matching", IEEE Conference on Computer Vision and Pattern Recognition, Vol. 2, pp. 44-51, (2000)

[14] Conradsen K, Pedersen J, "Analysis of two-dimensional electrophoresis gels", Biometrics, Vol. 48, pp. 1273-1287, (1992)

[15] Cui X, Churchill GA, "Statistical tests for differential expression in cDNA microarray experiments", Genome Biology, Vol. 4, 4, Article 210, (2003)

[16] Dobbin K, Shih JH, Simon R, "Statistical design of reverse dye microarrays", Bioinformatics, Vol. 19, 7, pp. 803-810, (2003)

[17] Dowsey AW, Dunn MJ, Yang GZ, "The role of bioinformatics in two-dimensional gel electrophoresis", Proteomics, Vol. 3, pp. 1567-1596, (2003)

[18] Efrat A, Hoffmann F, Kriegel K, Schultz C, Wenk C, "Geometric algorithms for the analysis of 2D-electrophoresis gels", Journal of Computational Biology, Vol. 9, pp. 299-315, (2002)

[19] Fodor IK, Nelson DO, Alegria-Hartman M, Robbins K, Langlois RG, Turteltaub KW, Corzett TH, Mccutchen-Maloney SL, "Statistical challenges in the analysis of two-dimensional difference gel electrophoresis experiments using DeCyder", Bioinformatics, Vol. 21, 19, pp. 3733-3740, (2005)

[20] Fournier D, "Analyse de bactériophages", Travail de diplôme, EIG (Ecole d'Ingénieurs de Genève), HES, http://eig.unige.ch/~kocher/RAPPORT_DE_DIPLOME_2002.PDF, (2002)

[21] Glasbey CA et Mardia KV, "A review of image warping methods", Journal of Applied Statistics, Vol. 25, pp. 155-171, (1998)

[22] Gonzalez RC, Wood RE, "Digital Image Processing, 2nd Edition", Prentice Hall, (2002)

[23] Gold S, Anand R et al., "New Algorithms for 2D and 3D Point matching: Pose estimation and correspondence", Pattern Recognition, Vol. 31, 8, pp. 1019-1031, (1998)

[24] Guo XH, Chen SH, "The structure and thermodynamics of protein-SDS complexes in solution and the mechanism of their transport in gel electrophoresis process", Chemical Physics, Vol. 149, pp. 129-139, (1990)

[25] Gustafsson JS, Blomberg A, Rudemo M, "Warping two-dimensional electrophoresis gel images to correct for geometric distortions of the spot pattern", Electrophoresis, Vol. 23, pp. 1731-1744, (2002)

[26] Hatfield GW, Hung S, Baldi P, "Differential analysis of DNA microarray gene expression data", Molecular Microbiology, Vol. 47, 4, pp. 871-877, (2003)

[27] Horgan GW, Glasbey CA, "Uses of digital image analysis in electrophoresis", Electrophoresis, Vol. 16, pp. 298-305, (1995)

[28] Huber W, von Heydebreck A, Sultmann H, Poustka A, Vingron M, "Variance stabilization applied to microarray data calibration and to the quantification of differential expression", Bioinformatics, Vol. 18, pp. S96-S104, (2002)

[29] Jungblut P, Thiede B, Zimny-Arndt U, Muller EC et al., "Resolution power of two-dimensional electrophoresis and identification of proteins from gels", Electrophoresis, Vol. 17, pp. 839-847, (1996)

[30] Kaczmarek K, Walczak B, Jong S, Vandeginste BGM, "Baseline reduction in two-dimensional gel electrophoresis images", Acta chromatographica, Vol. 15, pp. 82-96, (2005)

[31] Kaczmarek K, Walczak B, Jong S, Vandeginste BGM, "Preprocessing of two-dimensional gel electrophoresis images", Proteomics, Vol. 4, pp. 2377-2389, (2004)

[32] Karp NA, Kreil DP, Lilley KS, "Determining a significant change in protein expression with DeCyder during a pair-wise comparison using two-dimensional difference gel electrophoresis", Proteomics, Vol. 4, pp. 1421-1432, (2004)

[33] Kreil DP, Karp NA, Lilley KS, "DNA microarray normalization methods can remove bias from differential protein expression analysis of 2-D difference gel electrophoresis results", Bioinformatics, Vol. 20, pp. 2026-2034, (2004)

[34] Kultima K, Scholz B, Alm H, Sköld K, Svensson M, Crossman AR, Bezard E, Andrén PE, Lönnstedt I, "Normalization and expression changes in predefined sets of proteins using 2D gel electrophoresis: A proteomic study of L-DOPA induced dyskinesia in an animal model of Parkinson's disease using DIGE", BMC Bioinformatics, 7:475, (2006)

[35] Lim, Jae S., "Two-Dimensional Signal and Image Processing", Englewood Cliffs, NJ: Prentice Hall, pp. 536-540, (1990)

[36] Locke BR, Trinh SH, "When can the Ogston-Morris-Rodbard-Chrambach model be applied to gel electrophoresis?", Electrophoresis, Vol. 20, 17, pp. 3331-3334, (1999)

[37] Luhn S, Berth M, Hecker M, Bernhardt J, "Using standard positions and image fusion to create proteome maps from collections of two-dimensional gel electrophoresis images", Proteomics, Vol. 3, pp. 1117-1127, (2003)

[38] Marquardt D, "An Algorithm for Least-Squares Estimation of Nonlinear Parameters". SIAM Journal on Applied Mathematics, Vol. 11, pp. 431–441, (1963)

[39] Miller MJ et al., "Strategies and techniques for testing the precision, reliability and reproducibility of computerized two-dimensional gel electrophoresis analysis systems", Applied and Theoretical Electrophoresis, Vol. 1, 3, pp. 127-135, (1989)

[40] Nadaraya EA, "On nonparametric estimates of density functions and regression curves", Theory Prob. Appl., Vol. 10, pp. 186-190, (1965)

[41] Nishihara JC, Champion KM, "Quantitative evaluation of proteins in one- and two-dimensional polyacrylamide gels using a fluorescent stain", Electrophoresis, Vol. 23, 14, pp. 2203-2215, (2002)

[42] Pietrogrande MC, Marchetii N, Dondi F, Righetti PG, "Spot overlapping in two-dimensional polyacrylamide gel electrophoresis separations: A statistical study of complex protein maps", Electrophoresis, Vol. 23, pp. 283-91, (2002)

[43] Pleißner KP, Hoffmann F, Kriegel K, Wenk C, Wegner S, Sahlström A, Oswald H, Alt H, Fleck E, "New algorithmic approaches to protein spot detection and pattern matching in two-dimensional electrophoresis gel databases", Electrophoresis, Vol. 20, 4-5, pp. 755-765, (1999)

[44] Prehm J, Jungblut P, Klose J, "Analysis of two-dimensional electrophoretic protein patterns using a video camera and a computer. II. Adaptation of automatic spot detection to visual evaluation", Electrophoresis, Vol. 8, pp. 562-572, (1987)

[45] Rabilloud T, "Two-dimensional gel electrophoresis in proteomics: Old, old fashioned, but it still climbs up the mountains", Proteomics, Vol. 2, pp. 3-10, (2002)

[46] Raman B, Cheung A, Marten MR, "Quantitative comparison and evaluation of two commercially available, two-dimensional electrophoresis image analysis software packages, Z3 and Melanie", Electrophoresis, Vol. 23, 14, pp. 2194-2202, (2002)

[47] Rodbard D, Chrambach A, "Unified theory of gel electrophoresis and gel filtration", Proceedings of the National Academy of Sciences, Vol. 65, 4, pp. 970-977, (1970)

[48] Rogers M, Graham J, Tonge RP, "Using statistical image models for objective evaluation of spot detection in two-dimensional gels", Proteomics, Vol. 3, 6, pp. 879-896, (2003)

[49] Rosengren AT, Salmi JM, Aittokallio T et al., "Comparison of PDQuest and Progenesis software packages in the analysis of two-dimensional electrophoresis gels", Proteomics, Vol. 3, pp. 1936-1946, (2003)

[50] Rousseeuw PJ, Leroy AM, "Robust Regression & Outlier Detection", Wiley & Sons, New-York, 329 p., (1987)

[51] Seillier-Moiseiwitsch F, Trost DC et Moiseiwitsch J, "Statistical methods for proteomics", Biostatistical Methods in Molecular Biology, Looney S (éditeur), Vol. 184, pp. 51-80, (2002)

[52] Serra J, "Image Analysis and mathematical Morphology" Ac. Press, Tome 1 (1982), Tome 2 (1988)

[53] Skolnick MM, Sternberg SR, Neel JV., "Computer programs for adapting two-dimensional gels to the study of mutation", Clinical Chemistry, Vol. 28, 4, 969-978, (1982)

[54] Takahashi K, Nakazawa M, Watanabe Y, Konagaya A, "Fully-Automated Spot Recognition and Matching Algorithms for 2-D Gel Electrophoretogram of Genomic DNA", Genome Informatics, Vol. 9, pp. 161-172, (1998)

[55] Tyson JJ, Haralick RH, "Computer analysis of two-dimensional gels by a general image processing system", Electrophoresis, Vol. 7, pp.107-113, (1986)

[56] Ünlü M, Morgan ME, Minden JS, "Difference gel electrophoresis: a single gel method for detecting changes in cell extracts", Electrophoresis, Vol. 18, pp. 2071-2077, (1997)

[57] Vincent L, Soille P, "Watersheds in digital spaces: An efficient algorithm based on immersion simulations", IEEE Tran. on Pattern Analysis and Machine Intelligence, Vol. 13, pp. 583-598, (1991)

[58] Wheelock AM, Buckpitt AR, "Software-induced variance in two-dimensional gel electrophoresis image analysis", Electrophoresis, Vol. 26, pp. 4508-4520, (2005)

[59] Wilkins MR, Williams KL, Appel RD , Hochstrasser DF, "Proteome Research: new frontiers in functional genomics", Springer-Verlag, Heidelberg, pp. 1-12, (1997)

[60] Wilkins MR, Williams KL, Appel RD , Hochstrasser DF, "Proteome Research: new frontiers in functional genomics", Springer-Verlag, Heidelberg, pp. 187-220, (1997)

[61] Yan JX, Devenish AT, Wait R, Stone T, Lewis S, Fowler S, "Fluorescence 2-D difference gel electrophoresis and mass spectrometry based proteomic analysis of E. coli.", Proteomics, Vol. 2, pp. 1682-1698, (2002)

[62] Yang YH, Dudoit S, Luu P, Lin DM, Peng V, Ngai J, Speed TP, "Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation", Nucleic Acids, Vol. 30, 4, e15, (2002)

IX. ANNEXES

A. Tableaux des indicateurs de l'efficacité d'appariement des taches protéiques suivant la méthode employée

Approche classique sans alignement				
Image	NbSpots	NbAppariements	%Appariement	%Erreurs
06008 M csw	1115	630	52%	13%
06008 T csw	1302	709	54%	15%
06009 T csw	1304	877	67%	9%
06009 M csw	1325	1325	100%	0%
06051 M csw	1055	573	48%	23%
06051 T csw	1047	552	47%	22%
06052 M csw	1041	581	49%	15%
06052 T csw	1347	663	50%	20%
06053 M csw	1141	670	54%	22%
06053 T csw	1230	617	48%	44%
06054 M csw	1074	637	53%	24%
06054 T csw	1296	636	49%	38%

Tableau 18 : Indicateurs de qualité d'appariement pour l'approche « classique sans alignement ».

Approche classique avec alignement				
Image	NbSpots	NbAppariements	%Appariement	%Erreurs
06008 M caw	1110	716	59%	12%
06008 T caw	1277	767	59%	16%
06009 T caw	1316	899	68%	10%
06009 M caw	1314	1314	100%	0%
06051 M caw	1086	609	51%	12%
06051 T caw	1047	579	49%	17%
06052 M caw	1065	583	49%	26%
06052 T caw	1312	711	54%	16%
06053 M caw	1127	671	55%	17%
06053 T caw	1259	688	53%	16%
06054 M caw	1112	697	57%	19%
06054 T caw	1306	689	53%	16%

Tableau 19 : Indicateurs de qualité d'appariement pour l'approche « classique avec alignement ».

Approche pattern commun intra-gel				
Image	NbSpots	NbAppariements	%Appariement	%Erreurs
06008 M pc1	1423	956	66%	6%
06008 T pc1	1423	927	64%	6%
06009 T pc1	1471	1471	100%	0%
06009 M pc1	1471	1471	100%	0%
06051 M pc1	1169	728	55%	17%
06051 T pc1	1169	753	57%	16%
06052 M pc1	1339	811	58%	20%
06052 T pc1	1339	835	59%	19%
06053 M pc1	1421	839	58%	18%
06053 T pc1	1421	808	56%	16%
06054 M pc1	1482	901	61%	11%
06054 T pc1	1482	893	60%	11%

Tableau 20 : Indicateurs de qualité d'appariement pour l'approche « patron commun intra-gel ».

Approche pattern commun inter-gel				
Image	NbSpots	NbAppariements	%Appariement	%Erreurs
06008 M pc2	1390	1389	100%	0%
06008 T pc2	1390	1387	100%	0%
06009 M pc2	1390	1390	100%	0%
06009 T pc2	1390	1389	100%	0%
06051 M pc2	1390	1383	99%	0%
06051 T pc2	1390	1389	100%	0%
06052 M pc2	1390	1388	100%	0%
06052 T pc2	1390	1389	100%	0%
06053 M pc2	1390	1389	100%	0%
06053 T pc2	1390	1389	100%	0%
06054 M pc2	1390	1388	100%	0%
06054 T pc2	1390	1387	100%	0%

Tableau 21 : Indicateurs de qualité d'appariement pour l'approche « patron commun inter-gels ».

Approche ProGenesis (SameSpots)				
Image	NbSpots	NbAppariements	%Appariement	%Erreurs
06008 M pc2	1640	1640	100%	0%
06008 T pc2	1640	1640	100%	0%
06009 M pc2	1640	1640	100%	0%
06009 T pc2	1640	1640	100%	0%
06051 M pc2	1640	1640	100%	0%
06051 T pc2	1640	1640	100%	0%
06052 M pc2	1640	1640	100%	0%
06052 T pc2	1640	1640	100%	0%
06053 M pc2	1640	1640	100%	0%
06053 T pc2	1640	1640	100%	0%
06054 M pc2	1640	1640	100%	0%
06054 T pc2	1640	1640	100%	0%

Tableau 22 : Indicateurs de qualité d'appariement pour l'approche logiciel Progenesis (SameSpots).

B. Procédure de mise en œuvre de ProDIGE

Cette procédure a été rédigée à l'attention du personnel du laboratoire de protéomique de bioMérieux. Elle montre, par l'exemple, l'utilisation courante de l'outil ProDIGE (sous Matlab) dans le contexte du workflow IDADIGE.

Traitement des images

- Créer son répertoire de travail
- Lancer Matlab
- Lancer ProDIGE (écrire "prodige" dans la fenêtre de commande et appuyer sur entrée)
- Sélectionner son répertoire de travail
- Sélectionner « Prétraitement d'images »
- Sélectionner les images brutes à traiter

  Remarques:

  - Un répertoire "Prétraitement" vient d'être créé dans le répertoire de travail, qui contient une copie des images brutes

  - Les images venant d'être sélectionnées font partie de la sélection sujette aux traitements. À tout moment il est possible de redéfinir cette sélection (option « 1 »).

  - La sélection de travail est affichée dans la fenêtre de commande
- Afficher la sélection de travail: option « 2 ». Préciser ensuite le nombre d'images par figure désiré (4 par exemple).

  Remarques:

  - Suite aux aléas de manipulation, les images peuvent être pivotées ou bien en négatif. Ce n'est pas le cas ici, mais il est utile de tester les options de mise en forme qui ne sont pas destructives (on peut passer indéfiniment en négatif, ou bien pivoter 36 fois l'image de 90°, l'information portée par l'image est conservée).

  - La suppression du fond et du bruit est destructive.

  - La suppression du bruit n'est à utilisée que pour les images très bruitées (poussières, petites bulles, etc.), c'est à dire pratiquement jamais.

  - Les figures ne sont pas fermées automatiquement entre chaque visualisation. Les fermer manuellement si nécessaire.
- Étaler les niveaux de gris sur toute la plage d'intensité puis réafficher les images. Constater l'amélioration du contraste.

  Remarque: Un suffixe "_dm" a été ajouté aux noms des images
- Supprimer le fond. ProDIGE affiche une image et donne la main à l'utilisateur. Il faut cliquer sur deux points pour définir un segment correspondant grossièrement au diamètre du spot le plus gros de la série d'images de la sélection de travail.
- Réafficher les images. Constater que, visuellement, les images sont peu modifiées.

  Remarques:

  - Le traitement peut prendre plusieurs minutes.

  - La suppression du fond est plus ou moins visible selon les images et les séries expérimentales.

  - Un suffixe "_bs" a été ajouté aux noms des images
- Tester les options de mise en forme non destructives (puis revenir au format standard)

  Remarques :

- À l'issue ces étapes, le dossier prétraitement contient les images prétraitées.

- Les images intermédiaires ont été supprimées. Si l'on n'est pas satisfait de ce prétraitement, il est nécessaire de supprimer manuellement les images du répertoire prétraitement et de réimporter les images brutes.

Fusion des images

– Lancer Prodige (ou revenir au premier menu s'il est déjà lancé)
– Sélectionner « Fusion d'images »
– Sélectionner les images alignées (pour l'analyse différentielle donc pas les images Cy2)

Remarques :

- Les images alignées sont disponibles ici :
\\Ison\outil_interne_proteomique$\Labo\Personnel\Fabien\TP_ProDIGE\Images_Alignées

- Un répertoire « fusion » est créé dans le répertoire de travail et les images alignées y sont copiées

- Étaler les niveaux de gris (en permettant une légère saturation) des images de la sélection de travail, c'est-à-dire des images que l'on veut fusionner (redéfinir la sélection si nécessaire avec l'option « définir la sélection »).

Remarques :

- Le fait de permettre une légère saturation lors de l'étalement de la dynamique permet de mettre les images globalement au même niveau d'intensité. En effet certaines images sous-exposées ne peuvent pas être mise à niveau par un étalement sans saturation à cause d'une minorité (un seul suffit) de pixel atteignant la limite de la dynamique.

- Mettre la valeur proposée par défaut sauf si cas particulier.

- Paramètre « fraction de pixels autorisés à saturer en basse intensité » (défaut 0.001, classiquement entre 0.001 et 0.005) : une valeur trop faible entraîne une mise à niveau des images potentiellement insuffisante, une valeur trop grande provoque une saturation trop forte des spots qui impacte la suite du traitement.

- Un suffixe "_dms" est ajouté aux noms des images. Ces images ne servent qu'à la fusion.

- Afficher les images obtenues.
- Sélectionner « Fusionner les images sélectionnées ».
- Donner un nom pour l'image de fusion caractéristique du jeu d'images considéré. Observer le répertoire fusion du répertoire de travail.
- Conserver les images ayant servi à la fusion. Observer le répertoire fusion du répertoire de travail.
- Afficher l'image de fusion.

Analyse des images

- Lancer ImageMaster 2D.
- Créer un nouveau workspace (le nommer « TP_ProDIGE » et le sauvegarder en local).

Remarque: En pratique les « workspaces » sont sauvegardés ici : « \\Ison\outil_interne_proteomique$\Labo\Workspace IM 6 ». Il faudra veiller à garder une cohérence au niveau des lieux de sauvegarde.

- Créer un projet au sein de ce workspace (le nommer Janvier_2006 et le sauvegarder dans le même répertoire que le workspace).

Remarque: En pratique les projets sont sauvegardés dans le même répertoire que pour le Workspace (ou un sous répertoire).

- Cliquer droit sur l'icône « Gel » du projet et sélectionner « importer gel ».
- Choisir un facteur de réduction de 1 et le format « TIFF ».
- Sélectionner les images Cy3 et Cy5 obtenues à l'issue de l'étape d'alignement.

Remarque: Les images alignées sont disponibles ici: \\Ison\outil_interne_proteomique$\Labo\Personnel\Fabien\TP_ProDIGE\Images_Alignées

- Dans la nouvelle fenêtre, accepter la dénomination par défaut et le dossier de sauvegarde des fichiers propriétaires (.mel) générés par ImageMaster 2D.

Remarque: Par défaut, les fichiers sont sauvegardés dans le répertoire des images sélectionnées.

- Dans la nouvelle fenêtre, attribuer le cyanine (Cy3 ou Cy5) correspondant à chaque image.

Remarque: Maintenant les fichiers images (.mel) apparaissent dans la structure du workspace.

- Répéter l'opération d'importation afin d'inclure dans le projet l'image de fusion se trouvant dans le répertoire « Fusion » du dossier de travail. Lui attribuer arbitrairement le cyanine Cy2.
- Sélectionner toutes les images de dossier « Gels » de la structure du workspace, cliquer droit et les ouvrir dans une nouvelle feuille (« Open in new Worksheet »).
- Empiler les images, ajuster le contraste, faire défiler les images et juger de la qualité de l'alignement.

Remarque: L'alignement est une étape primordiale dont dépend la qualité de l'analyse. Il est important de juger de sa qualité. Si elle est jugée insuffisante, l'étape d'alignement doit être recommencée quitte à écarter certains patients dont les gels sont problématiques.

- Détecter les spots sur le gel de fusion.
- Retravailler le patron de détection jusqu'à satisfaction (suppression des spots en bordure, des spots artefactuels, modification de contours, ajout de spots).
- Annoter les spots.
- Sélectionner tous les spots et les copier sur les autres images (Cy3 et Cy5).

Remarque: Cette opération de copie des spots est assez (très) longue sous ImageMaster.

- Copier les annotations sur tous les gels.

Remarque: Contrôler que les annotations se sont bien rattachées automatiquement aux bons spots.

- Sélectionner toutes les images Cy3 et Cy5 (exclure l'image de fusion) et sélectionner tous les spots.

- Exporter les données : « Export gel data to XML », une fenêtre s'ouvre, cocher toutes les cases afin d'exporter toutes les informations. Une nouvelle fenêtre s'ouvre, y préciser que l'on exporte toutes les données sélectionnées.
- Donner un nom explicit (ex : Export_janvier_2006) et sauvegarder le fichier xml dans le dossier de travail.
- Sauvegarder le workspace et quitter ImageMaster 2D.

Remarque:

- Pour l'exemple, on peut éviter cette étape d'analyse des images en considérant, pour la suite du workflow, le fichier xml disponible ici :
\\Ison\outil_interne_proteomique$\Labo\Personnel\Fabien\TP_ProDIGE

Traitement des données

- Lancer Prodige (ou revenir au premier menu s'il est déjà lancé).
- Sélectionner « Analyse des données ».
- Sélectionner « Importer les données brutes à partir d'un fichier xml ».
- Donner un nom explicit au fichier Excel qui servira de sauvegardes et de moyen de visualisation des données et des futurs traitements.
- Sélectionner le fichier XML d'export issu de l'analyse d'image.

Remarques:

- L' import des données XML par matlab peut prendre plusieurs minutes.
- Un fichier Excel supplémentaire est créé, il porte le nom spécifié par l'utilisateur augmenté du suffixe « _rawdata ». Il contient les données brutes déjà contenues dans le fichier principal ainsi que des données telles que les contours des spots qui ne sont pas directement utiles à l'utilisateur. Par ailleurs, il est important de ne pas changer les noms des deux fichiers Excel car ProDIGE respecte une nomenclature.

- Sélectionner « préciser la structure DIGE ».
- Sélectionner le fichier Excel principal que l'on vient de créer (pas celui avec l'extension « _rawdata »).
- Excel est lancé automatiquement. Sélectionner, sur le feuillet « Vol Brut », l'ensemble des cellules de données à considérer, colonne A (identifiants) et ligne 1 (nom des images) comprises.

Remarque:Le nombre d'images considérées doit être un multiple de 4 puisque notre schéma expérimental comporte une répétition en Dye-Swap pour chaque patient

- Revenir sous Matlab et cliquer sur OK dans la petite boite de dialogue.
- Répondre aux questions dans la fenêtre de commande Matlab afin de préciser la structure DIGE.

Remarque: Un feuillet « Vol Brut Struct » a été ajouté dans le fichier Excel qui contient les données réorganisées. Les en-têtes des colonnes résument l'information essentielle et sont rangées de manière à correspondre au schéma DIGE en Dye-Swap

- Sélectionner « Normaliser les données ».
- Sélectionner le fichier Excel principal.
- Sélectionner l'ensemble des cellules à considérer sur le feuillet « Vol Brut Struct ».

- Revenir sous Matlab et cliquer sur OK dans la boite de dialogue.
- Ne pas structurer les données (ceci, vient d'être fait, l'option est là pour des cas d'usage particulier de ProDIGE).
- Affecter la valeur la plus faible non nulle aux spots de volume nul.
- Les régressions LTS (Least Trimmed Squares) et Loess (régression polynomiale locale) sont alors effectuées sur les différentes figures M vs. A (M = log2(Tumeur)-log2(Muqueuse) et A = [log2(Tumeur)+log2(Muqueuse)]/2). Des visualisations sont affichées.
- Choisir la régression qui corrige le mieux les biais systématiques observables sur les nuages de points.

Remarque: Si les corrections sont trop sévères, il peut être préférable de ne pas les appliquer.

- Appuyer sur « Entrée », une correction par quantile modérée est alors visible. Juger de l'opportunité de l'appliquer ou non.

Remarques:

- L'impact de la normalisation par quantile est visible grâce aux vecteurs sur les figures M vs. A.

- L'opportunité peut-être jugée en observant la déviation des points d'intensité forte (A élevé). Cette déviation ne doit pas entraîner de changement radical dans l'interprétation

- Moins le jeu de données contient de point plus les différentes normalisations doivent être utilisées avec précaution (le nombre de spots standard est ~2000).

- Observer les différentes visualisations disponibles puis les fermer.

Analyse des données

- Toujours dans le menu « Analyse des données », sélectionner l'option « Analyse différentielle des données (t-test régulé)».
- Sélectionner le fichier Excel de travail.
- Sélectionner toutes les cellules concernées par l'analyse du feuillet « log2(Vol)_N ».
- Revenir sous Matlab et cliquer sur OK dans la petite boite de dialogue.
- Ne pas structurer les données (ceci a déjà été fait : l'option est là pour des cas d'usage particulier de ProDIGE).
- Renseigner les paramètres : donner la valeur 9 au « nombre VO de pseudos observations associées à la variance de fond à considérer ».

Remarques :

- Ce paramètre permet de rendre plus robuste l'estimation de la variance de la quantité protéique d'un spot pour chacune des deux classes. L'estimation n'est pas réalisée avec les seules valeurs disponibles pour le spot dans chaque classe car généralement le nombre de ces valeurs est faible (~5) et ne permet pas une estimation robuste. La méthode utilisée ici est basée sur des principes Bayesiens (cf Baldi et Long, 2001) et permet une estimation de la variance basée non seulement sur la variance propre du spot considéré mais également sur la variance des spots de la même classe d'intensité, la variance de fond. Ceci permet donc une estimation plus robuste de la variance et donc, également, une plus grande cohérence entre les

analyses menées sur des expériences différentes. Le paramètre V0 pondère la variance propre du spot par rapport à la variance de fond. Plus V0 est grand plus on accorde d'importance à la variance de fond. Afin de renseigner ce paramètre, il faut auparavant s'être fixé un nombre N de répétitions idéal et maximal, dans chacune des classes, que l'on ne dépassera jamais (quitte à prendre une marge). Si n est le nombre de répétitions effectif de l'expérience considérée, alors il faut choisir V0 tel que V0+n=N. Par exemple, si l'on choisit N=15 et que le nombre de répétitions dans chacune des classes est n=6 alors il faut prendre V=9.

- Donner la valeur 200 au « nombre total de points du voisinage (en intensité) de chaque spot à considérer ».

- Donner la valeur 0.05 au « seuil de significativité des marqueurs potentiels ».

- Donner un nom à chacune des classes étudiées.
- Sélectionner les images à utiliser pour la visualisation des résultats pour chacune des classes. Prendre de préférence les images prétraitées voire les images archivée lors de la fusion afin d'obtenir une meilleure visualisation (un meilleur contraste).

Ecole Nationale Supérieure des Mines
de Saint-Étienne

N° d'ordre :

Fabien Bernard

Titre de la thèse
IDADIGE: A PROCESS FOR 2D-GEL IMAGES TREATMENT AND ANALYSIS IN THE CONTEXT OF THE SEARCH FOR PROTEIN BIOMARKERS
Spécialité
Image, Vision, Signal
Mots clefs
Electrophoresis, differential, Bidimensional, gel, biomarker, image, workflow, merging, common pattern, at-risk profile, variance stabilisation.

Résumé
This thesis is part of the NODDICCAP project that was initiated by bioMérieux (France) with the aim of developing new tools for screening, diagnosis and evaluation of colorectal cancer prognosis using proteomic tools. The development of these new tools requires the identification of biomarkers that are discriminatory and specific for colorectal cancer.
The aim of this work was to optimise images and datas analyses methods, for the identification of colorectal cancer biomarkers by the use of Differential In-Gel Electrophoresis, a variant of bidimensional gel electrophoresis.
Initially areas for improvement in the traditional methods of 2-D gel electrophoresis image treatment were identified. This led us to reconsider the strategy, and to seek innovative methods for image and data treatment, or to adapt methods issued from closely related research areas. Method selection was guided by the constraints linked to the biological and technological context, as well as the evaluation of their effectiveness when compared to the methods usually used.
The main advances gained from this work are the definition of the experimental set up, the strategic approach for image analysis, and the statistical interpretation of data.
The choice of a standard cell line allowed better comparison of data between experiments.
The strategic analysis of image analysis was improved by the use of a unique detection template. The creation of this template was performed by using a novel method of image merging that enabled accurate representation of each single spot for every image analysed. Finally, methods for statistical analysis of data were proposed that improved ratios studies by considering the range of each spot intensity. Moreover, the specification of an at-risk profile by a biologist allows, for example, to focus more attention on highly expressed spots. Together, all these methods put in place, from image acquisition to discovery and visualisation of potential protein markers, form the Image and Data Analysis for Differential In Gel Electrophoresis (IDADIGE) workflow. This workflow exploits different software and many functions put in place in Matlab and regrouped under the name of ProDIGE.
The proteomics laboratory at bioMérieux currently uses the IDADIGE workflow, which has allowed the discovery of new colorectal biomarkers that now must be validated in a biological context.

Ecole Nationale Supérieure des Mines
de Saint-Étienne

N° d'ordre :

Fabien Bernard

Titre de la thèse
IDADIGE : PROCÉDÉ DE TRAITEMENT DES IMAGES DE GELS D'ÉLECTROPHORÈSE BIDIMENSIONNELLE DIFFÉRENTIELLE DANS LE CONTEXTE DE LA RECHERCHE DE MARQUEURS PROTEIQUES

Spécialité
Image, Vision, Signal
Mots clefs
Électrophorèse, différentielle, Bidimensionnelle, gel, biomarqueur, image, procédé, fusion, patron commun, profil de risque, stabilisation de la variance.

Résumé
Le sujet de cette thèse s'inscrit dans le cadre du projet NODDICCAP initié par l'entreprise bioMérieux et visant le développement de Nouveaux Outils pour le Dépistage, le DIagnostic, l'évaluation du pronostic et le suivi du Cancer Colorectal par une Approche Protéomique. Le développement de ces nouveaux outils passe nécessairement par l'identification de marqueurs tumoraux discriminants et spécifiques du cancer colorectal.
L'objectif de la thèse a été d'optimiser l'analyse des images et des données issues de la technologie DIGE (Differential In-Gel Electrophoresis), afin de permettre la découverte de marqueurs potentiels du cancer colorectal.
Après avoir identifié les maillons faibles de la chaîne de traitement classique des images de gel d'électrophorèse 2D, nous avons été amenés à reconsidérer les approches utilisées et à rechercher des méthodes de traitement d'image et de données innovantes, ou bien existantes mais issues de domaines voisins. Le choix des méthodes a été guidé par l'évaluation de leur efficacité en comparaison aux méthodes classiquement employées, et également par les contraintes liées au contexte biologique et technologique.
Les principales avancées issues de ce travail sont la définition du schéma expérimental, l'approche stratégique de l'analyse d'images ainsi que l'analyse statistique des données.
En ce qui concerne le schéma expérimental, le choix d'une lignée cellulaire comme standard commun a permis un meilleur recoupement des données entre différentes expériences.
L'analyse stratégique de l'analyse d'image a été améliorée grâce à l'utilisation d'un patron de détection unique. Ce patron unique a été réalisé à l'aide d'une méthode de fusion d'images originale permettant une juste représentativité de chacune des tâches protéiques de l'ensemble des images considérées.
Enfin, les méthodes pour l'analyse statistique des données ont tenu compte de l'intensité des tâches protéiques grâce à une régulation de la variance lors de la comparaison des ratios. Par ailleurs, la spécification par le biologiste d'un profil de risque a permis, par exemple, de porter une plus grande attention aux protéines fortement exprimées.
L'ensemble des méthodes mises en place depuis l'acquisition des images jusqu'à la découverte et la visualisation des marqueurs protéiques potentiels constitue le workflow IDADIGE (Image and Data Analysis for Differential In Gel Electrophoresis). Ce workflow exploite différents logiciels ainsi que plusieurs fonctions implémentées sous Matlab et regroupées sous le nom ProDIGE.
L'exploitation par le laboratoire de protéomique de bioMérieux du workflow IDADIGE a été utilisée en routine et a permis la découverte de marqueurs protéiques du cancer colorectal qui doivent maintenant être validés biologiquement.