Histopathologie spectrale du cancer colique : développements d'outils chimiométriques pour le traitement automatisé des images tissulaires en spectroscopie moyen-infrarouge

Warda Boutegrabet

▶ **To cite this version:**

Warda Boutegrabet. Histopathologie spectrale du cancer colique : développements d'outils chimiométriques pour le traitement automatisé des images tissulaires en spectroscopie moyen-infrarouge. Médecine humaine et pathologie. Université de Strasbourg, 2022. Français. NNT : 2022STRAJ013 . tel-04461183

HAL Id: tel-04461183
https://theses.hal.science/tel-04461183

Submitted on 16 Feb 2024

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

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

UNIVERSITÉ DE STRASBOURG

ÉCOLE DOCTORALE des Sciences de la Vie et de la Santé (ED 414)

U1113 INSERM - Interface de Recherche Fondamentale et Appliquée en Cancérologie(IRFAC)

Bio Spectroscopie Translationnelle (EA7506 BioSpecT)

THÈSE présentée par :

Warda BOUTEGRABET

soutenue le : **18 Mars 2022**

pour obtenir le grade de : **Docteur de l'université de Strasbourg**

Discipline/ Spécialité : Médecine translationnelle

TITRE de la thèse

Histopathologie spectrale du cancer colique: Développements d'outils chimiométriques pour le traitement automatisé des images tissulaires en spectroscopie moyen-infrarouge

THÈSE dirigée par :

Mme GUENOT Dominique	Directrice de recherche, Université de Strasbourg
Mr PIOT Olivier	Professeur des universités, Université de Reims-Champagne Ardenne

RAPPORTEURS :

Mr DUPONCHEL Ludovic	Professeur des universités, Université de Lille 1 Sciences et Technologies
Mr GOORMAGHTIGH Erik	Professeur des universités, Université Libre de Bruxelles

---

EXAMINATEURS :

Mr LAQUERRIERE Patrice	Professeur des universités, Université de Strasbourg
Mme TFAILI Sana	Maitre de conférences, Université Paris sud
Mr GOBINET Cyril	Maitre de conférences, Université de Reims-Champagne Ardenne

Je dédie cette thèse à mes exemples de courage et de persévérance,
mes parents et mon grand frère Foudil.

Remerciements

Ce travail de thèse a été effectué dans le cadre d'une collaboration entre deux laboratoires de recherche : l'unité de recherche 1113 INSERM Interface Recherche Fondamentale et Appliquée en Cancérologie (IRFAC) de Strasbourg dirigée par le Dr Jean-Noël FREUND et l'unité de recherche Bio Spectroscopie Translationnelle (EA7506 BioSpecT) dirigée par le Pr Olivier PIOT.

A Monsieur le Docteur Jean-Noël FREUND et Monsieur le Professeur Olivier PIOT,

Je vous remercie de m'avoir accueilli dans vos laboratoires, merci pour l'expérience enrichissante et pleine d'intérêt que j'ai pu avoir durant mes quatre années de thèse. Je vous remercie également pour la confiance que vous m'avez accordée en me donnant l'occasion à de multiples reprises de présenter mes travaux dans des congrès nationaux et internationaux. Soyez assurés de toute mon estime et de mon profond respect.

A Monsieur Professeur Ludovic DUPONCHEL,

Je vous adresse mes remerciements sincères pour l'honneur que vous me faites en acceptant d'être rapporteur scientifique de ce travail de thèse. Je vous remercie de l'intérêt que vous portez à ce travail et du temps consacré à lire le manuscrit et à le juger. Soyez assuré de mon respect et de ma gratitude.

A Monsieur le Professeur Erik GOORMAGHTIGH,

C'est un honneur de vous voir rapporteur de mon travail. Je vous remercie sincèrement pour votre disponibilité et d'avoir accepté cette invitation. Soyez également assuré de tout mon respect et de ma profonde gratitude.

A Monsieur le Professeur Patrice LAQUERRIERE,

Je suis très sensible à l'honneur que vous me faites d'accepter d'examiner mon travail et de siéger parmi les membres du jury. Veuillez accepter l'expression de ma profonde reconnaissance.

A Madame le Docteur Sana TFAILI,

Je vous adresse mes remerciements les plus sincères pour avoir accepté d'être membre du jury. Je suis honorée de votre présence dans ce jury de thèse. Veuillez croire en l'expression de ma gratitude.

La réalisation de ce travail a été possible grâce au concours de plusieurs personnes à qui je voudrais témoigner toute ma reconnaissance et adresser toute ma gratitude.

A Madame le Docteur Dominique GUENOT,

Je vous remercie de m'avoir intégré au sein de votre équipe de recherche « Développement, cancer et cellules souches » et d'avoir dirigé ce travail de thèse et assurer son bon déroulement. Merci pour votre rigueur scientifique et le partage de votre savoir, vos conseils et remarques constructives et la disponibilité à toute épreuve dont vous avez fait preuve à mon égard. Merci

de m’avoir fait découvrir le monde de la recherche et d’avoir participé à développer mon esprit scientifique. Soyez assurée de mon sincère respect et de ma profonde gratitude.

A Monsieur le Professeur Olivier PIOT,

Je t'exprime toute ma reconnaissance d’avoir dirigé ce travail de thèse. Merci pour le partage de ton expertise en spectroscopie, ta rigueur scientifique et ta disponibilité, ainsi que pour ton écoute infaillible. Je te remercie également pour la patience et la confiance que tu m’as témoigné tout au long de ces années et pour tous tes précieux conseils et remarques constructives qui m’ont permis d’ailleurs d’améliorer grandement la qualité de mes travaux.

Merci pour tes qualités humaines, tes paroles réconfortantes lors de mes débuts, tes encouragements, le soutien moral (et matériel dans les moments délicats, merci de m’avoir aidé à m’installer sur Reims et d’avoir été mon garant !).

Je me souviendrai des bons moments partagés ensemble au laboratoire grâce à ta bonne humeur et ta générosité. Soit assuré, Olivier, de toute mon estime et de ma profonde gratitude.

A Monsieur le Docteur Cyril GOBINET,

Quant à mon encadrant de thèse durant ces quatre années, je ne pourrai jamais te remercier assez pour cela. Tu as toujours été disponible, à l’écoute très attentive de mes nombreuses questions et idées.

Je t’adresse Cyril mes remerciements pour le partage de ton expertise, et les compétences que tu m’as apprises en traitement de données, en programmation, pour ta patience et ta rigueur scientifique. Merci de m’avoir appris à prendre du recul, à être critique, à me remettre en cause, à organiser mes idées.

Merci de m’avoir aidé à surmonter les petits (et grands) moments de découragement. Je tiens à te remercier également pour ton soutien moral ininterrompu, et tes précieux conseils sont pour beaucoup dans le résultat final de ce travail.

Je n’oublierai pas ta vision très imagée et enthousiaste qui nous a permis de travailler toujours dans la bonne humeur. Je te manifeste toute ma gratitude pour m’avoir encadré durant cette thèse. Bien sûr tu resteras une référence pour moi. Soit assuré, Cyril de toute mon estime et de ma profonde gratitude.

Je tiens à remercier **Monsieur le Professeur Gérard THIEFIN** et **Monsieur le Docteur Jean-Noël FREUND** d’avoir été les membres de mes comités de thèse pour le suivi de cette thèse et notamment vos différentes critiques et remarques scientifiques qui m’ont appris à bien valoriser mes travaux.

J’adresse mes remerciements à nos collaborateurs cliniciens au sein des services d’Hépato-gastroentérologie et Cancérologie Digestive et de Biopathologie du CHU de Reims : Professeur **Olivier BOUCHE**, Docteur **Camille Boulagnon-Rombi** pour la mise à disposition des échantillons humains et des données cliniques. Je remercie **Nicole BOULAND**, **Saviz NAZRI**

et **Elisabeth MARTINE** pour leur disponibilité et leur aide à la réalisation des coupes tissulaires indispensables à la réalisation de mon travail de thèse.

Ce travail n'aurait pu être mené à bien sans l'aide de différents financeurs qui, au travers de leur soutien matériel, ont reconnu mon travail et m'ont fait confiance : **la Région Grand Est**, **le Cancéropôle Grand Est**, **la Fondation ARC**.

Je ne veux pas oublier les personnels du laboratoire BioSpecT qui ont su, par leurs questions ou leurs conseils, me permettre de toujours pousser un peu plus loin mon raisonnement. Les rapports humains dont j'ai profité à leurs côtés ont fait naître de réels liens d'amitié qui à mes yeux n'a pas de prix.

Je remercie tout particulièrement mon bon ami de bureau, monsieur **Mohammed ESSENDOUBI**. Merci de m'avoir apporté beaucoup de conseils et d'encouragements pendant ces dernières années.

Merci à **Hamid MORJANI** pour ta passion communicative et tes encouragements, tes conseils dont j'ai pu largement profiter. Merci pour les bonnes tartes aux pommes et les gâteaux partagés à de nombreuses occasions!

Merci à **Abdel BELJEBBAR** pour tes critiques scientifiques lors des réunions des doctorants qui m'ont incitée à bien argumenter mes travaux.

Merci à **Ganesh SOCKALINGUM** pour vos conseils et nos nombreuses discussions.

Christine et **Laurence**, **Valérie** merci pour vos conseils, vos explications constructives sur l'utilisation des techniques de laboratoire.

Guillaume, **Marie Pierre**, **Céline,** merci pour vos conseils et les bons moments partagés ensemble, et merci d'avoir rendu les journées à BioSpecT plus sportives!

Si j'ai eu tellement de plaisir à venir travailler à BioSpecT c'est grâce à la bonne ambiance qu'y ont mis mes collègues des laboratoires U1113 INSERM IRFAC et BioSpecT que je remercie pour leur investissement dans les réunions des doctorants, les pots, les sorties hors des laboratoires, leur bonne humeur et les bons moments plus agréables partagés ensemble. Je tiens à remercier tout particulièrement :

Angila et **Fatima** pour leurs nombreuses relectures et corrections de mon anglais. Elles ont toujours répondu présentes, même dans les cas les plus désespérés.

Nicolas d'avoir été disponible pour les discussions scientifiques sur les méthodes chimiométriques et le partage de ton expertise. Merci pour les bons moments partagés ensemble et d'avoir rendu les journées à BioSpecT plus agréables et plus sportives!

Ouma d'avoir toujours été disponible quand j'avais besoin d'une oreille bienveillante, merci pour les éclats de rire, merci pour les bons moments partagés ensemble et d'avoir été mon binôme dans toutes nos idées farfelues, merci de m'avoir apporté ton soutien et tes encouragements.

Elie, Imane, Almar merci pour votre générosité et la patience que vous m'avez accordée au moment de la rédaction de ce manuscrit

Charles, Pierre, Julien, Aimée, Sherine, Stéphane, Lise pour votre bonne humeur et vos encouragements.

Enfin, je suis redevable à mes parents, **Aicha** et **Salah**, pour leur soutien moral et matériel sans failles et leur confiance indéfectible dans mes choix, leurs encouragements face aux épreuves de la vie. Sans vous je ne serais pas ici à écrire ces quelques lignes.

D'autre part, je remercie du plus profond de mon cœur mes très chers, mon grand frère **Foudil**, ma belle-sœur **Ahlem** et ma bien-aimée **Tess** et mon porte bonheur **Sidrah** toujours présents à mes côtés à tout moments. Merci d'avoir supporté mes sautes d'humeur et mes angoisses maladives durant la rédaction de ce manuscrit. Merci également pour votre soutien quotidien et vos encouragements inestimables. Mon frère Foudil, sans ta présence et ton soutien constant, je ne serai pas là où je suis aujourd'hui.

Dans une famille nombreuse, c'est chouette d'être la chouchoutée! Je voudrais remercier chaque membre de cette famille, ma sœur, **Chafia**, mes frères **Rachid**, **Rabeh**, **Nabil**, **Massoud** et mes belles sœurs **Nabila** et **Sihem**, mes chers neveux et nièces **Maissen**, **Hossam**, **Ilyen** dont, malgré la distance, les attentions et soutiens sans failles n'ont jamais cessé de m'accompagner tout au long de ces années.

Les moments difficiles restent gravés dans nos mémoires, tout comme le soutien des personnes qui ont aidé à les surmonter. Un merci tout particulier à mes oncles **Amar** et **Baya** d'avoir assuré le rôle de deuxièmes parents pour mon frère Foudil et moi.

Une pensée à la mémoire de mon oncle **Saïd** dont je n'oublie pas les efforts réalisés pour la réussite de mes études.

Je n'oublie pas non plus mes amies **Massilia**, **Theziri**, **Ikram, Tamazight** et **Salsabil** qui n'ont jamais cessé de me soutenir et de m'encourager. Merci pour votre amitié de longue date. Merci pour le temps partagé ensemble et grâce à qui j'ai passé d'excellents moments.

Table des matières

Liste des tableaux

Liste des figures

Liste des abréviations

CCR	Cancer Colorectal
APC	Polypose Adénomateuse Colis
FCA	Foyers Cryptiques Aberrants
HNPCC	Cancer Colorectal héréditaire sans polypose
EGFR	Récepteur du facteur de croissance épidermique
VEGFR	Récepteurs du facteur de croissance endothélial vasculaire
CIMP	Phénotype de méthylateur d'îlots CpG
CIMP-H	CIMP-élevé
CIMP-L	CIMP-faible
CMS	Consensus Molecular Subtypes
KRAS	V-Ki-ras2 Homologue de l'oncogène viral du sarcome de rat Kirsten
NRAS	Neuroblastoma RAt Sarcoma virus
BRAF	B-Raf proto-oncogènes, serine/threonine kinase
MSH2	Homologue MutS 2
MLH1	Homologue Mut-L 1
MSH6	Homologue MutS 6
MSI	Instabilité des microsatellites
MSI-H	MSI-élevé
MSI-L	MSI-faible
MSS	Microsatellite stable
MYC	Proto-oncogène MYC, facteur de transcription bHLH
CIN	Instabilité chromosomique
LOH	Perte d'Hétérozygotie
PAF	Polypose Adénomateuse Familiale
RER	Erreur de Réplication
MMR	Gènes de réparation de mésappariements (Mis-Match Repair)
PIK3CA	Phosphatidylinositol-4,5-bisphosphate 3-kinase sous-unité catalytique alpha
TP53	Protéine tumorale p53
Wnt	Wingless site
SMAD4	Membre de la famille SMAD 4
TNM	Tumor Node Metastasis
IRM	Imagerie par Résonance Magnétique

CT scan	densitométrie
PET	Tomographie par Emission de Positons
FDG	Fluorodésoxy Glucose
PFS	Survie Sans Progression
OS	Survie globale
CaF2	Fluorure de calcium
IR	Infra-Rouge
FTIR	Transformée de Fourier d'Infrarouge
MCT	Mercure Cadmium Telluride
MSC	Multiplicative Signal Correction
EMSC	Extended Multiplicative Signal Correction
MW-U	Mann–Whitney-U
AG	Algorithme Génétique
ACP	Principal Component Analysis
RF	Random Forest
SG	Savitzky-Golay
SNV	Standard Normal Variate
MCR-ALS	Multivariate Curve Resolution-Alternating Least Squares
PLS	Partial Least Squares
QCL	Quantum Cascade Lasers

Préambule

Le cancer colorectal (CCR) est un problème majeur de santé publique, puisque même si la mortalité est en constante baisse depuis les années 1980 grâce à l'association de la chimiothérapie et de thérapies ciblées, la survie relative à 5 ans reste faible pour les patients atteints d'un CCR métastatique (survie globale de 63.5% tous stades confondus, 47% pour les stades III et 14% pour les stades métastatiques IV qui correspondent au stade de dissémination dans les organes à distance). Ce faible taux de survie des stades avancés peut s'expliquer par une absence de réponse ou le développement d'une résistance aux traitements. Les résistances innées sont généralement associées à la présence de mutations dans des gènes oncogéniques et/ou à l'hétérogénéité moléculaire intratumorale, qui se caractérise par l'expression de plusieurs profils moléculaires différents au sein de la même tumeur. La résistance acquise quant à elle, serait une conséquence des traitements en permettant la sélection, la survie et la prolifération de sous-clones mineurs préexistants dans la tumeur initiale et portant une/des mutation(s) de résistance. Des mutations oncogéniques ont été identifiées dans le CCR, comme par exemple une mutation du gène BRAF, facteur de mauvais pronostic quel que soit le stade de développement de la tumeur, ou encore le gène KRAS dont les mutations ne sont pas favorables à la réponse de la tumeur aux thérapies ciblant les voies de survie et de prolifération. Puisqu'il s'agit d'une pathologie très hétérogène, au stade métastatique, la personnalisation de la prise en charge médico-chirurgicale est un enjeu majeur. A ce jour, la prédiction d'une réponse à un traitement s'appuie sur des critères moléculaires (identification de mutations), morphologiques (densité tumorale, degré de différenciation, présence de mucus, ...), et l'expression de biomarqueurs plasmatiques qui sont en cours d'étude afin de prédire l'efficacité des thérapies. Cependant, ces critères requièrent des technologies et manipulations dont les résultats ne sont pas immédiats et pas encore suffisamment fiables pour poser un diagnostic adapté à chaque patient. Il apparait donc crucial de développer une approche applicable en routine clinique, permettant d'identifier des marqueurs prédictifs de la réponse thérapeutique et prenant en compte la variabilité moléculaire associée à l'hétérogénéité tumorale.

L'histopathologie spectrale (SHP) est une approche innovante pour caractériser des échantillons de tissus. Cette méthode combine l'imagerie microscopique par absorption moyen-infrarouge (IR) et un traitement chimiométrique des données spectrales multivariées. Elle permet de distinguer les principales structures histologiques qui constituent les tissus normaux et les tissus tumoraux. Par conséquent, l'approche SHP peut non seulement détecter automatiquement les

cellules cancéreuses, mais aussi caractériser le microenvironnement tumoral. Une hétérogénéité spectrale peut être mise en évidence au sein de ces différentes structures, ouvrant la voie à une caractérisation de l'hétérogénéité moléculaire des tissus tumoraux. Des développements menés au sein de l'unité BioSpecT ont permis aussi d'analyser directement des échantillons inclus en paraffine, sans nécessiter de déparaffinage chimique. Néanmoins, cette histopathologie spectrale nécessite actuellement que l'opérateur fixe certains paramètres de manière parfois subjective selon l'expérience de cet opérateur. Il est donc nécessaire de perfectionner la méthodologie de façon à proposer une méthode automatisée et optimisée, entièrement indépendante de toute intervention de l'opérateur. Cette étape est cruciale pour tirer profit du potentiel de l'imagerie spectrale infrarouge, et ainsi permettre d'identifier des marqueurs spectroscopiques de nature numérique à valeur diagnostique, notamment pour le cancer colique métastatique humain où la prédiction de la réponse thérapeutique constitue un réel enjeu médical.

Par conséquent, l'objectif de ce travail de thèse a été de développer une méthodologie visant à optimiser l'histopathologie spectrale infrarouge appliquée à des échantillons tumoraux humains dont le but ultime sera l'identification de marqueurs spectraux prédictifs de la réponse au traitement. De tels marqueurs aideront les cliniciens à identifier les lésions susceptibles d'être résistantes au traitement, ouvrant ainsi la voie à une médecine de précision en sélectionnant une thérapie personnalisée à chaque individu.

Dans ce travail de recherche centré sur des développements méthodologiques de l'histopathologie spectrale infrarouge, nous nous sommes intéressés à deux sujets en particulier :

1/ l'automatisation de l'étape de prétraitement des données infrarouges par EMSC (Extended Multiplicative Signal Correction) pour une identification optimale des spectres tissulaires lors du déparaffinage numérique des coupes histologiques (Article #1). La méthodologie mise au point sur les échantillons inclus en paraffine a ensuite été transférée à la correction du signal de l'OCT (Optimal Cutting Temperature) dans le cas des cryosections.

2/ la sélection non-supervisée des variables spectrales (nombres d'onde) discriminantes en incorporant au sein d'un algorithme génétique une fonction objectif quantifiant la séparabilité entre des clusters de données et leur compacité. Ce développement a d'abord été testé sur des banques de données publiques (Article #2), puis évalué sur des tissus de côlon.

Pour chacun de ces deux développements, des jeux de données simulées mimant des enregistrements tissulaires ont également été créés.

La description de ces développements originaux est complétée dans ce manuscrit par une introduction décrivant succinctement la problématique médicale en question, à savoir la prédiction de la réponse au traitement dans le cancer du côlon métastatique, le principe de la spectroscopie moyen-infrarouge et de l'histopathologie spectrale. Une discussion sur les expérimentations réalisées et les perspectives envisagées est également présentée.

Liste des communications

Publications dans des journaux internationaux

L'ensemble de ce travail de thèse a permis la production de travaux scientifiques et de participation à des congrès nationaux et internationaux :

En 2021, une publication est parue dans un journal scientifique international à comité de lecture (Article #1, Chapitre III) :
Boutegrabet, W., Guenot, D., Bouché, O., Boulagnon-Rombi, C., Marchal Bressenot, A., Piot, O., Gobinet, C. Anal. Chem. Automatic identification of paraffin pixels on FTIR images acquired on FFPE human samples. Anal Chem, 2021 Mar 2;93(8):3750-3761. doi: 10.1021/acs.analchem.0c03910.

Un article est en cours de finalisation pour soumission (Article #2, Chapitre V) :
Boutegrabet, W., Piot, O., Guenot, D., Gobinet, C. Unsupervised feature selection by a genetic algorithm for mid-infrared spectral data.

Participations aux congrès

- Communications orales

Boutegrabet W., Guenot D., Bouché O., Boulagnon-Rombi C., Marchal Bressenot A., Piot O., Gobinet C. Optimisation de l'histologie spectrale appliquée à des images spectrales infrarouges de coupes de cancer colique humain. 11ème Forum du Cancéropôle Est, 15-16 novembre 2018, Reims, France.

Boutegrabet W., Guenot D., Bouché O., Boulagnon-Rombi C., Marchal Bressenot A., Piot O., Gobinet C. Automation of outlier removal for the improvement of IR spectral histology applied

to human côlon cancer samples. Workshop on Machine Learning and Chemometrics in Biospectroscopy (BioSpecMLC 2019), 18-21 August 2019, Minsk, Belarus. (Une bourse de soutien d'une somme de 800€ par le Cancéropôle Est a été accordée).

Boutegrabet W., Guenot D., Bouché O., Boulagnon-Rombi C., Marchal Bressenot A., Piot O., Gobinet C. Methodological development for the improvement of IR spectral histopathology applied to human côlon cancer samples. Congrès Chimiométrie 2020, 27-29 janvier 2020, Liège, Belgique.

Boutegrabet W., Guenot D., Bouché O., Boulagnon-Rombi C., Marchal Bressenot A., Piot O., Gobinet C. An automatic method for removal of paraffin pixels for the improvement of IR spectral histology applied to human côlon cancer samples. Journée rémoise des jeunes chercheurs en santé (JRJCS 2020), 22 octobre 2020, Reims, France.

Boutegrabet W., Guenot D., Bouché O., Boulagnon-Rombi C., Marchal Bressenot A., Piot O., Gobinet C. Automation of outlier removal for the improvement of IR spectral histology applied to human côlon cancer samples. 16th International Conference on Laser Applications in Life Sciences, 1-2 avril 2022, Nancy, France.

- Communications affichées

Boutegrabet W., Guenot D., Bouché O., Boulagnon-Rombi C., Marchal Bressenot A., Piot O., Gobinet C. Optimisation de l'histologie spectrale appliquée à des images spectrales infrarouges de coupes de cancer colique humain. Journée des Jeunes Chercheurs de la SFR CAP-Santé (JJCS 2018), 18 octobre 2018, Reims, France.

Boutegrabet W., Guenot D., Bouché O., Boulagnon-Rombi C., Marchal Bressenot A., Piot O., Gobinet C. Automation of outlier removal for the improvement of IR spectral histology applied to human côlon cancer samples. 18th European Conference on Spectroscopy of Biological Molecules (ECSBM 2019), 19-22 August 2019, Dublin, Ireland.

Boutegrabet W., Piot O., Guenot D., Gobinet C. Development of a new unsupervised feature selection based on genetic algorithm: application on the FTIR images of human côlon cancer. Journées Jeunes Chercheurs en Cancérologie - Fondation ARC (JJC2021), 18-19 novembre 2021, Paris, France.

Chapitre I : Introduction

I. A. Cancer colorectal

I. A. 1. Généralités et épidémiologie

Le cancer colorectal (CCR) est l'une des tumeurs malignes les plus répandues au monde et se développe dans la muqueuse du côlon ou du rectum. Dans le monde, plus de 1,9 millions de personnes développent un CCR chaque année (dont 20% avec métastases au moment du diagnostic) et environ 935 000 décès ont été rapportés en 2020 (1). En France, le CCR représente près de 12 % de tous les décès par cancer, plus spécifiquement chez les 65 ans et plus (2). Ces données font de ce cancer, la troisième maladie la plus fréquemment diagnostiquée chez l'homme (11,2 % en France) et la principale cause de décès par cancer (9.3% du total des décès) après le cancer de la prostate (14.1%) et du poumon (14.3 %). Chez la femme, le CCR occupe la deuxième place après le cancer du sein (24.5 % du nombre total de cas (3). Il convient de noter qu'en fonction du degré de développement économique et des modes de vie associés, la mortalité et l'incidence du CCR varient d'un pays à l'autre. En effet, en liant le taux d'incidence à l'Indice de Développement Humain (IDH), l'incidence du CCR chez les hommes et les femmes est environ 6 à 7 fois plus élevée dans les zones à IDH très élevé par rapport aux zones à faible IDH (4). En outre, les pays avec un IDH très élevé ont un taux d'incidence de 335,3 et 267,6 pour 100 000 personnes, tandis que dans les pays/régions faibles, les taux d'incidence pour les hommes et les femmes sont respectivement de 104,3 et 128,0 pour 100 000 personnes. La Figure 1 montre la distribution de l'incidence du cancer colique (A) et du rectum (B) dans différents pays pour les hommes et les femmes.

Cependant, le taux d'incidence varie également avec l'âge au moment du diagnostic, le sexe et la race. Le sexe est un facteur de risque élevé de développer un CCR, les hommes étant plus susceptibles de développer un CCR que les femmes (5). L'âge est également un facteur de risque important dans le développement du cancer, puisque le CCR affecte principalement les personnes adultes de plus de 50 ans), ce risque augmentant de 1 % à 4 %/an avec l'âge (6). Dans le cas où il existe des antécédents familiaux de CCR au premier degré, ce risque est doublé.

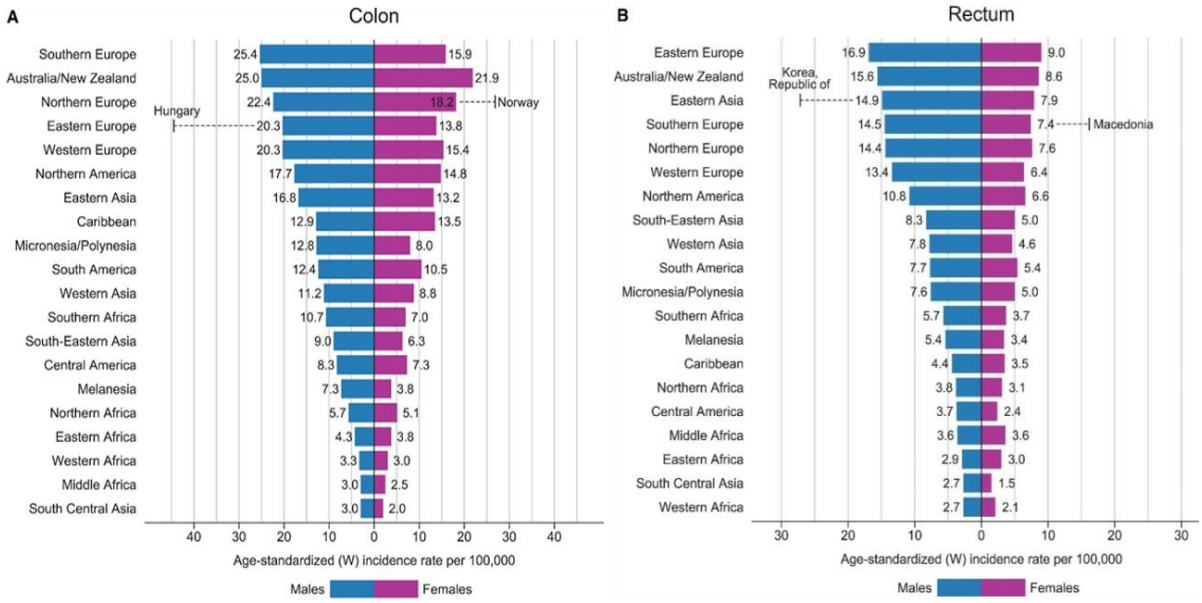


Figure 1. Diagramme à barres des taux d'incidence normalisés selon l'âge par région et par sexe pour les cancers du (A) côlon et (B) du rectum (y compris l'anus) en 2020. Les taux sont indiqués par ordre décroissant du taux mondial (W) normalisé selon l'âge. Source : GLOBOCAN 2020 (1).

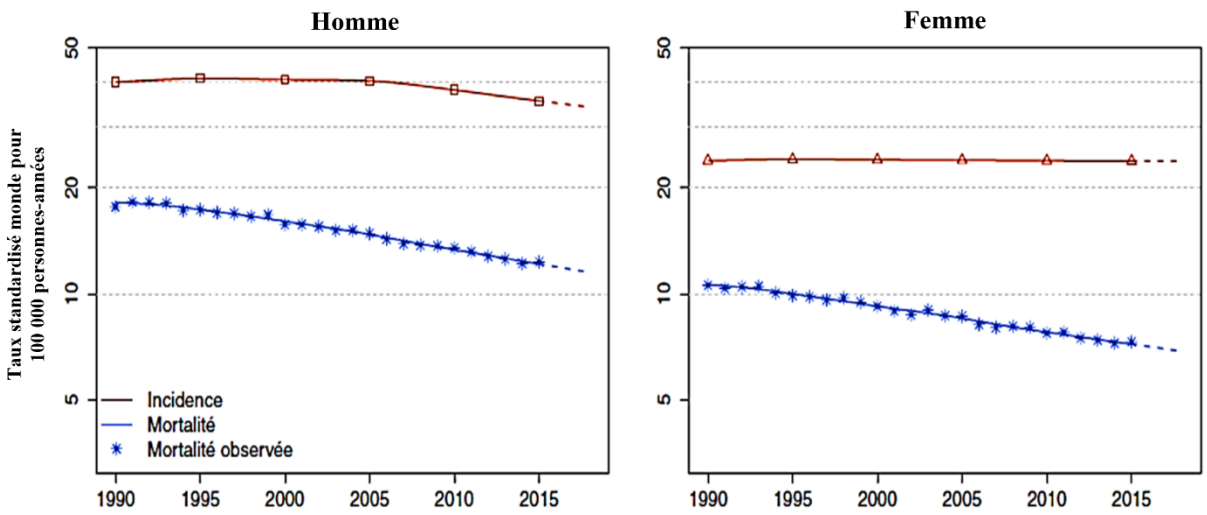


Figure 2. Taux d'incidence et de mortalité par cancer colorectal en France de 1990 à 2018 (2).

En termes de race, de nombreuses disparités raciales/ethniques sont observées dans le risque de développer un cancer à un stade avancé et le risque de mortalité, et de grandes variations existent entre les différents sous-groupes ethniques (7–9). La diminution relative de la mortalité par CCR observée dans les pays à fort IDH comme la France (Figure 2) peut s'expliquer par la

mise en place de programmes de dépistage précoce et l'adoption des nouvelles associations thérapeutiques (10–12).

I. A. 2. Méthodes de diagnostic de CCR

Le diagnostic d'un CCR peut se faire de deux manières différentes. Dans le premier cas, le patient présente des symptômes cliniques (présence de sang dans les selles, douleurs abdominales) et est diagnostiqué sur la base de ces symptômes. Le second cas correspond au résultat d'un dépistage régulier effectué par le patient asymptomatique de 50 à 74 ans. Depuis 2015, un test immunologique quantitatif (FIT ou Fecal Immunological Test) est très utilisé dont la sensibilité pour le carcinome est de 91 % et de 40 % pour l'adénome avancé.

I. A. 3. Facteurs de risque

Des études épidémiologiques cumulatives ont confirmé que l'apparition du CCR est associée à de multiples facteurs, notamment la prédisposition génétique, le régime alimentaire riche en graisse, la sédentarité, le surpoids, l'obésité, l'inactivité physique, le tabagisme et l'abus d'alcool (13). En conséquence, comme le rapporte le World Cancer Research Fund/American Institute for Cancer Research, une alimentation riche en viande rouge ou transformée, ainsi que les boissons alcoolisées représentent le risque le plus élevé de développer un CCR (14–16). Tous ces facteurs peuvent en partie expliquer la prévalence élevée des CCR sporadiques dans les pays développés (11).

I. A. 4. Physiopathologie du CCR

La plupart des CCR surviennent sporadiquement à la suite d'une carcinogenèse en plusieurs étapes, caractérisée par des altérations génétiques (mutations, anomalies du nombre de chromosomes) et des altérations épigénétiques (méthylation de l'ADN ; miRNA) ; (17,18). Dans la majorité des CCR, ce sont des mutations génétiques et des modifications épigénétiques (19) qui conduisent à la formation d'un carcinome (Figure 3). Toutefois, il a été noté que le risque de la transformation de la muqueuse normale en muqueuse cancéreuse dépend de la taille, du degré de dysplasie et de la nature des altérations génétiques des polypes/adénomes, ou la présence d'une composante villeuse.

Si la majorité des CCR sont des cancers sporadiques (95%), il existe 5% des CCR d'origine héréditaire et qui résulte généralement de la présence d'une mutation génique, entrainant l'activation d'un oncogène ou la perte d'un gène suppresseur de tumeur dans les cellules

germinales. L'acquisition de nouvelles mutations/anomalies chromosomiques au cours de la vie des individus va favoriser le développement d'un CCR.

Pour ce qui concerne les CCR héréditaires, on retrouve le syndrome de Lynch est un CCR héréditaire sans polypose (Hereditary non polyposis colorectal cancer, HNPCC) et correspond à 3% des CCR héréditaires. Ce syndrome est causé par des mutations de gènes de réparation de mésappariements de l'ADN (MisMatch Repair : MMR) (20,21). De même, la polypose adénomateuse familiale (FAP), responsable de 1 à 2% des CCR héréditaires. Cette dernière est causée par les mutations germinales du gène suppresseur de tumeur Adenomatous Polyposis Coli (APC) et se caractérise par une hyper prolifération de l'épithélium dans la muqueuse colique qui entraine la formation de plusieurs centaines de polypes (22).

I. A. 4. 1. Les lésions précurseurs

La première étape du développement du CCR est l'apparition d'une hyper-prolifération de l'épithélium (hyperplasie) décelable au niveau des cryptes qui évolue vers un cancer selon une séquence polype-adénome-carcinome (Figure 3). Parmi les polypes, définition macroscopique, on distingue 3 types de lésions :

- les polypes non néoplasiques (20 à 35 % des personnes âgées de plus de 50 ans), les polypes néoplasiques pouvant évoluer en carcinome invasif (70 à 75 % des polypes). Pour les lésions adénomateuses, le risque de cancer croît avec le nombre, la taille des adénomes (> 1 cm) et la proportion du contingent villeux. En effet, elles peuvent être classées selon leur risque croissant de transformation tumorale. D'autre part, il existe également un autre type de polype à fort potentiel tumoral, les polypes dentelés sessiles ou polypes festonnés (5-10% des polypes). Ces polypes se développent à partir d'événements moléculaires et histologiques distincts des adénomes tubulaires (23). Suivant les publications, de 65 à 90% des CCR proviendraient d'adénomes conventionnels et de 10 à 35% d'adénomes festonnés (24).

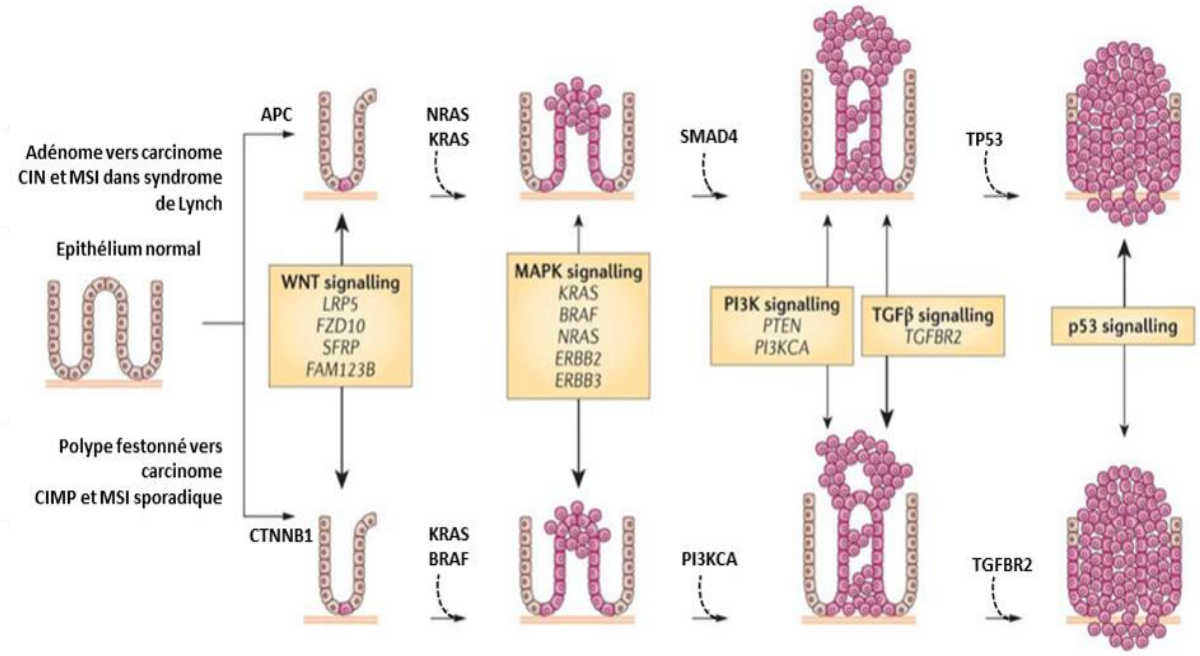


Figure 3. Séquences de progression du polype vers le carcinome. La voie « classique » ou traditionnelle (haut) implique le développement d'adénomes tubulaires qui qui peuvent évoluer vers des adénocarcinomes. Une voie alternative (bas) implique des polypes dentelés et leur progression vers le cancer colorectal dentelé. D'après Kuipers et al, 2015 (25).

I. A. 4. 2. Les carcinomes colorectaux

Il existe des différences anatomiques et fonctionnelles entre la partie proximale 30-40% des CCR) et la partie distale (20%) du côlon (et 30-40% rectum) dont l'origine embryonnaire (partie proximale dérive de l'intestin primitif moyen ; la partie distale de l'intestin primitif postérieur). La physiologie diffère avec des variations dans la concentration en sels biliaires, la composition et la densité en bactéries, le métabolisme de l'épithélium. De même, des études transcriptomiques ont mis en évidence des gènes différentiellement exprimés dans le côlon proximal et le côlon distal (26). Ces différences justifient l'émergence de voies de carcinogenèse différentes avec des caractéristiques différentes pour les deux localisations (27) (Figure 4).

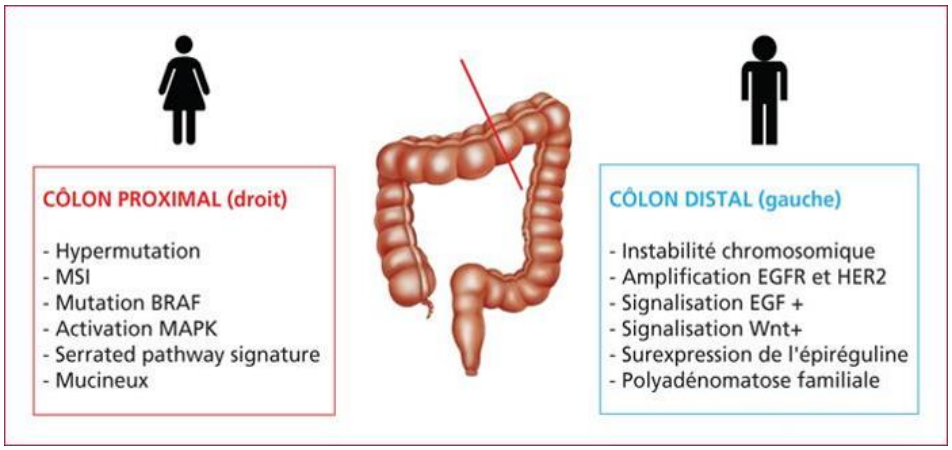


Figure 4. Caractéristiques cliniques et moléculaires des tumeurs proximales et distales. L'instabilité chromosomique est plus souvent observée dans le côlon distal que l'instabilité des microsatellites et le phénotype méthylateur sont plus fréquemment observés dans le côlon proximal.

I. A. 5. Classification TNM des CCR

La classification est essentielle pour établir un pronostic (prévision de l'issue de la maladie) et déterminer les choix de traitement. La classification dite « Tumor, Nodes, Metastasis » ou TNM a été mise en place par l'Union Internationale Contre le Cancer (UICC) et les trois lettres symbolisent la propagation de la tumeur : (T) : taille de la tumeur primitive, (N) présence et nombre de ganglions régionaux envahis (N pour node en anglais) et (M) présence (M1) ou absence de métastases (M0), le foie étant la cible préférentielle des métastases dans 75% des cas. La classification TNM permet de classer les CCR en cinq stades qui sont indiqués sur la Figure 5.

L'intérêt majeur de la classification TNM est sa forte association au pronostic. En effet, la survie globale à 5 ans pour des patients diagnostiqués entre 2010 et 2016 et par stades est de 91 % pour les stades I et II, 72 % pour le stade III et 14 % au stade IV (28).

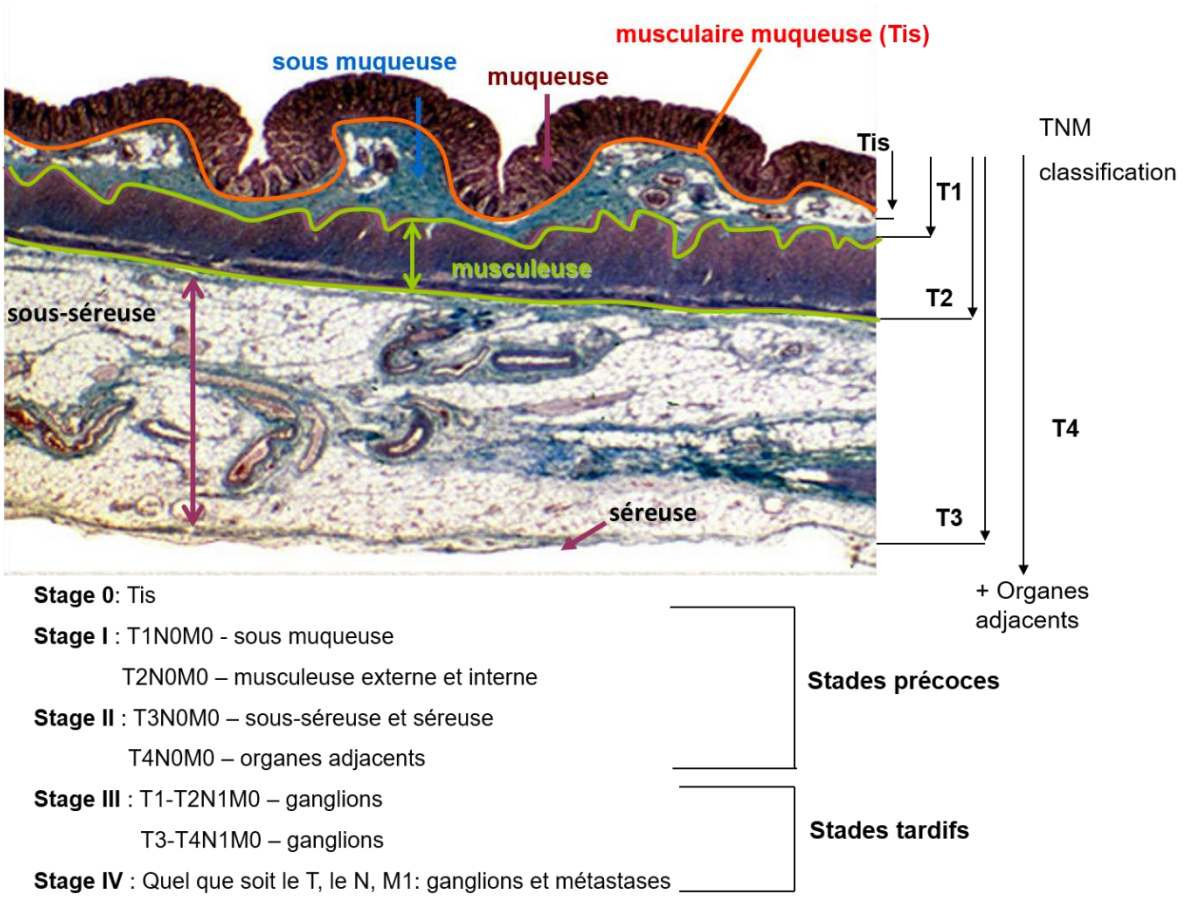


Figure 5. Classification UICC des cancers colorectaux en fonction du statut TNM de la tumeur. Tis : tumeur in situ, intra-épithéliale. T1, atteinte de la muqueuse et sous-muqueuse ; T2, infiltration de la musculeuse muqueuse sans dépassement ; T3, atteinte de toute l'épaisseur de la paroi ; T4, atteinte des organes adjacents. N0, absence d'atteinte ganglionnaire ; N1, atteinte de 1 à 3 ganglions péri tumoraux ; N2, atteinte de plus de 4 ganglions. M0, absence de métastases ; M1, présence de métastases à distance.

I. A. 6. Caractérisation moléculaire des CCR

La progression tumorale résulte d'une accumulation séquentielle d'altérations par mutations de gènes suppresseurs de tumeur, d'oncogènes, de gènes impliqués dans la réparation de l'ADN ou par l'acquisition de gains ou pertes de chromosomes/fragments de chromosomes (29). Sur la base de ces observations, un modèle décrivant la cinétique d'apparition des modifications moléculaires conduisant au CCR a été proposé (19). Ces séquences différencient deux voies de carcinogenèse : la plus fréquente étant l'instabilité des chromosomes (CIN ou microsatellite stable, MSS) et la seconde, l'instabilité des loci de microsatellites (MSI ou MIN, ou Replication ERror RER), les microsatellites étant séquences di-, tri- ou tétra-nucléotidiques répétées et hautement polymorphes, réparties dans les séquences codantes et non codantes du génome. Une

troisième voie d'instabilité d'origine épigénétique a été décrite plus récemment et définit le phénotype méthylateur (CIMP) et se caractérise par la méthylation des îlots CpG de l'ADN (30); (Figure 3).

I. A. 6. 1. Le phénotype CIN

Ce phénotype dans lequel l'intégrité de l'ADN est perturbée, représente 80-85% des CCR sporadiques. On le retrouve principalement dans le côlon distal, caractérisé par des altérations chromosomiques qui génèrent des mutations, des pertes/gains de chromosomes ou fragments de chromosomes, conduisant à un nombre anormal de chromosomes ou aneuploïdie.

Plus de 70% des CCR de phénotype CIN présentent une perte allélique en 18q entraînant la perte d'expression de gènes suppresseurs de tumeur de la voie de signalisation du TGF-β qui régule la prolifération cellulaire et l'induction de la transition épithelio-mésenchymateuse.

Retrouvée dans 70% des CCR CIN, la perte d'un allèle au locus 5q21 où est localisé le gène suppresseur de tumeur APC, est très fréquemment associée à des mutations du second allèle (31). Les mutations conduisent à l'apparition d'un codon stop et à la synthèse d'une protéine tronquée non fonctionnelle. Le gène APC code pour une protéine impliquée dans l'adhésion et la migration cellulaire. Les mutations du gène APC entrainent l'activation constitutive de la voie canonique Wnt/β-caténine impliquée dans le processus de cancérogenèse colique (29).

Le gène suppresseur de tumeur TP53 dont la protéine p53 est le «gardien du génome», bloque le cycle cellulaire, permettant la réparation des lésions de l'ADN, et induisant la mort cellulaire si les lésions n'ont pu être réparées. 75% des cas de CCR présentent une perte allélique au niveau du locus 17p13, où se situe TP53 et cette perte allélique est généralement associée à une mutation inactivatrice sur le second allèle (32).

I. A. 6. 2. Le phénotype MSI

L'instabilité de microsatellites caractérise 15 à 20% des CCR sporadiques et 95% des syndromes héréditaires de Lynch. Ils sont plutôt localisés dans le côlon proximal et l'instabilité résulte d'une défaillance du système de réparation des mésappariements de l'ADN (MMR). C'est l'accumulation dans l'ADN tumoral, de mutations somatiques engendrées par le phénotype MSI qui participe au processus de transformation multi-étapes conduisant à générer une tumeur MSI diploïde (33).

L'acquisition de la déficience du système MMR repose sur une perte d'expression d'une des 4 protéines MMR : MLH1, MSH2, MSH6 ou PMS2; les cas sporadiques sont essentiellement liés à la perte d'expression de MLH1 dans la tumeur par hyperméthylation de son promoteur, riche en îlots CpG (34).

Le syndrome de Lynch est essentiellement associé à des mutations germinales des gènes MLH1 ou MSH2 dans 90% des cas. Une copie d'un gène MMR est donc constitutionnellement inactivée dans les cellules germinales.

Dans les tumeurs MSI, c'est l'instabilité génétique élevée (environ une centaine de mutations /mégabase en moyenne) qui est à l'origine de nombreux évènements somatiques, principalement dans les microsatellites. Ces altérations somatiques de microsatellites participent au développement tumoral lorsqu'elles modifient la fonction de gènes ayant un rôle dans l'oncogenèse. Le statut MSI des tumeurs confère un meilleur pronostic au patient (35) et de nombreuses études rétrospectives incluant des CCR tous stades confondus, ont rapporté des survies globales et sans récidive significativement meilleures pour les CCR MSI par rapport aux CCR CIN (36).

I. A. 6. 3. Le phénotype CIMP

Cette voie d'instabilité d'origine épigénétique se caractérise par un phénotype méthylateur ou CIMP par méthylation des îlots CpG de l'ADN, ce qui entraîne une répression transcriptionnelle du gène. Dans le CCR, cette hyperméthylation cible environ 10% de gènes qui sont normalement exprimés dans les cellules dont le gène hMLH1 du système MMR (30). Les CCR CIMP ont une localisation plutôt proximale, sont plus fréquentes chez les femmes, se retrouvent chez des patients d'un âge avancé et sont d'un grade tumoral élevé. Ce phénotype représente environ 20-30 % de l'ensemble des CCR et près de 2/3 des tumeurs MSI. Il peut être présent dans les polypes où il est plutôt associé aux polypes festonnés (37). Une classification des CCR en fonction de la fréquence de méthylation des îlots CpG a permis de distinguer les CCR CIMP-high, CIMP-low et non-CIMP (38); (Figure 6), chacun associé à des mutations de gènes spécifiques (Table 1). Le phénotype CIMP-H est un facteur de mauvais pronostic et est associé à une survie sans progression et une survie globale plus courte que le phénotype CIN ou MSI associé (39).

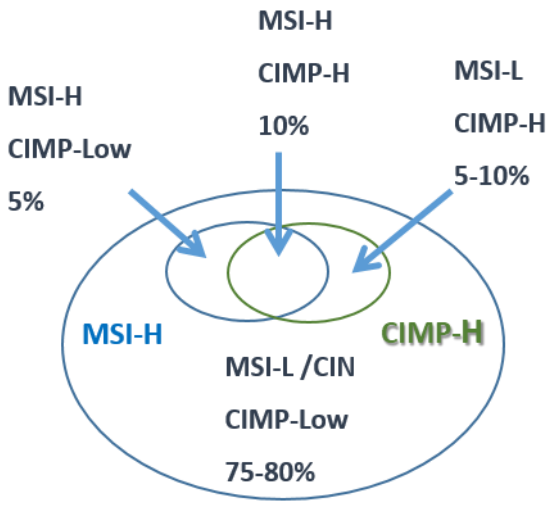


Figure 6. Les sous-types de cancer colorectal selon le statut MSI et CIMP. D'après Ogino et Goel, 2008 (40).

	NON CIMP	CIMP-Low	CIMP-High
Localisation	distal< proximal		
Ratio H/F	hommes = femmes	hommes > femmes	femmes > hommes
Mutations			
BRAF	sauvage	sauvage	muté
KRAS	muté	muté	sauvage
TP53	70%	30%	10%
Phénotype associé	CIN	CIN	MSI

Table 1. Caractéristiques moléculaires des tumeurs CCR de phénotype CIMP. D'après Ogino et al, 2007 (38).

I. A. 6. 4. Classification CMS

Le CCR est un cancer très hétérogène et des sous-typages moléculaires basés sur l'expression de gènes ont été développés. Un consortium international d'experts a dégagé, à partir de 6 classifications moléculaires et plus de 4 000 patients, la classification CMS pour Consensus Molecular Subtypes qui définit 4 groupes de cancers colorectaux (41) caractérisés par des facteurs moléculaires, biologiques et cliniques (Table 2). La classification CMS a une valeur pronostique importante puisque pour un CCR non métastatique (stades de 0 à III), le pronostic est favorable pour les tumeurs du sous-groupe CMS-1 et dans une moindre mesure du sous-groupe CMS-2 mais en situation métastatique (stade IV), c'est le sous-groupe CMS-1 qui est lié au plus mauvais pronostic (41); (Figure 7).

CMS 1 MSI – Immunité	CMS 2 Canonique	CMS3 Métabolisme	CMS4 Mésenchymateux
14%	37%	13%	23%
MSI, CIMP-high Hypermuté	Forte variation du nombre de copies de gènes	MSI mixte -CIMP-low Nombre copies de gènes - low	Forte variation du nombre de copies de gènes
BRAF muté		KRAS muté	
Infiltration et activation immunitaire	Voie WNT et activation gène MYC	Dérégulation métabolique	Infiltration stromale Voie TGF + Angiogenèse
Mauvaise survie et récidive			Mauvaise survie globale et survie avec récidive

Table 2. Taxonomie proposée du CCR, reflétant des différences biologiques significatives dans les sous-types moléculaires basés sur l'expression des gènes. CIMP, phénotype méthylateur de l'îlot CpG ; MSI, instabilité des microsatellites. D'après Guinney et al, 2015 (41).

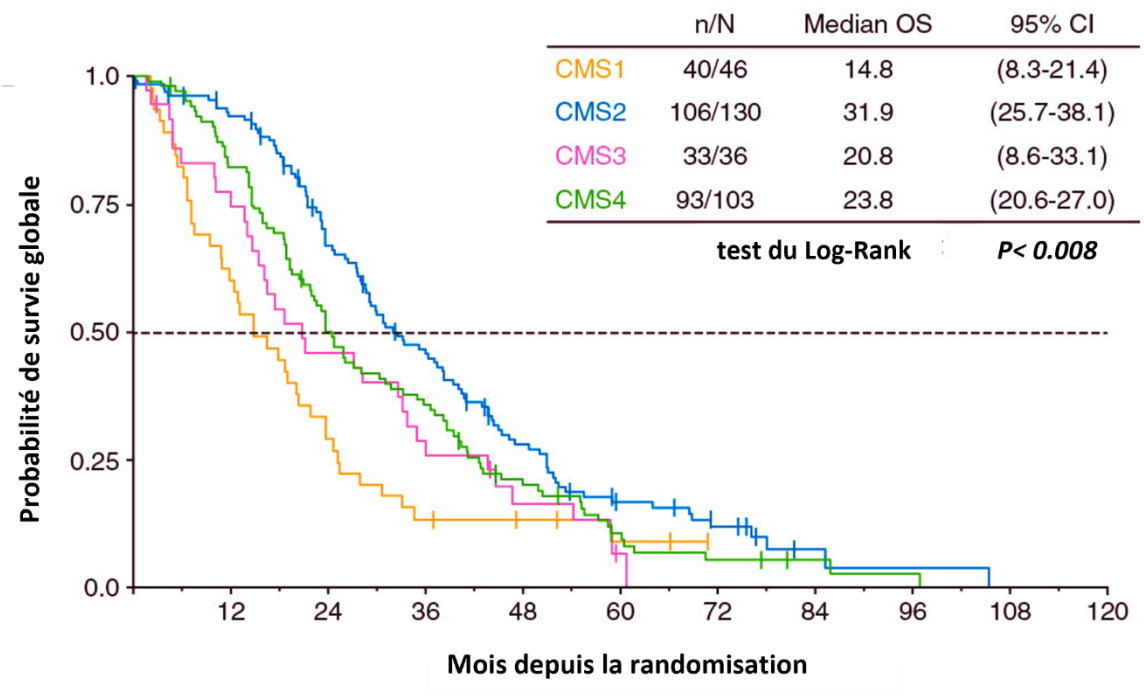


Figure 7. Impact pronostique de la classification CMS en situation métastatique. D'après Stinzing S et al. 2019 (42).

I. A. 7. Voies de signalisation impliquées dans la carcinogenèse colique

I. A. 7. 1. Voie Wnt/APC/β-caténine

La voie Wnt est une voie de signalisation majeure dans l'embryogenèse et la régulation de l'homéostasie intestinale et les protéines centrales sont la protéine APC et la β-caténine. En l'absence de signal mitogène, la β-caténine est séquestrée dans un complexe cytoplasmique avec la protéine APC, grâce à une sérine-thréonine glycogène-kinase (GSK3b) qui permet la dégradation de la β-caténine par le protéasome. Lorsqu'un signal mitogène est délivré par le proto-oncogène Wnt ou en l'absence d'une protéine APC fonctionnelle, la GSK3b est inhibée et la β-caténine est transloquée vers le noyau où elle permet la transcription de gènes favorisant la prolifération cellulaire.

La perte de fonction de la protéine APC est une conséquence de la mutation du premier allèle puis l'inactivation du deuxième allèle, soit par une seconde mutation, soit par la perte d'un des deux bras longs du chromosome 5 (Figure 8). Dans les cancers CIN, la voie de signalisation Wnt est principalement activée par une inactivation bi-allélique du gène APC alors que dans les cancers MSI des mutations activatrices de la β-caténine ou inactivatrices de l'axine 2 sont décrites (19,43).

I. A. 7. 2. Voie RAS/MAPK

La fonction de la voie de signalisation RAS/MAPK est d'intégrer les signaux extracellulaires et de coordonner une réponse appropriée pour contrôler la croissance, la survie et la différenciation cellulaire. L'activation aberrante de cette voie est un événement oncogénique majeur, très répandu dans de nombreux cancers (44), dont 55 % des cancers du côlon et du rectum avec des mutations des gènes KRAS, NRAS ou BRAF (45).

I. A. 7. 2. 1. Oncogène KRAS

La cascade Ras/MAPK transmet des signaux extrinsèques venant de récepteurs situés à la surface des cellules, en particulier de signaux mitogènes, via les récepteurs à tyrosine kinase comme l'EGFR ou le VEGFR. Leur surexpression affecte l'adhésion et le cycle cellulaires. Dans cette famille, une mutation de l'oncogène KRAS est observée dans presque 50% des CCR sporadiques de phénotype CIN. La mutation de l'oncogène KRAS est un événement précoce

dans la carcinogenèse puisqu'elle est observée avec la même fréquence dans le cancer et les adénomes de plus de 1cm.

I. A. 7. 2. 2. Oncogène BRAF

D'autres acteurs majeurs de la voie des MAPK sont fréquemment mutés tel que le gène BRAF avec la mutation V600E. Cette mutation ponctuelle BRAFV600E entraîne un changement de la conformation de la protéine BRAF qui devient constitutivement active. L'incidence de ces mutations dans le CCR est estimée entre 4 % et 18 % des cas, entre 5 % et 10 % des cas de CCR métastatique. Elle est plus fréquemment associée aux lésions proximales (60 à 80 % et moins de 5 % pour des lésions distales), chez des patients âgés, de sexe féminin et survient surtout dans les cancers sporadiques de phénotype MSI (50-80%, 5-10 % des CIN) ; (46). Les mutations BRAFV600E sont également associées au phénotype CIMP, entraînant une hyperméthylation du promoteur du gène MLH1 qui se trouve ainsi inactivé. Weisenberger et al, (2006) ont montré que la mutation BRAF est fortement associée au phénotype MSI-H sporadique (76% des CCR), CIMP-H (77% des CCR) et dans 94 % des CCR CIMP-H/MSI-H (47). Il faut noter que les mutations KRAS et BRAFV600E sont mutuellement exclusives (48).

I. A. 7. 3. Voie TGF-β/SMAD

La superfamille du TGF-β comprend différents membres dont les BMPs (Bone Morphogenic Proteins), les activines et les TGF-β. Ce sont des morphogènes avec une fonction importante au niveau de l'épithélium digestif en inhibant la prolifération (49). La liaison du TGF-β à son récepteur entraine une cascade d'évènements qui induisent la transcription de gènes régulateurs du cycle cellulaire. Dans les CCRs, on retrouve une mutation inactivatrice du gène TGFBRII dans 90% des tumeurs MSI suggérant que la perte de signalisation du TGF- β joue un rôle majeur au cours de la cancérogenèse MSI. Par contre, dans les tumeurs de phénotype CIN, ces mutations sont rares (50) mais la voie du TGF est inhibée par une perte d'hétérozygotie en 18q qui s'observe dans 70% des CCR CIN. Au niveau moléculaire, la perte d'activité de la voie de signalisation TGF-β/SMAD induit la transition épithélio-mésenchymateuse, favorable à la dissémination métastatique (51).

I. A. 7. 4. Gène suppresseur de tumeur TP53

La protéine p53, dénommée « gardienne du génome », est un facteur de transcription qui intervient dans la régulation du cycle cellulaire, de l'apoptose et la réplication de l'ADN (52).

Des mutations inactivatrices et/ou une perte allélique de ce gène suppresseur de tumeur sont retrouvées dans environ 50% des carcinomes coliques tous stades confondus.

Il faut noter que certaines mutations, telles que celles d'APC et SMAD4 par exemple, sont communes à tous les sous-groupes moléculaires - suggérant un rôle central dans le CCR en général - tandis que d'autres sont limitées à un sous-groupe (par exemple, BRAF dans les CCR de phénotype CIMP/MSI.

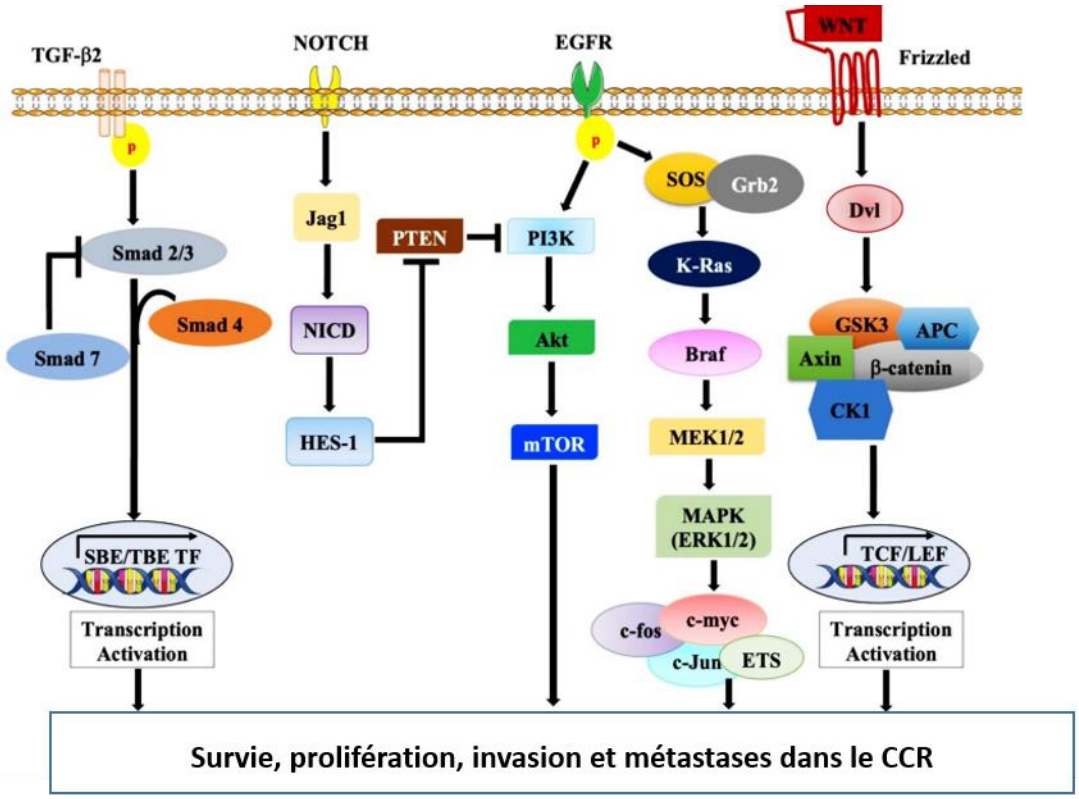


Figure 8. Représentation schématique des voies moléculaires impliquées dans la pathogenèse du CCR. D'après Malki et al, 2020 (53).

I. A. 8. Traitements des CCR

Le traitement standard du CCR est la résection chirurgicale par cœlioscopie de la tumeur primitive associée à un curage ganglionnaire d'au moins 12 ganglions.

Pour un cancer de stade I, la chirurgie seule est indiquée. Globalement, les stades I et II, de meilleur pronostic, ne subissent qu'une simple chirurgie.

Pour les tumeurs de stade III, une chimiothérapie est recommandée après la chirurgie dans le but de réduire le risque de récidive, il s'agit alors de chimiothérapie adjuvante. Enfin, pour les

CCR métastasés, donc de stade IV, la chimiothérapie est systématiquement prescrite, que ce soit avant une première opération chirurgicale afin de la faciliter (traitement néo-adjuvant), à la place de la chirurgie lorsque celle-ci n'est pas envisageable ou après la chirurgie (traitement adjuvant).

L'ensemble des thérapies ciblées actuellement recommandées par le National Comprehensive Cancer Network est illustré sur la Figure 10.

I. A. 8. 1. Les chimiothérapies conventionnelles

- La chimiothérapie adjuvante (après chirurgie) : protocoles FOLFOX 4 ou LV5FU2 en cas de risque élevé de récidive, qui correspond au stade IIb. Pour le stade III, le traitement de référence est le FOLFOX 4 (Oxaliplatine + LV5FU2 qui est la combinaison d'acide folinique (ou leucovérine) et de 5-fluorouracile (5-FU) ou le CAPOX (ou XELOX : capécitabine et oxaliplatine).
- La chimiothérapie néoadjuvante (avant chirurgie) : dans le cas de présence de métastases, elle a pour but de réduire le volume tumoral et de tester la sensibilité de la tumeur et des métastases à la chimiothérapie. Le protocole de référence est le FOLFOX 4, seul ou associé à des thérapies ciblées qui sont des médicaments dirigés contre des cibles moléculaires spécifiques (récepteurs, gènes ou protéines jouant un rôle dans la cancérogenèse).
- Pour un CCR de stade 4 ou une récidive, les protocoles peuvent être :
    - Folfiri – leucovorine (qui réduit la toxicité du traitement), 5-fluorouracil et irinotécan
    - Folfox – leucovorine, 5-fluorouracil et oxaliplatine
    - Folfoxiri – leucovorine, 5-fluorouracil, oxaliplatine et irinotécan
    - CAPOX (XELOX) – capécitabine et oxaliplatine
    - CAPIRI – capécitabine et irinotécan

Les résultats des traitements thérapeutiques se sont améliorés avec l'avènement des inhibiteurs à petites molécules et de points de contrôle immunitaires. Cependant, de nombreux patients ne répondent pas avec ces molécules en monothérapie. Par conséquent, ces thérapies sont utilisées en combinaison avec la chimiothérapie conventionnelle pour augmenter l'efficacité clinique par des effets synergiques potentiels.

I. A. 8. 2. Les thérapies ciblées

Elles ciblent une molécule spécifique dans la cellule, plutôt que de cibler la prolifération ou la survie d'un ensemble de cellules. Parmi ces thérapies on retrouve :

- Le cétuximab (Erbitux), un anticorps monoclonal chimérique (humain à 60%) qui inhibe la voie de signalisation de l'EGFR; il est préconisé pour le traitement des CCR ayant le gène KRAS sauvage uniquement. Il peut être associé au Folfox ou Capox. Il agit sur la survie et la prolifération.
- Le panitumumab (Vectibix), également un anticorps monoclonal (100% humain) dirigé contre les récepteurs de l'EGF – Comme le cétuximab, il est indiqué pour les CCR sans mutation du gène KRAS (54,55).
- Le bévacizumab (Avastin), anticorps monoclonal (90% humain) dirigé contre les récepteurs du VEGF est habituellement associé au Folfiri, au Folfox ou au CAPOX. Il bloque l'angiogenèse.
- Le régorafénib (Stivarga) est un inhibiteur de tyrosine kinases non spécifique. Il est utilisé pour les cas de CCR métastatiques réfractaires aux autres traitements.
- D'autre part, l'encorafénib, un inhibiteur de BRAF utilisé en association avec le cétuximab, est un traitement de 2e ligne et plus particulièrement pour le CCR métastatique porteur d'une mutation BRAF V600E.

I. A. 8. 3. L'immunothérapie

La base de l'immunothérapie consiste à surmonter les mécanismes qui interviennent dans la tolérance immunitaire aux auto-antigènes tumoraux et à bloquer la réponse immunosuppressive qui se produit dans le microenvironnement tumoral. En effet, l'échappement des cellules tumorales à la réponse de l'hôte caractérise plusieurs types de cancers. Ce processus a comme principale origine l'inactivation et l'épuisement des cellules lymphocytaires T via l'activation des récepteurs de points de contrôles immunitaires à la surface des cellules T : le récepteur PD-1 et son ligand PD-L1, et l'antigène 4 des lymphocytes T cytotoxiques (CTLA-4); (56); (Figure 9).

Avec l'immunothérapie, il s'agit d'améliorer la surveillance immunitaire et de bloquer la tentative de la tumeur d'échapper à sa détection par les lymphocytes T en utilisant des inhibiteurs des « checkpoints immunitaires ». Plusieurs molécules ont été approuvées pour traiter entre autres, le mélanome ou le cancer bronchique, comme le pembrolizumab (90% humain) ou le nivolumab (100% humain), des anticorps inhibiteurs de PD-1.

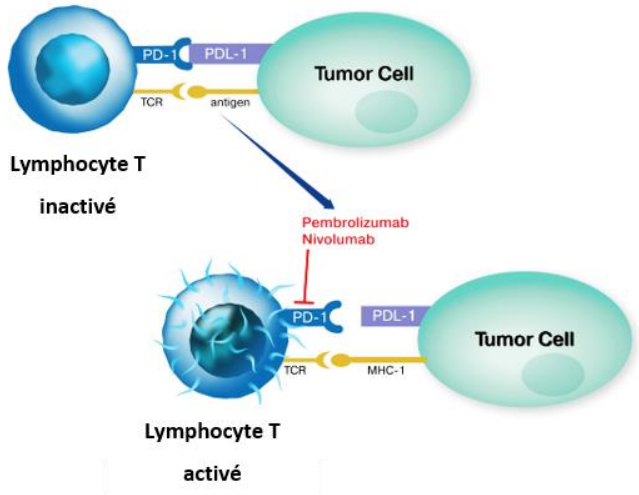


Figure 9. Inhibition de PD-1 par le pembrolizumab et le nivolumab. PD-1 : protéine de mort cellulaire programmée 1 ; PD-L1, ligand de PD-1 ; TCR, récepteur des lymphocytes T. D'après Gianchecchi et al, 2013(56).

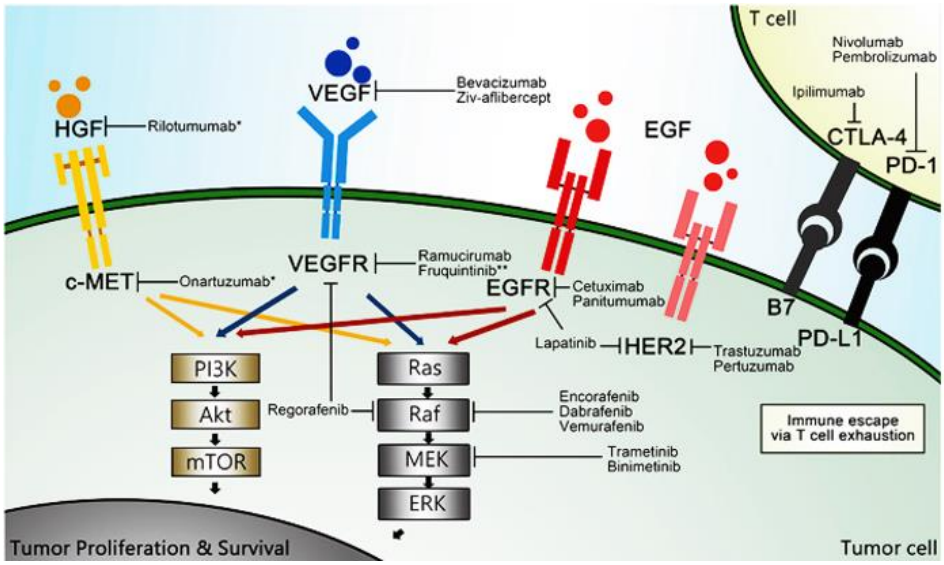


Figure 10. Aperçu des agents ciblés recommandés par le National Comprehensive Cancer Network. D'après Xie et al, 2020 (57).

Le CCR MSI-H représente une très petite minorité de cas et de nombreux essais cliniques en cours explorent les effets synergiques potentiels de la combinaison de la chimiothérapie ou de la thérapie ciblée avec l'immunothérapie chez les patients ayant un CCR MSI. Bien que les études soient encore en développement, la classification CMS pourrait guider le développement et l'application de thérapies puisque les patients dans les quatre classifications CMS pourraient avoir différents mécanismes d'évasion immunitaire, ce qui permettrait de proposer une thérapie ciblée sur mesure.

I. A. 9. Facteurs prédictifs de réponse au traitement

Les marqueurs pronostiques sont utiles pour évaluer le risque de récidive ou de décès d'un patient. Cependant, ils ne prédisent en rien la sensibilité de la tumeur à un traitement donné. Dans la pratique, il est fréquent que plusieurs options thérapeutiques soient possibles, telles que des chimiothérapies ou des thérapies ciblées. La présence d'un/de marqueur(s) moléculaire(s) permettrait d'orienter la prescription médicale pour une meilleure efficacité thérapeutique.

I. A. 9. 1. Marqueurs moléculaires

Le rôle de KRAS en tant que biomarqueur prédictif de la réponse à une thérapie anti-EGFR est maintenant bien reconnu puisque les patients avec un CCR muté pour le gène KRAS ne tirent aucun bénéfice du cétuximab puisqu'ils ont une survie globale plus faible par rapport aux patients KRAS de type sauvage (58). En effet, les résultats de l'étude OPUS ont montré un effet délétère des mutations de KRAS sur l'association FOLFOX plus cétuximab, la survie sans progression étant de 5,5 mois avec le traitement combiné contre 8,6 mois avec le FOLFOX seul (59).

Comme pour les mutations de KRAS, les mutations de la protéine BRAF, située en aval de KRAS confèrent une résistance aux thérapies anti-EGFR (cétuximab ou panitumumab) dans le groupe de patients avec le gène KRAS sauvage (60), les 2 mutations étant mutuellement exclusives (61).

Le 5-fluorouracile (5-FU) est largement utilisé depuis la fin des années 1950, et reste le premier choix d'agent chimio-thérapeutique pour les patients atteints de CCR à un stade avancé. Webber et al. ont mené une méta-analyse sur 16 études, pour évaluer la valeur prédictive du phénotype MSI sur la réponse au 5-FU de patients atteints de CCR de tous les stades (62). Ils ont observé un effet bénéfique de la chimiothérapie à base de 5-FU chez les patients ayant des tumeurs MSS alors que l'effet du traitement n'est pas statistiquement significatif pour les CCR MSI-H (62). Si le statut MSI comme marqueur de bon pronostic est plus certain que son statut prédictif, le rationnel moléculaire sous-jacent n'est pas encore expliqué.

I. A. 9. 2. Classification CMS et localisation

L'impact prédictif de la classification CMS a été suggéré par les données exploratoires de 2 essais cliniques qui indiquent que pour les tumeurs CMS-1, localisées préférentiellement à droite, l'association Folfiri + bévacizumab serait meilleure en termes de survie globale et de

survie sans progression que l'association Folfiri + cétuximab (42). Pour les tumeurs CMS-2 et 4, localisées préférentiellement dans le côlon gauche, il existe un effet favorable du cétuximab par rapport au bévacizumab en termes de survie globale et de survie sans progression (Figure 11).

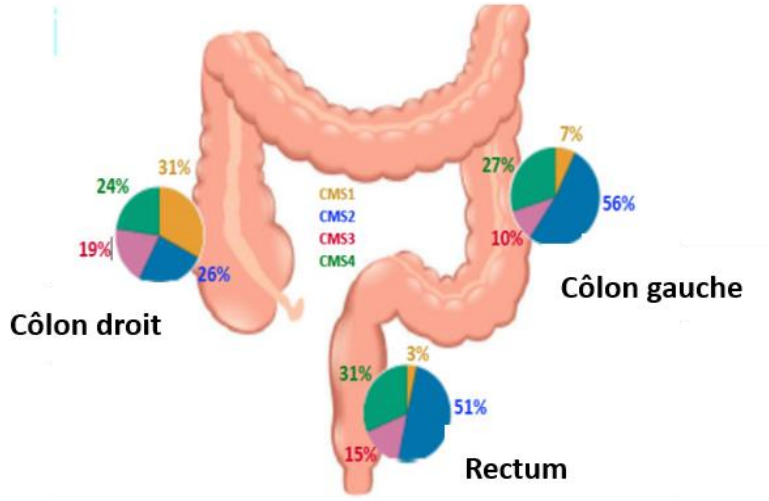


Figure 11. Répartition des groupes CMS sur différents sites tumoraux. (n = 2 651). D'après Guinney et al, 2015 (41).

I. A. 9. 3. Tumeurs immunogènes

L'efficacité de l'immunothérapie a été démontrée pour traiter des tumeurs solides auparavant difficiles à traiter, comme le mélanome et le cancer du poumon. Une charge de mutations élevée dans une tumeur est apparue comme un facteur de réponse à l'immunothérapie dans plusieurs cancers (63). Une hypothèse de cette efficacité serait que dans ces cancers, le niveau élevé de néo-antigènes produits par le grand nombre de mutations favoriserait une infiltration accrue des lymphocytes T (64); en effet, les tumeurs MSI ont une importante infiltration immunitaire (33). Malgré tout, l'efficacité est limitée puisque seulement 3 % à 7 % des patients avec CCR MSI-H métastatiques répondent à ces traitements (65).

Par contre, ces thérapies sont inefficaces pour les CCR CIN car dans ces tumeurs, la faible charge de mutations et l'absence d'infiltration de cellules immunitaires sont proposées comme des mécanismes de résistance immunitaire.

I. A. 10. Techniques d'évaluation de la réponse au traitement

En fonction du type de traitement, le seuil significatif pour prédire la réponse/progression chez les patients varie selon les traitements. Par exemple avec les thérapies ciblées (anti-VEGF ou

anti-EGFR) n’induisent souvent qu’un faible changement de taille, tandis que la survie des patients est nettement améliorée (66). L’imagerie fonctionnelle multiparamétrique, comme la densitométrie (CT scan), l’IRM pondérée en T2 ou l’IRM à contraste dynamique (imagerie dynamique rehaussée par un agent de contraste) pourraient mieux caractériser la tumeur résiduelle, et prédire la réponse thérapeutique tôt après le début de traitement (67). Cependant, parfois il est difficile de faire la distinction entre une fibrose induite par le traitement et la tumeur résiduelle sur la base des images d’IRM pondérées en T2. Aussi, des techniques d'imagerie fonctionnelle et moléculaire, telles que la tomographie par émission de positons (TEP) utilisant le fluor 18-fluorodésoxyglucose (FDG, analogue du glucose) est plus simple (68). La TEP au FDG est une technique d'imagerie moléculaire qui visualise et quantifie les processus métaboliques dans les cellules cancéreuses. Par cette technique, les changements métaboliques indicatifs d'une réponse tumorale, peuvent survenir avant tout changement structurellement détectable comme la taille de la tumeur. Elle est très utile dans l'évaluation et la surveillance de la réponse métabolique tumorale des CCR traités par une thérapie ciblée (69–71).

Une autre approche est la spectroscopie infrarouge, méthode analytique qui permet d'analyser avec précision la composition chimique d'un échantillon avec peu d’exigence de préparation. L’effet de traitement anti-cancéreux sur le tissu ou des cellules tumorales s'accompagne de modifications de la composition biochimique globale du tissu, ainsi que de modifications morphologiques cellulaires codées et observées à travers des changements dans le spectre infrarouge. C’est dans cette optique que s’inscrit ce travail de thèse, évaluant le potentiel prédictif de la spectroscopie infrarouge combinée à la chimiométrie pour identifier des caractéristiques spectrales robustes d’une tumeur, afin d’identifier celles dont les changements moléculaires seraient prédictifs d’une bonne/mauvaise réponse au traitement.

I. B. Spectroscopie infra-rouge

I. B. 1. Interaction rayonnement – matière et spectroscopie infrarouge

La spectroscopie optique consiste à analyser l'interaction entre la lumière et la matière en mesurant la composition du rayonnement à différentes fréquences (ou énergies).

I. B. 1. 1. Rayonnement et spectre électromagnétiques

Le rayonnement électromagnétique permet la propagation d'énergie grâce aux variations périodiques, de fréquence ν, d'un champ électrique et d'un champ magnétique. Le rayonnement est dit monochromatique s'il correspond à une radiation de fréquence ν précisément définie, ou polychromatique s'il est composé de radiations de fréquences diverses.

Selon la mécanique quantique, le rayonnement électromagnétique présente une double nature : ondulatoire et corpusculaire. Il peut être ainsi considéré comme une onde, l'onde électromagnétique, ou un flux de particules de masse nulle, les photons. Une onde peut être caractérisée par sa fréquence ν ou, de façon inversement proportionnelle, par sa période temporelle T_p. La fréquence présente l'intérêt d'être directement proportionnelle à l'énergie. Par ailleurs, la longueur d'onde λ correspond à la périodicité spatiale.

Dans le vide, ces grandeurs physiques sont liées par la relation :

$$\lambda = c/\nu = T_p c \qquad \text{Éq (1)}$$

avec c (3×10^8 m.s^{-1}) étant la célérité de la lumière dans le vide.

On définit aussi le nombre d'onde σ, généralement exprimé en cm^{-1}, par la relation :

$$\sigma = 1/\lambda \qquad \text{Éq (2)}$$

avec λ en cm.

Selon la relation de Planck-Einstein pour le modèle corpusculaire des ondes électromagnétiques, l'énergie transportée par un photon est proportionnelle à la fréquence selon la relation :

$$E = h\nu \qquad \text{Éq (3)}$$

avec h étant la constante de Planck ($6{,}63.10^{-34}$ J.s).

Ainsi, il est possible de classer les différentes régions du spectre électromagnétique en fonction de l'énergie transportée (Figure 12). Le domaine infrarouge (IR), localisé entre la région du visible et la région des micro-ondes, s'étend de 0,75 µm à 1000 µm en termes de longueur d'onde ou de 10 à 13000 cm^{-1} en nombre d'onde. Il se décompose lui-même en domaines proche IR (13000- 4000 cm^{-1}), IR moyen (4000-400 cm^{-1}) et IR lointain (400-10 cm^{-1}).

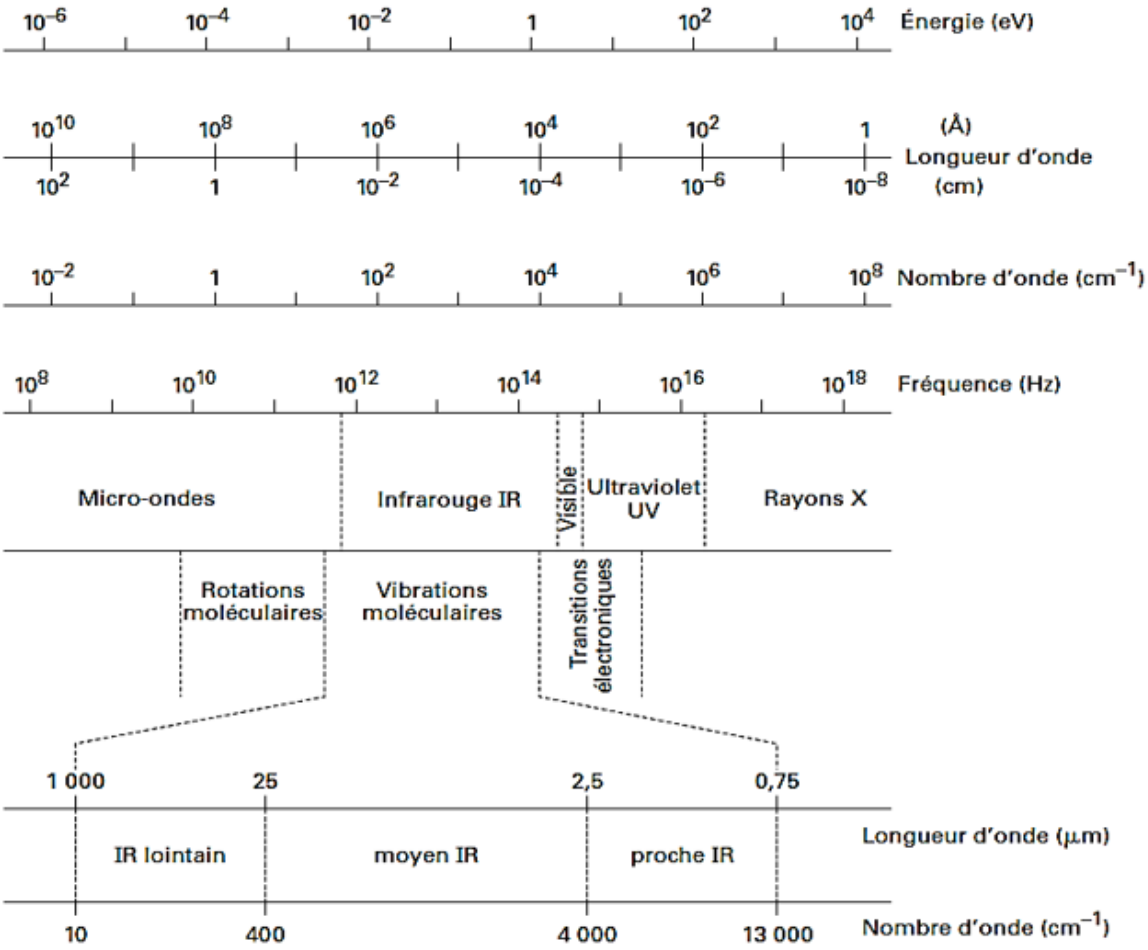


Figure 12. Les divers domaines spectraux du rayonnement électromagnétique (72).

I. B. 1. 2. Interaction matière-rayonnement

Une onde électromagnétique, véhiculant une énergie liée à sa fréquence, interagissant avec une molécule peut échanger de l'énergie avec cette dernière. On peut distinguer trois phénomènes principaux à la suite d'une telle interaction (Figure 13) :

- Une absorption du photon incident si celui-ci est porteur d'une énergie correspondant à la différence d'énergie de deux niveaux énergétiques. La molécule sera alors portée à un état excité.
- Une émission radiative consécutive à une absorption, correspondant au retour de la molécule à un état fondamental. C'est le cas de l'émission de fluorescence.
- Une diffusion, c'est-à-dire la déviation dans diverses directions de l'onde électromagnétique. Elle peut se faire avec ou sans changement de fréquence c'est-à-dire de façon inélastique ou élastique.

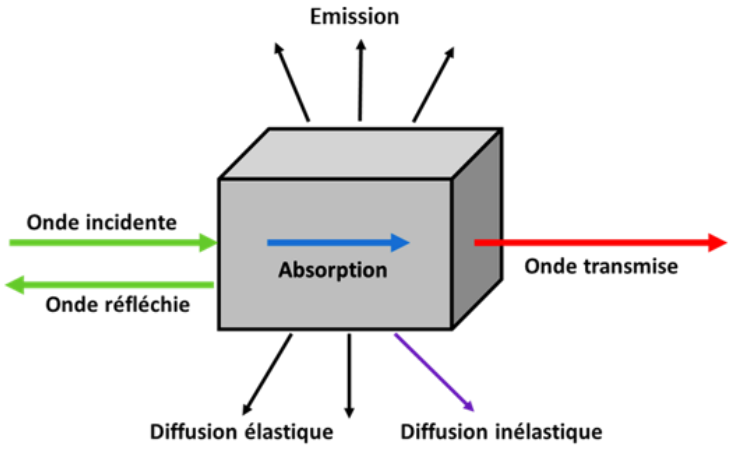


Figure 13. Représentation schématique des différents phénomènes optiques émanant d'une interaction lumière/matière.

I. B. 1. 3. Niveaux d'énergie d'une molécule

D'après la mécanique quantique, l'énergie d'une molécule E_m, à l'exception de son énergie cinétique, est quantifiée : elle dépend de nombres entiers, nommés nombres quantiques. E_m correspond à la somme de l'énergie des électrons E_e (liée à leur nombre, la forme de la molécule et leur état d'excitation), de l'énergie vibrationnelle des liaisons interatomiques E_v (liée à la masse et à l'arrangement des atomes), et de l'énergie E_r due à la rotation de la molécule (dans le cas des gaz) :

$$E_m = E_e + E_v + E_r \qquad \text{Éq (4)}$$

avec $E_e > E_v > E_r$.

Schématiquement, la Figure 14 montre les gammes d'énergie correspondant à ces différentes contributions et leur correspondance avec l'énergie des domaines du spectre électromagnétique.

Les niveaux vibrationnels correspondent à la gamme d'énergie du domaine spectral du moyen infrarouge.

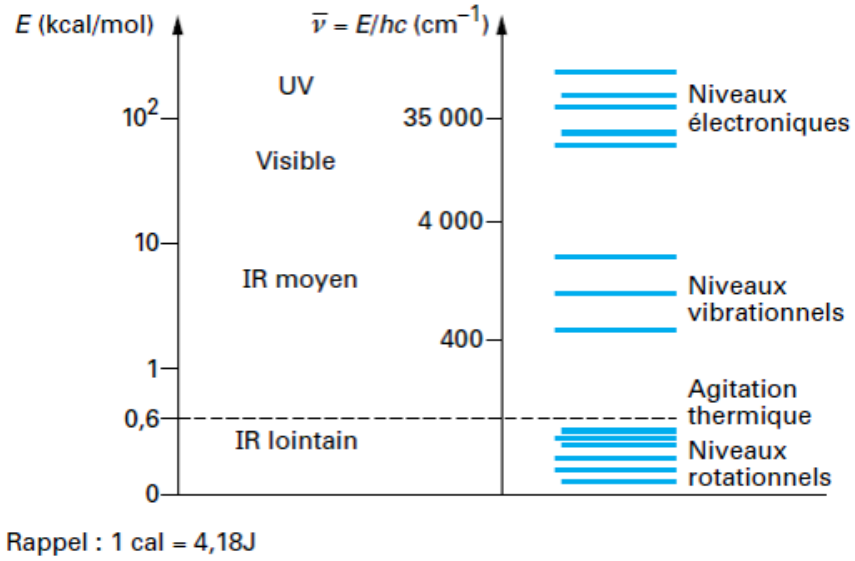


Figure 14. Valeurs respectives des contributions électroniques, vibrationnelles et rotationnelles d'une molécule (72).

I. B. 1. 4. Modes de vibration moléculaire

Une molécule formée de N atomes possède 3N degrés de liberté, soit 3N mouvements possibles dans l'espace. Parmi eux, trois représentent la translation de la molécule dans son ensemble (le long des trois axes du repère x, y, z) et trois autres définissent la rotation de la molécule autour de chacun de ces axes (73). Les autres possibilités de mouvements (soit 3N-6 ou 3N–5 si la molécule est linéaire) sont associées aux mouvements internes de vibration d'une molécule.

Deux modes principaux de vibrations moléculaires sont à distinguer (Figure 15) :

- Les modes d'élongation ou de valence qui entraînent une variation de longueur des liaisons interatomiques, sans modification des angles formés par ces liaisons.
- Les modes de déformation qui, à l'inverse, entraînent une modification des angles formés par les liaisons sans modification des longueurs de ces liaisons.

Pour chaque mode, dans le cas de groupements chimiques constitués d'au moins deux liaisons moléculaires identiques, on distinguera les vibrations symétriques et asymétriques, selon que les mouvements se font en phase ou en opposition de phase entre chaque liaison.

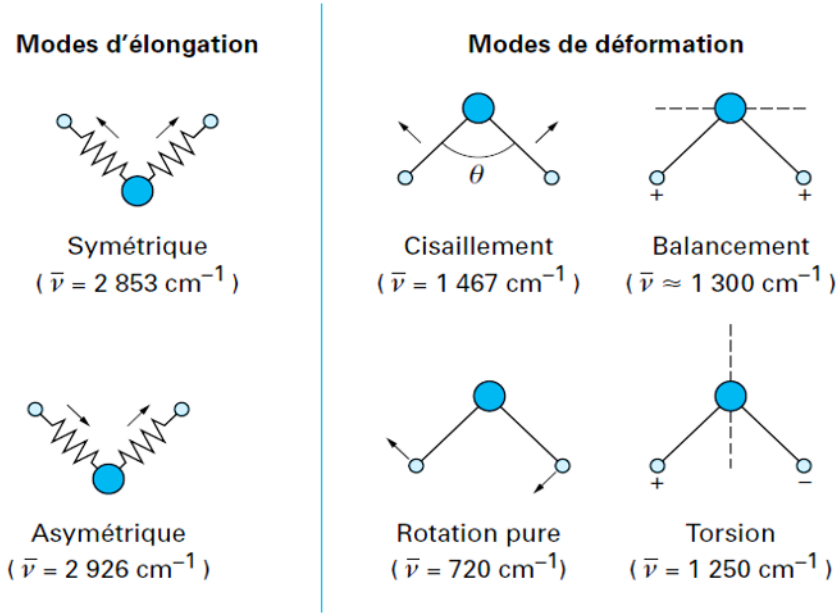


Figure 15. Modes de vibration : exemple de vibrations au niveau du groupement CH_2 d'une molécule (72).

I. B. 1. 5. Spectroscopies vibrationnelles

On distingue différents types de spectroscopies optiques selon la correspondance des énergies du rayonnement et de la molécule (Figure 14) : la spectroscopie d'absorption UV-visible, la spectroscopie d'émission de fluorescence et la spectroscopie vibrationnelle (spectroscopies infrarouge et Raman).

Alors que les spectroscopies d'absorption UV-visible et d'émission de fluorescence analysent les transitions entre niveaux énergétiques électroniques (rayonnement de longueur d'onde dans les domaines du visible et de l'ultraviolet), la spectroscopie vibrationnelle analyse des transitions entre niveaux énergétiques vibrationnels (rayonnement des domaines de l'infrarouge). La spectroscopie vibrationnelle permet alors de déterminer la composition chimique d'une substance. De plus, elle peut donner accès à des informations sur l'organisation structurelle de molécules simples et de systèmes plus complexes tels que les cellules, les tissus ou les biofluides (74–76).

Les techniques de spectroscopie vibrationnelle incluent la spectroscopie d'absorption infrarouge (IR) et la spectroscopie de diffusion Raman. Ces deux techniques sont complémentaires entre elles dans la mesure où certains modes de vibrations sont seulement actifs en Raman, d'autres uniquement en infrarouge. De plus, la forte absorption de l'eau dans l'infrarouge interdit l'analyse d'échantillons hydratés en spectroscopie infrarouge

contrairement à l'analyse Raman. Au niveau instrumental, chaque modalité nécessite des systèmes spécifiques notamment du fait des propriétés des optiques en verre, fortement absorbantes dans l'infrarouge et transparentes dans le visible (domaine utilisé en spectroscopie Raman).

I. B. 2. Principe de la spectroscopie infrarouge

Les domaines électromagnétiques concernés en spectroscopie vibrationnelle sont le « moyen infrarouge » (4000-400 cm^{-1}) et le proche IR (13000-4000 cm^{-1}), qui correspondent respectivement à des vibrations des modes fondamentaux tels que l'élongation ou la déformation (Figure 12) et à des harmoniques ou des combinaisons des modes fondamentaux. Par la suite, nous aborderons exclusivement la région du moyen infrarouge.

I. B. 2. 1. Approches physiques du phénomène d'absorption infrarouge

I. B. 2. 1. 1. Approche classique

Un rayonnement électromagnétique peut se coupler avec tout mouvement moléculaire faisant intervenir une variation de la liaison moléculaire.

Ainsi certains mouvements de vibration d'une molécule peuvent être mis en résonnance par une onde électromagnétique de même fréquence et donner lieu à un phénomène d'absorption de ce rayonnement. Pour que ce phénomène ait lieu, il faut que cette absorption induise une variation du moment dipolaire μ de la molécule.

Cette règle de sélection est associée aux propriétés de symétrie d'une liaison moléculaire. Cette règle est à la base de la différence entre la spectroscopie infrarouge et la spectroscopie Raman dans la mesure où la diffusion Raman ne s'observe que lorsque le mouvement de vibration entraine une variation de la polarisabilité moléculaire (73).

I. B. 2. 1. 2. Approche quantique

Dans cette approche, l'énergie de vibration moléculaire est quantifiée en niveaux discrets. A chaque niveau ν est associée une probabilité de distribution d'énergie autour de ce quantum. En cas d'interaction entre un photon et une molécule, différents phénomènes peuvent être engendrés (Figure 16).

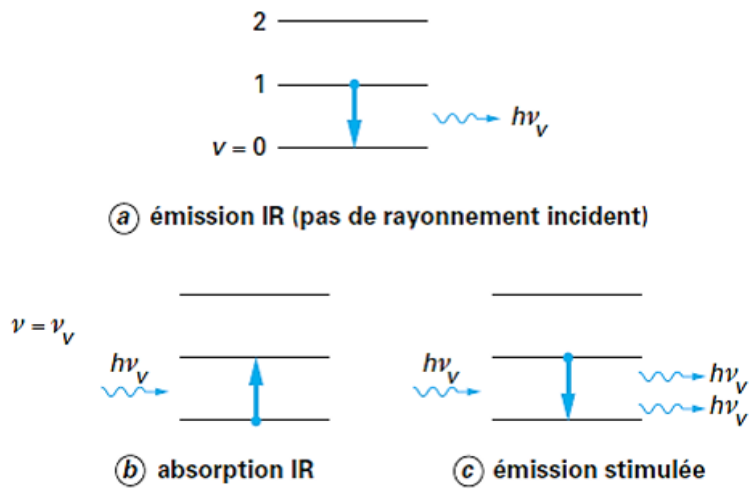


Figure 16. Interaction entre un photon et la matière caractérisée par des niveaux d'énergie vibrationnelle.

Si la fréquence de l'onde électromagnétique incidente ν est de l'ordre de grandeur des fréquences de vibration moléculaire, la transition la plus probable est $\nu = \nu + 1$. C'est le phénomène d'absorption infrarouge (illustration b de la Figure 16). A noter également la possibilité de $\nu = \nu - 1$ correspondant au phénomène d'émission stimulée à la base du rayonnement laser (illustration c de la Figure 16). L'émission infrarouge sans rayonnement incident (illustration a de la Figure 16) correspond à l'émission thermique avec transformation d'énergie thermique en radiation ; lorsque le corps est à température ambiante, l'émission se fait dans le domaine du moyen infrarouge.

I. B. 2. 2. Spectre et absorption infrarouge

Lorsque l'énergie du rayonnement lumineux traversant l'échantillon est égale à l'énergie de vibration d'une liaison moléculaire, une absorption du rayonnement a lieu engendrant une diminution de son intensité. Un spectre infrarouge présente donc des valeurs d'absorbance (ou de transmittance) en fonction de nombres d'onde, unité proportionnelle à l'énergie du rayonnement. Les valeurs d'absorbance et de transmittance obéissent à la loi de Beer-Lambert suivant l'équation suivante :

$$I = I_0 e^{-\varepsilon c l} \qquad \text{Éq (5)}$$

avec : I l'intensité du rayonnement transmis, I_0 l'intensité du rayonnement incident, ε le coefficient d'extinction moléculaire de la substance absorbante à la fréquence considérée, c la concentration du milieu en substance absorbante et l la longueur du trajet optique.

La transmittance T_{IR} est définie par :

$$T_{IR} = I/I_0 \qquad \text{Éq (6)}$$

Connaissant la transmittance, l'absorbance peut être déduite suivant la formule :

$$A = \log_{10}(I_0/I) \qquad \text{Éq (7)}$$

Un spectre infrarouge apporte une information qualitative quant à la composition moléculaire de l'échantillon analysé puisque chaque bande du spectre correspond à un groupe de vibrations caractéristiques des molécules de l'échantillon, comme l'indique la Figure 17 pour un échantillon biologique (77). La loi de Beer-Lambert apporte également une information quantitative puisque l'absorbance, en un nombre d'onde donné, est proportionnelle à la concentration des molécules absorbantes. Cependant, cette proportionnalité avec la concentration est limitée à des valeurs d'absorbance situées en général entre 0,3 et 2 par rapport à la sensibilité de détection des systèmes.

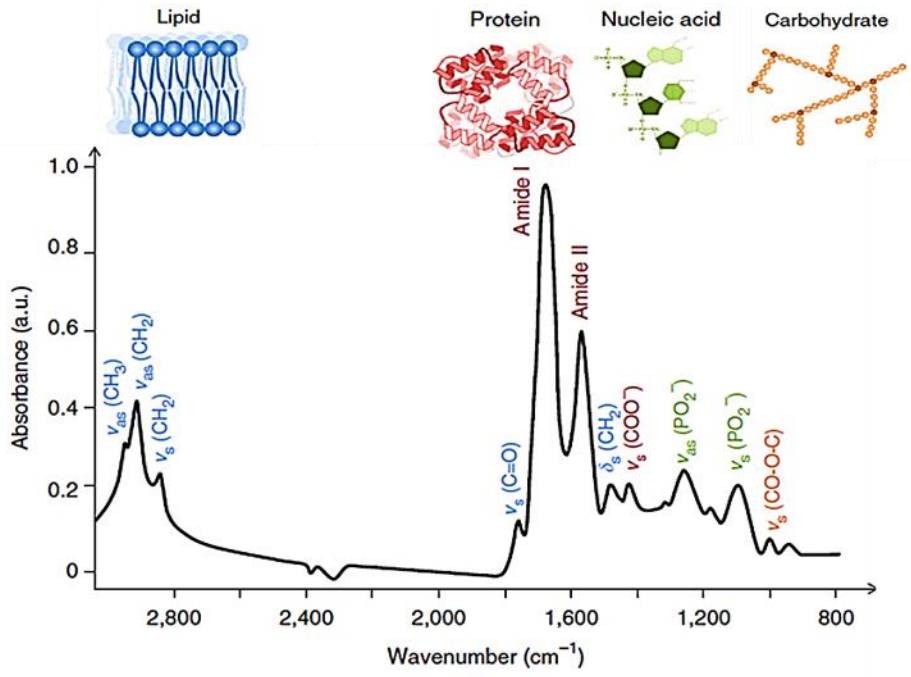


Figure 17. Spectre infrarouge d'un échantillon biologique avec des exemples d'attribution biomoléculaire de bandes de 800 à 3000 cm^{-1} (78).

I. B. 3. Instrumentation et imagerie spectrale IR

La mesure d'un spectre d'absorption moyen-infrarouge est traditionnellement réalisée par un spectromètre infrarouge à transformée de Fourier (FTIR) : instrumentation comportant cinq éléments importants (Figure 18) :

- une source polychromatique dans le moyen IR,
- un interféromètre (de Michelson) comprenant les miroirs fixe et mobile,
- le compartiment de mesure de l'échantillon soit en transmission sur un support transparent (KBr, ZnSe, CaF2, silicium), soit en réflexion sur un support réfléchissant (aluminium, diamant, Kevley),
- un détecteur photosensible soit pyroélectrique, soit photoélectrique,
- un convertisseur analogique/numérique pour convertir le signal analogique en signal numérique exploitable par un ordinateur.

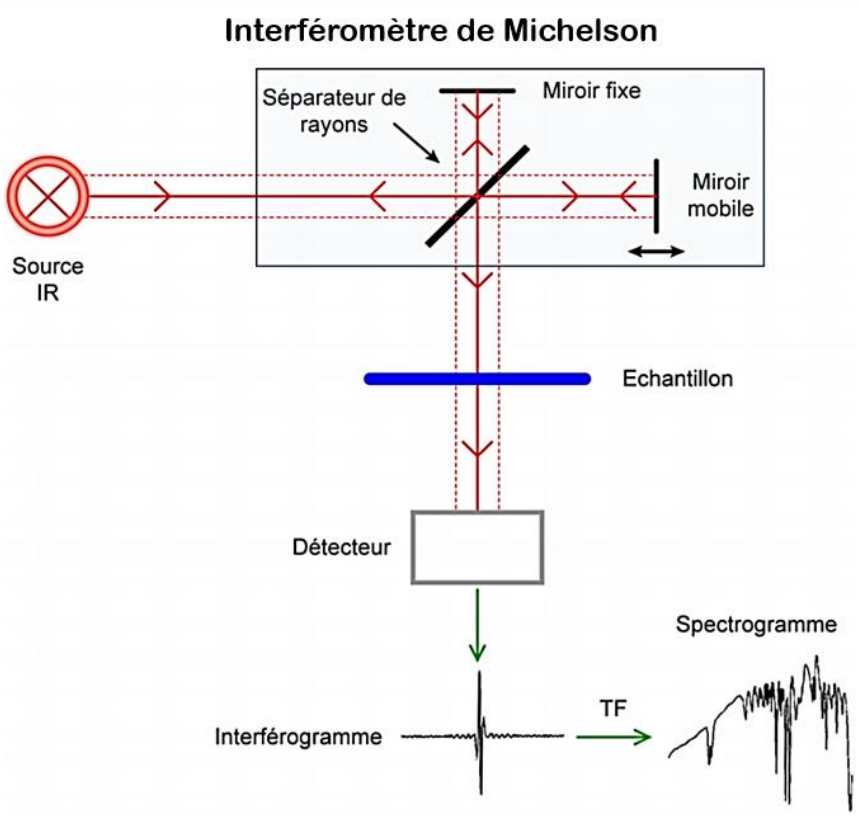


Figure 18. Schéma instrumental d'un spectromètre infrarouge à transformée de Fourier (FTIR).

La source de rayonnement infrarouge (tel le Globar) est obtenue en portant à haute température (1500°C) un bâtonnet en carbure de silicium (SiC).

L'interféromètre de Michelson est un dispositif permettant de générer des interférences entre les 2 faisceaux provenant respectivement du miroir fixe et du miroir mobile. Par ce procédé, il devient possible d'analyser les différentes longueurs d'onde de la source polychromatique simultanément.

Cet interféromètre comprend une lame séparatrice semi-transparente divisant le rayon incident en deux faisceaux d'égale intensité. Le premier faisceau est réfléchi vers un miroir fixe, alors que le second est transmis vers un miroir mobile.

Après réflexion, les deux demi-faisceaux se recombinent au niveau de la lame séparatrice. Selon le déplacement du miroir mobile, et donc la différence de chemin optique parcourue entre les deux demi-faisceaux, il va se former des interférences constructives ou destructives. Le faisceau reconstitué arrive sur un détecteur MCT (Mercury Cadmium Telluride) où l'ensemble des interférences sont traduites sous la forme d'un interférogramme (intensité du faisceau en fonction du déplacement du miroir mobile).

La transformée de Fourier (algorithme mathématique, TF) permet de passer d'une représentation spatiale (interférogramme qui est fonction de la distance du miroir mobile) à une représentation dans le domaine fréquentiel (spectre qui est fonction du nombre d'onde). Lors d'une acquisition, deux interférogrammes sont enregistrés, avec et sans échantillon. La différence des deux spectres ainsi obtenus génère un spectre d'absorption infrarouge de l'échantillon.

Pour l'imagerie tissulaire à l'échelle microscopique, le spectromètre infrarouge est couplé à un microscope optimisé pour la gamme spectrale infrarouge. Ce microscope est équipé d'un objectif de type Cassegrain fonctionnant sur le principe d'un télescope, c'est-à-dire par réflexions successives du rayonnement sur des miroirs concaves et convexes. En général, ces objectifs présentent une ouverture numérique (NA) comprise entre 0,5 et 0,7. L'ouverture numérique et la longueur d'onde λ du rayonnement déterminent la résolution spatiale de la mesure, c'est-à-dire la capacité de l'instrument de mesure à séparer deux objets voisins. En effet, cette limite de résolution est donnée par le critère de Rayleigh (79,80).

La distance angulaire $\Delta\theta$ entre deux objets doit vérifier :

$$\Delta\theta > 1.22\frac{\lambda}{\mathrm{D.\,NA}} \qquad \text{Éq (8)}$$

pour une ouverture circulaire de diamètre D.

Dans le cas du rayonnement moyen-infrarouge, la longueur d'onde est de l'ordre de 10 µm, ce qui induit une résolution spatiale de même ordre de grandeur. Dans notre étude, nous avons utilisé un micro-imageur infrarouge, en mode transmission, équipé d'un détecteur multicanaux dont chaque élément correspond à une taille de pixel nominale 6,25x6,25 μm^2 au niveau de l'échantillon.

La résolution spatiale peut être améliorée en utilisant une mesure en réflexion totale atténuée (ATR) (81–84). En effet, cette technique repose sur l'utilisation d'un cristal d'indice de réfraction optique (n) élevé, ce qui induit à une grande ouverture numérique NA, et donc une meilleure résolution spatiale. Cet élément ATR peut être monté sur le système de micro-imagerie infrarouge mais nous ne l'avons pas employé pour nos analyses.

Une image spectrale IR constitue donc un cube de données hyperspectrales (84) composé de deux dimensions spatiales (x, y) correspondant aux pixels de l'image et une dimension spectrale représentée par l'intensité d'absorption infrarouge à chaque la longueur d'onde (λ) du spectre (Figure 19). Appliquée à l'analyse des tissus, la résolution spatiale est suffisante pour accéder à des données à l'échelle cellulaire. Cette imagerie est potentiellement riche en informations biochimiques spécifiques du statut physiopathologique de l'échantillon tissulaire analysé. Ainsi, elle vient compléter l'analyse histologique conventionnelle qui repose sur l'utilisation de colorants ou sondes fluorescentes pour caractériser la morphologie des structures tissulaires et détecter des molécules d'intérêt.

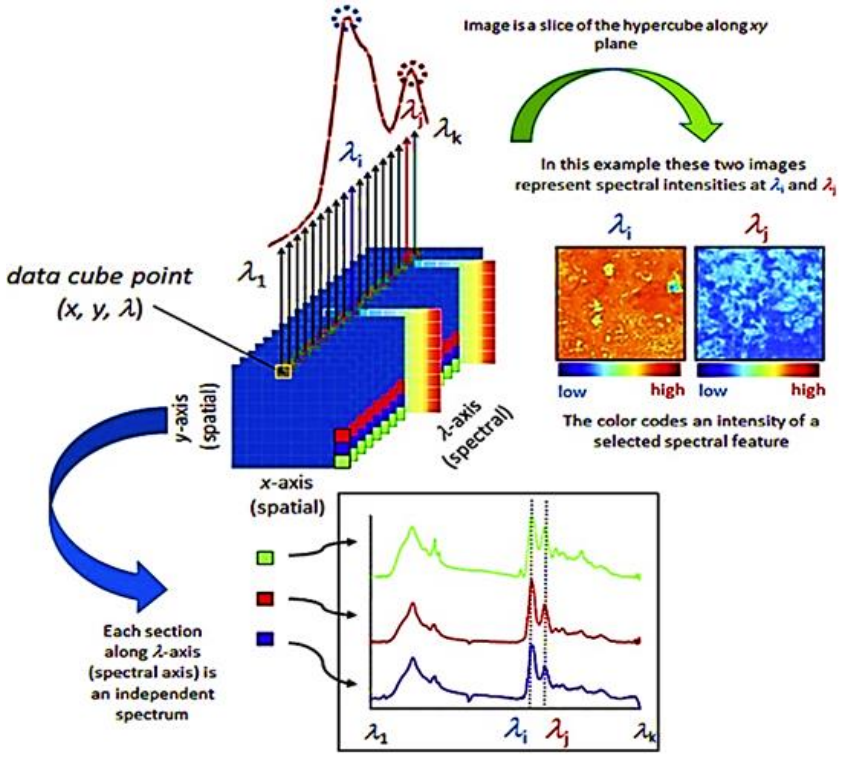


Figure 19. Les différents modes de mesure en imagerie spectrale infrarouge (84).

I. C. Chimiométrie en spectroscopie IR

Afin d'exploiter la richesse des données issues de la spectroscopie IR, il est nécessaire de leur appliquer des méthodes d'analyses statistiques multivariées, également appelées méthodes chimiométriques. Les outils de chimiométrie classiquement utilisés en spectroscopie infrarouge sont généralement appliqués en trois étapes : en premier lieu, un prétraitement des données spectrales acquises, puis une réduction de données par sélection de variables, et enfin une classification statistique.

I. C. 1. Prétraitement des données spectrales

Comme le dit la célèbre expression anglaise « garbage in, garbage out », des données de mauvaise qualité vont engendrer des résultats de mauvaise qualité en sortie d'un algorithme. Or, les données spectrales brutes ne contiennent pas que des informations utiles. Par conséquent, une étape de prétraitement des données est donc nécessaire pour extraire le signal originaire de l'échantillon analysé en atténuant les effets de phénomènes indésirables, tels que la variabilité de l'épaisseur de l'échantillon, la diffusion de la lumière, et les interférences spectrales dues par exemple au milieu d'inclusion, dans le cas d'analyse d'échantillons biologiques, qui peuvent altérer la qualité des informations d'intérêt. Les méthodes de prétraitement des spectres FTIR peuvent être divisées en deux catégories : les méthodes de filtrage et les méthodes basées sur des modèles (81).

Les méthodes de filtrage sont, comme leur nom l'indique, basées sur l'application de filtres spécialement conçus pour une tâche. Par exemple, le calcul des spectres dérivés permet de corriger une partie de la ligne de base due à la diffusion de la lumière mais également d'améliorer la résolution des bandes infrarouges, c'est-à-dire de résoudre le problème du recouvrement des bandes spectrales IR. Dans la pratique, une dérivée seconde est généralement appliquée aux spectres IR. Cependant, cette opération de dérivation des spectres s'accompagne d'une augmentation du bruit. Pour atténuer le bruit, un filtre de Savitzky-Golay (SG) est généralement appliqué car sa formulation mathématique permet lisser les spectres tout en les dérivant et en préservant les caractéristiques principales des bandes infrarouges (positions des maximums et largeurs à mi-hauteur).

Les méthodes de prétraitement basées sur des modèles consistent à quantifier et séparer dans les spectres les contributions et composantes physiques de celles chimiques tout en cherchant à conserver au maximum l'information contenue dans le spectre (81). Le premier modèle utilisé en spectroscopie infrarouge est la Multiplicative Scatter Correction (MSC) développée pour corriger les spectres IR de la diffusion multiplicative de la lumière (85). Ce modèle a ensuite été généralisé sous le nom Extended Multiplicative Signal Correction (EMSC) (86) afin de corriger d'autres effets indésirables présents dans les spectres infrarouge, tels que la ligne de base en incluant un polynôme dans le modèle. Ce modèle général EMSC a ensuite été utilisé pour corriger les spectres IR des effets de dispersion non linéaires de la lumière (l'effet de diffusion du Mie) par des particules sphériques de même taille que la longueur d'onde du rayonnement IR (87–89). Le modèle EMSC a également été décliné en une version capable de prétraiter des spectres IR acquis sur des échantillons inclus en paraffine, sans déparaffinage chimique préalable. Dans ce modèle, les interférences spectrales causées par ce milieu d'inclusion sont prises en compte, permettant ainsi de neutraliser la variabilité du signal spectral de la paraffine en tout point de l'échantillon (75,90).

En règle générale, une fois prétraitées, les données doivent être normalisées. La normalisation des spectres a pour but de corriger les effets de variation d'épaisseur de l'échantillon analysé en appliquant des opérations de translation et de compression/dilatation des intensités afin de les exprimer sur une même échelle de valeur. Son but est donc d'homogénéiser tous les spectres à une échelle d'intensité commune. Il existe plusieurs façons de normaliser les spectres infrarouges. La normalisation min-max transforme chaque spectre dans une échelle d'intensité variant de 0 à 1, c'est-à-dire que l'amplitude maximale du spectre vaut 1 et que son intensité minimale est ramenée à 0. La normalisation vectorielle commence par soustraire l'intensité moyenne d'un spectre. Puis cette normalisation divise ce spectre centré par sa norme euclidienne. La normalisation Standard Normal Variate (SNV) est la plus utilisée en spectroscopie infrarouge et consiste à centrer chaque spectre, puis à le diviser par son écart-type. Ces normalisations ont l'avantage de corriger à la fois les effets additifs et multiplicatifs. Le modèle d'EMSC permet également une normalisation sur le spectre de référence choisi par l'utilisateur en divisant chaque spectre corrigé par le coefficient de régression estimé de ce spectre de référence.

Le choix des prétraitements à appliquer et de leurs paramètres est primordial car, mal choisis, ils peuvent générer des artéfacts ou des distorsions, ou encore engendrer une perte

d’information sur le système biologique étudié (91). L’enchaînement de plusieurs prétraitements peut en particulier engendrer une cascade et une amplification des erreurs générées par chaque étape du prétraitement. Il est donc souvent recommandé de limiter le nombre de ces étapes (91), ou alors de se tourner vers un prétraitement par un modèle regroupant plusieurs prétraitements à la fois afin de s’affranchir du problème d’amplification d’erreur. La polyvalence de l’EMSC, la simplicité de son modèle et son principe de corrections multiples « tout-en-un » en font un outil très prisé pour prétraiter les données spectrales infrarouge.

I. C. 2. Réduction de données par sélection et extraction de variables

La spectroscopie infrarouge est une modalité qui génère des spectres riches en information et composés de plusieurs centaines de variables. Parmi ces variables, toutes ne sont pas utiles pour répondre à la question biologique posée, et souvent des corrélations existent entre les variables pertinentes. Il est donc primordial, surtout dans un contexte d’analyse supervisée, de réduire la dimensionnalité des données pour améliorer les performances des modèles de classification supervisée, réduisant ainsi la complexité du modèle et le temps de calcul, et facilitant l'interprétation biologique des résultats. Deux grands types de techniques existent pour atteindre ce but, à savoir la sélection et l’extraction de variables.

La sélection de variables consiste à identifier, parmi les variables d’origine, un sous-ensemble de variables explicatives de la question biologique posée. Trois grandes familles de techniques existent pour réaliser une sélection de variables : les filtres, les méthodes englobantes « wrappers » et les méthodes intégrées « embedded » (92). Les filtres mettent en œuvre des tests statistiques univariés et ne requièrent donc pas d’être associés à une classification supervisée. Par exemple, la sélection de variables par filtre a été appliquée en imagerie spectrale IR, en utilisant un test U de Mann-Whitney, pour identifier les nombres d'onde les plus discriminants entre le côlon cancéreux et sain (93). Les « wrappers » évaluent des sous-ensembles de variables par une méthode de classification supervisée. Le principal avantage de ces méthodes est leur caractère multivarié permettant de prendre en compte les dépendances entre variables. Par contre, ces méthodes sont très gourmandes en temps de calcul. Par exemples, des algorithmes génétiques (GA) combinés à de l'analyse discriminante linéaire ont été appliqués afin d’identifier les régions spectrales les plus pertinentes pour la différenciation tissulaire sur des échantillons d'adénocarcinome de côlon humain (94) et afin de discriminer entre différents types de carcinomes cutanés ou entre différentes structures histologiques de la

peau (95). Les « embedded » sont des méthodes qui optimisent conjointement les sous-ensembles de variables et le classifieur. Ils ont les mêmes avantages que les « wrappers » mais avec un coût de calcul moindre.

L'extraction de variables a pour but de réduire la dimensionnalité des données en transformant et combinant les variables originales afin de créer de nouvelles variables moins nombreuses mais plus discriminantes. Il existe deux grands types de méthodes d'extraction de variables : les méthodes non-supervisées et les méthodes supervisées. Le but des méthodes non-supervisées est de représenter les données dans un nouvel espace de caractéristiques tout en conservant la meilleure reconstruction possible des données originales. La majorité de ces méthodes repose sur des factorisations matricielles cherchant une sous-structure des données. Tout spectre est alors décomposable en la somme pondérée de plusieurs composantes sous-jacentes estimées par ces algorithmes. Ces composantes représentent les nouvelles variables, et la projection des spectres sur ces variables représente les nouvelles coordonnées des spectres dans ce nouvel espace. Parmi ces méthodes, la plus appliquée en spectroscopie infrarouge est l'analyse en composantes principales (ACP) qui a permis, par exemple couplée avec de l'analyse discriminante linéaire (LDA) (96) ou des tests statistiques (97), de révéler des différences spectrales entre des échantillons de côlons sains et tumoraux. Depuis quelques années, une autre de ces méthodes par factorisation matricielle commence à émerger en spectroscopie infrarouge, à savoir la Multivariate Curve Resolution-Alternating Least Squares (MCR-ALS) qui repose sur une estimation itérative et alternée des composantes et de leurs concentrations en chaque spectre. L'originalité de cette méthode est l'introduction de contraintes diverses et spécifiques à l'application biologique qui permettent de limiter l'espace des solutions et donc faciliter l'estimation d'une solution. Cette méthode a été appliquée avec succès pour la segmentation d'images proche-infrarouge acquises sur des échantillons biologiques (98) ou encore pour la caractérisation des structures histologiques du poisson zèbre par imagerie infrarouge (99).

Les méthodes supervisées s'appuient sur une information supplémentaire, les labels des données, pour transformer les données dans un nouvel espace de dimension réduite maximisant les distances inter- et minimisant les variances intra-classes de labels. En spectroscopie infrarouge, la méthode la plus appliquée est la Partial Least Squares (PLS) qui a par exemple été appliquée avec succès pour l'identification de la dysplasie colorectale en conditions in vivo par endoscopie (100).

I. C. 3. Classification

Une fois les données prétraitées et réduites en dimension, la dernière étape de leur traitement consiste généralement en l'application d'algorithmes de classification. Ces outils mathématiques ont été développés pour automatiser la recherche de groupes d'objets dans des jeux composés de grandes quantités de données et/ou de grandes dimensions et/ou présentant de grandes similarités. Deux grandes catégories d'algorithmes de classification existent : non-supervisée et supervisée.

La classification non supervisée, également appelée clustering en anglais, est basée sur le regroupement d'individus en classes de manière à ce que les objets d'une même classe soient les plus similaires possibles et les plus dissimilaires des objets des autres classes. Le qualificatif « non supervisé » signifie que les données ne sont pas labellisées, c'est-à-dire que l'algorithme identifie les différentes classes d'individus à partir de la structure sous-jacente des données. Par contre, ces méthodes requièrent l'intervention de l'utilisateur pour leur paramétrage (nombre de clusters recherchés, métrique utilisée, …). La classification hiérarchique ascendante (Hierarchical Cluster Analysis ou HCA), les K-Means et les Fuzzy C-Means sont les méthodes les plus utilisées en imagerie IR pour identifier les structures histologiques au sein d'échantillons, par exemple, de côlon normal (101,102), de cancer du poumon (103) et d'adénocarcinomes colorectaux (104), sans connaissance a priori précise sur la répartition spatiale de leurs différentes structures. Une fois identifiées, ces structures peuvent être assignées à différentes structures histologiques par un anatomo-pathologiste pour ensuite alimenter une banque de signatures spectrales IR utilisée pour construire des modèles de prédiction (par classification supervisée) de la structure histologique d'échantillons inconnus.

La classification supervisée est basée sur l'exploitation de données d'entraînement labellisées pour construire un modèle de prédiction capable de distinguer les différentes classes d'individus. Une fois entraîné, ce classifieur peut prédire la classe d'appartenance d'un nouvel individu non labellisé. Les performances d'un classifieur dépendent principalement de la taille du jeu d'entraînement, de son échantillonnage au sein de la population mère, du choix de l'algorithme et de son paramétrage. En spectroscopie IR, les méthodes supervisées les plus couramment utilisées incluent les Support Vector Machine (SVM) qui sont appliqués pour prédire le risque de récidive du cancer de la prostate au moment du diagnostic initial (105), l'analyse discriminante linéaire (LDA) utilisée afin de réaliser un diagnostic automatique d'adénocarcinome du côlon en fonction des signatures spectrales des tissus du côlon (106,107),

les réseaux de neurones artificiels (ANN) pour l'identification automatique et fiable des structures tissulaires du cancer du côlon (108), ou les forêts aléatoires (RF) pour une caractérisation des tissus mammaires et le diagnostic du cancer du sein (109).

I. D. Histopathologie spectrale IR

I. D. 1. Définition

Au niveau méthodologique, le concept d'histopathologie spectrale a été défini ces deux dernières décennies comme la combinaison de l'imagerie spectrale infrarouge sur des échantillons tissulaires, sans coloration préalable, et de l'analyse chimiométrique afin de relier des différences spectrales observées sur différentes zones tissulaires à la morphologie du tissu et à des états et stades physiopathologiques spécifiques (110,111). En effet, le paradigme sur lequel l'histopathologie spectrale repose est que la transition des tissus ou cellules d'un état normal à un état pathologique s'accompagne de modifications de leur composition biochimique, qui peuvent se refléter par des changements de signature spectrale infrarouge. Ces changements moléculaires peuvent se produire avant même que la morphologie des structures soit altérée (69).

Outre le fait de ne pas utiliser de colorant, l'histopathologie spectrale présente l'avantage de pouvoir fournir une image « pseudo-couleurs » avec un fort contraste des structures tissulaires d'une façon automatisée et indépendante de l'opérateur, directement comparable à une image acquise sur une coupe colorée adjacente. Contrairement à l'histologie conventionnelle qui requiert l'expertise d'un anatomopathologiste pour caractériser finement les altérations tissulaires associées aux maladies, l'histopathologie spectrale peut donner un résultat objectif. L'implication d'un anatomopathologiste partenaire est néanmoins indispensable pour construire des modèles chimiométriques fiables et pertinents. En effet, la validation et l'interprétation des modèles mathématiques repose sur une confrontation des images infrarouges avec des données de référence (Gold Standard) telles que les colorations hématoxyline-éosine (95,112), l'immunohistochimie (107,111) ou les données cliniques propres au patient (Figure 20).

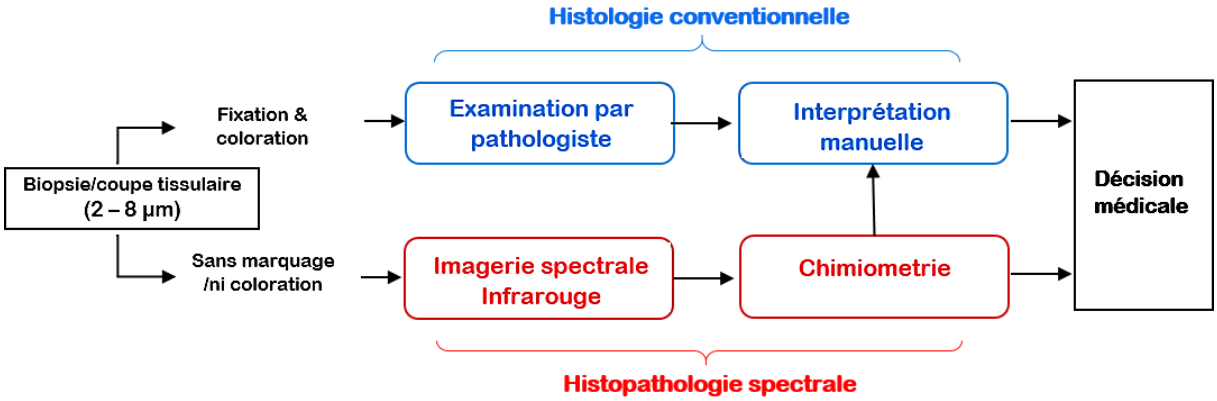


Figure 20. Schéma de principe de l'histopathologie spectrale infrarouge.

I. D. 2. Méthodologie

Ainsi, l'histopathologie spectrale s'avère une technique d'intérêt en anatomopathologie. Son exploitation comme outil de routine nécessite cependant d'optimiser le traitement des données spectrales pour envisager une exploitation optimale des informations biochimiques qu'elles contiennent. Les différentes méthodes de chimiométrie décrites dans la section précédente sont employées en routine en histopathologie spectrale.

La méthodologie d'histopathologie spectrale peut se résumer par le diagramme général présenté sur la Figure 21. Une fois préparés (congelés, inclus en paraffine, coupés en tranches fines), les échantillons sont déposés sur un support adapté à l'imagerie infrarouge. Une fois acquises avec le mode d'acquisition adapté à l'étude, les images spectrales sont prétraitées. Dans un contexte non-supervisé, ces images peuvent ensuite être décomposées par des méthodes d'extraction de variables pour deux objectifs principaux, à savoir (i) l'exploration des données pour identifier les spectres des constituants principaux de l'échantillon et leur distribution au sein de l'échantillon, (ii) l'alimentation directe des méthodes de classification en ne conservant que quelques composantes. L'application des classifications non-supervisées permet d'identifier les structures histologiques principales du tissu analysé. Les partitions ainsi obtenues peuvent ensuite être étudiées et annotées par un anatomopathologiste de façon à faire correspondre à chaque structure histologique contenue dans un échantillon une signature spectrale précise. Cette base de données spectrale labellisée peut servir à la construction de classifieurs supervisés. A l'issue de ces étapes, un modèle de prédiction est disponible afin d'identifier automatiquement une structure histologique (normale, tumorale, inflammatoire…) à partir de sa signature infrarouge, ouvrant ainsi la voie au développement d'un nouvel outil diagnostique.

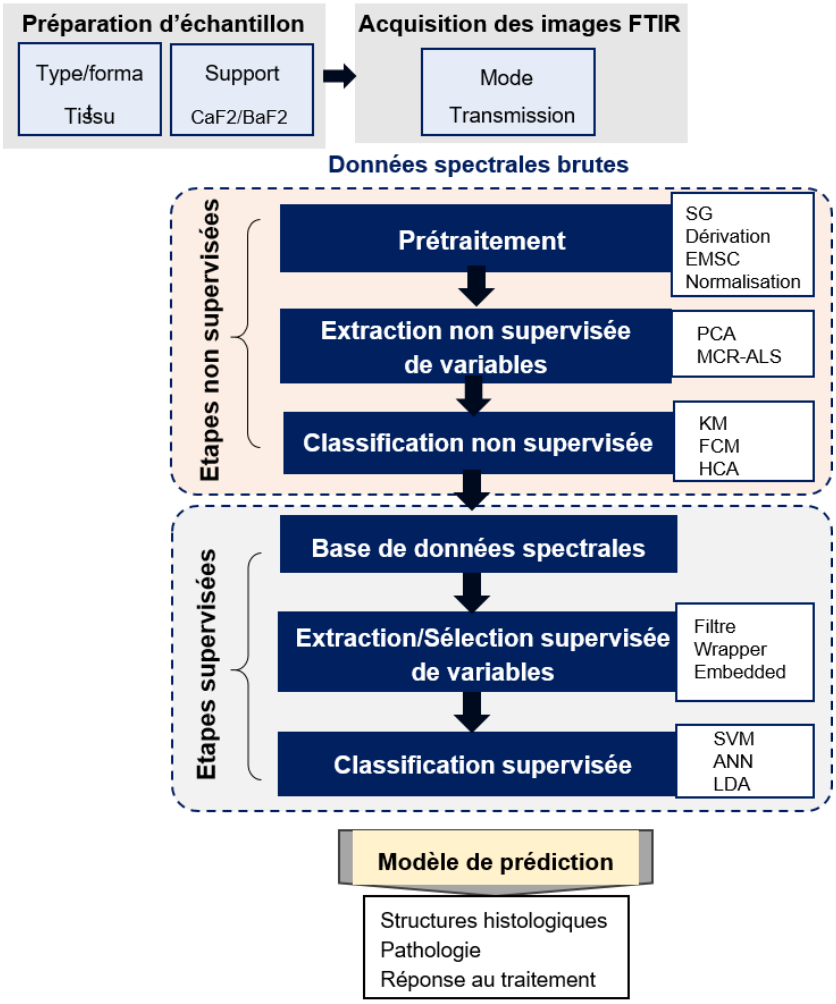


Figure 21. Diagramme général de l'histopathologie spectrale infrarouge.

I. D. 3. Applications de l'histopathologie spectrale dans la recherche biomédicale

S'appuyant sur des informations biochimiques plutôt que sur des changements morphologiques des tissus, l'histopathologie spectrale IR apparait comme une méthodologie d'intérêt dans le domaine biomédical. En effet, de nombreux travaux ont été menés à travers le monde afin d'explorer le potentiel de cette technique. Entre autres, elle a été appliquée avec succès pour :

(i) la caractérisation biomoléculaire de diverses pathologies, telles que l'athérosclérose (113) ou le diabète (114) ;

(ii) le diagnostic des lésions tumorales en entraînant des classifieurs à la distinction entre échantillons cancéreux et normaux, sur divers types de tissus tels que le col de l'utérus (115), le foie (116) le tissu mammaire (117,118), le poumon (119,120) ou encore le côlon (93,107,108,111,121,122). Le Table 3 résume certaines de ces études, en se concentrant sur les paramètres d'acquisition et les méthodes chimiométriques utilisées ;

(iii) pour le pronostic des lésions tumorales pour les cancers de la prostate (123), du côlon (124) ou du poumon (125);

(iv) pour l'évaluation de la réponse à la chimiothérapie de patients atteints d'un cancer de la vessie (126).

CANCER	RÉSOLUTION SPECTRALE	RÉSOLUTION LATÉRALE	MODE DE MESURE	MODE D'ACQUISITION	CHIMIOMETRIE	RÉFÉRENCE
SEIN	4 cm^{-1}	NC	Image	Transmission	Normalisation vectorielle, Mie-EMSC, ACP, Savitzky-Golay, HCA	(118)
	8 cm^{-1}	NC	Image	Transmission	Savitzky-Golay, EMSC, ACP, HCA	(117)
POUMON	2 cm^{-1}	6.25 µm	Image	Transmission	Dérivée seconde, normalisation, HCA	(119)
COLON	6 cm^{-1}	8,8 µm	Image	Transmission	Savitzky-Golay, Normalisation vectorielle, HCA, ANN	(108)
	8 cm^{-1}	6.25 µm	Image	Transmission	EMSC, KMeans, ACP, LDA	(121)
	4 cm^{-1}	6.25 µm	Image	Transmission	EMSC, KMeans, test U de Mann-Whitney, ACP, LDA	(93,107)
	4 cm^{-1}	5.5 µm	Image	Transflexion	Mie-EMSC,Savitzky-Golay, HCA, RF	(111)

Table 3. Exemples d'application l'histopathologie spectrale IR au diagnostic de pathologies cancéreuses. NC = non communiqué.

I. E. Objectifs

I. E. 1. Contexte

Dans les sections précédentes, nous avons pu voir que l'histopathologie spectrale est une méthode de caractérisation tissulaire qui permet de compléter les analyses plus conventionnelles, actuellement réalisées au sein des laboratoires d'anatomopathologie. Cette méthode trouverait un intérêt pour des questions médicales qui restent à ce jour sans réponse. C'est par exemple le cas pour la prédiction de la réponse au traitement, en particulier en ce qui concerne les cancers du côlon métastasés traités par chimiothérapie, pour lesquels les marqueurs histologiques ou moléculaires existants apparaissent difficilement généralisables. L'identification de marqueurs spectroscopiques, de nature numérique et porteurs d'informations biochimiques intrinsèques aux structures tissulaires (hétérogénéité tumorale, interaction cellules tumorales/microenvironnement), pourrait apporter une plus-value aux protocoles d'analyse tissulaire utilisés en clinique. Néanmoins, l'histopathologie spectrale reste aujourd'hui limitée au domaine de la recherche et n'est pas déployée comme outil diagnostique de routine.

I. E. 2. Objectifs

L'objectif principal de ma thèse porte sur des développements d'ordre chimiométrique pour rendre l'approche d'histopathologie spectrale plus robuste, c'est-à-dire indépendante de l'opérateur en automatisant et optimisant certaines étapes de traitement des images spectrales infrarouges. Ces développements sont nécessaires pour appliquer la méthode, de manière fiable, à l'identification de marqueurs spectroscopiques prédictifs de la réponse au traitement. Cette question biologique n'a pas pu être véritablement traitée dans ce travail de doctorat, mais les expérimentations réalisées permettent à présent d'envisager la construction de classifieurs mathématiques dédiés à la prédiction de la réponse des cancers du côlon métastasés à la chimiothérapie. Cette dernière étape de la chaine de traitement des données fera donc l'objet de perspectives.

Le positionnement des expérimentations réalisées au cours de ce doctorat est représenté sur la Figure 22 illustrant les étapes séquentielles de l'histopathologie spectrale. Le premier

développement, au niveau du prétraitement, a consisté à mettre au point une méthode de détection automatique des pixels tissulaires au sein des images spectrales. Il a porté sur les coupes de tissus inclus en paraffine (Chapitre III) puis sur les cryo-coupes de tissus congelés pour lesquelles le gel histologique (OCT) est employé (Chapitre IV). Le second développement a porté sur la construction d'un algorithme permettant une sélection non-supervisée de nombres d'onde discriminants (Chapitre V) de façon à optimiser le partitionnement des données spectrales. Ces développements ont été menés non seulement sur des données provenant d'échantillons biologiques mais également sur des jeux de données simulés, construits dans le but de mimer les signaux tissulaires.

Ensuite, une conclusion générale vient rappeler les principaux résultats et les perspectives, méthodologiques et applicatifs, qu'il nous semble pertinent de mentionner.

Il me parait important de noter que lorsque j'ai commencé cette thèse, il n'y avait pas encore de cohorte d'échantillons. La sélection des tissus a été réalisée en collaboration avec les cliniciens et pathologistes des services d'Hépatogastroentérologie, d'Oncologie Digestive et de Biopathologie du CHU de Reims pour les échantillons humains et avec l'unité IRFAC-U1113 INSERM de Strasbourg pour les échantillons de xénogreffes. Mes contributions à ce travail de thèse ont été la préparation et la mise en place d'une base de données d'images infrarouges de tumeurs du côlon et l'élaboration de solutions méthodologiques optimisant l'histopathologie spectrale.

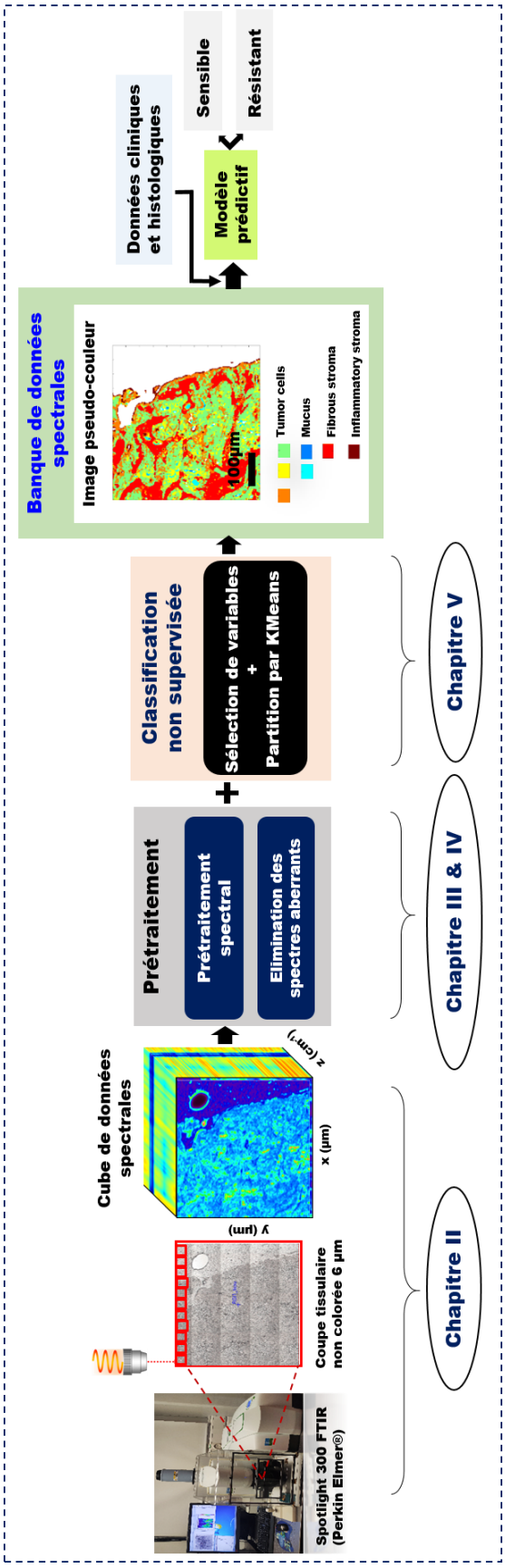


Figure 22. Pipeline général séquentiel des étapes méthodologiques développées durant la thèse.

Chapitre II : Matériels et méthodes

II. A. Choix de la cohorte et critères d'inclusion

II. A. 1. Echantillons humains

Dans ce travail, la cohorte sur laquelle est basée notre analyse a inclus des patients ayant une tumeur de stade IV réséquée dans le côlon distal, rectosigmoïde ou rectal (à l'exclusion du rectum irradié) et traités en première ligne par une chimiothérapie conventionnelle associant un anti métabolite analogue des bases pyrimidiques avec le 5-fluorouracile, l'acide folinique pour diminuer la toxicité des drogues, et un inhibiteur de l'ADN topo-isomérase 1 avec l'irinotécan (FOLFIRI). Cette thérapie a été complétée par une thérapie ciblée de type anti-angiogénique avec du bévacizumab (Avastin) qui consiste en des anticorps monoclonaux qui diminuent la vascularisation de la tumeur et donc son développement. Sachant que le statut mutationnel de certains gènes peut être déterminant dans la réponse aux thérapies ciblées, le statut des gènes KRAS, BRAF ainsi que le phénotype MSI ont été inclus pour chaque patient. L'analyse a porté sur 11 tumeurs de stade IV de patients présentant des métastases hépatiques synchrones avec un ratio homme/femme de 2.7 et en excluant une localisation de la tumeur primaire dans le rectum irradié (Table 4). Sur la base du critère de survie sans progression (PFS), il s'agit de comprendre ce qui différencie les tumeurs colorectales sensibles au traitement des résistantes. L'âge moyen est de 64 ans, les caractéristiques cliniques sont disponibles dans la Table 4. Pour cette étude, tous les échantillons ont été obtenus conformément aux directives éthiques avec le consentement éclairé de tous les patients (autorisation CPP n° AC-2019-3408).

La sélection des échantillons a été effectuée par les collaborateurs au sein des services d'Hépato-gastroentérologie et Cancérologie Digestive et de Biopathologie du CHU de Reims.

N° du patient	Localisation	Stade TNM	KRAS	BRAF	MSI	PFS (mois)	Age	Sexe
Patient # 1	sigmoïde	pT4N2M1	sauvage	sauvage	NC	>110,8	51	H
Patient # 2	rectum	T3N1M+	sauvage	NC	NC	43,3	49	H
Patient # 3	sigmoïde	T4N0M+	muté	NC	MSS	38,9	65	H
Patient # 4	JRS	pT3N2M+	sauvage	sauvage	NC	35,1	69	H
Patient # 5	JRS	T3N0M0	muté	NC	NC	3,5	85	F
Patient # 6	sigmoïde	T4N2M+	sauvage	sauvage	MSS	1,8	48	H
Patient # 7	droit puis gauche	pT3N2 et pT3N0	NC	NC	NC	1,2	70	H
Patient # 8	JRS	T+M+	sauvage	NC	NC	31,6	62	H
Patient # 9	sigmoïde	T3N0M1	sauvage	muté	MSI	38,7	66	H
Patient # 10	transverse	T4N2M+	muté	muté	MSI	20,7	66	F
Patient # 11	transverse gauche	T4N2M+	muté	mute	MSS, RER-	38,4	59	F

Table 4. Caractéristiques physiopathologiques des tumeurs de patients inclus dans la cohorte d'étude. NC = non communiqué.

II. A. 2. Echantillons de xénogreffes

En plus des tumeurs primaires humaines, des échantillons de xénogreffes ectopiques de tumeurs coliques humaines ont été récupérés à l'unité IRFAC U1113 de Strasbourg pour tester et optimiser les algorithmes méthodologiques développés. Il s'agit de 4 xénogreffes ayant des caractéristiques spécifiques (Table 5). La xénogreffe 36TP correspond à la tumeur primaire réséquée dans le côlon gauche du patient et la xénogreffe 36M1 est le résultat d'une métastase hépatique de la tumeur primaire du même patient. Aucun de ces 2 échantillons ne présente de mutation sur les gènes KRAS ou PI3K. La xénogreffe 40 est dérivée d'une tumeur du côlon droit de stade IV avec des métastases hépatiques synchrones, une mutation gain de fonction du gène codant la PIK3CA. La xénogreffe 45 est une métastase d'une tumeur hépatique. Tous ces patients ont reçu une thérapie ciblée avec du cétuximab. Les souris xénogreffées avec chacune de ces tumeurs ont été traitées au cétuximab et en fonction de la courbe de croissance tumorale, les xénogreffes 40 et 45 sont considérées comme résistantes au cétuximab alors que les

xénogreffes 36TP et 36M1 sont sensibles au cétuximab. Les caractéristiques de ces xénogreffes sont présentées dans la Table 5.

N° du Xénogreffe	Stade	Localisation	KRAS	Autre mutation	Age du patient	Sexe du patient	Traitement *
36TP	pT3N2M1	gauche	sauvage	NC	74	H	control/cetuximab
36M1	pT3N2M1	foie	sauvage	NC	74	H	control/cetuximab
40	pT4N1M1	droit	sauvage	PIK3CA	77	F	control/cetuximab
45	CLM	méta-hépatique	sauvage	NC	35	H	control/cetuximab

Table 5. Caractéristiques physiopathologiques des échantillons de xénogreffes inclus dans l'étude. * : le traitement correspond au traitement des souris xénogreffées avec chacune des tumeurs de patients. CLM = métastase de tumeur colique. NC = non communiqué.

II. B. Préparation des échantillons

A partir des 11 tumeurs humaines réséquées décrites ci-dessus, le service de pathologie du CHU de Reims a extrait 12 blocs de tumeurs humaines métastasées. De plus, dans la marge de tissu sain entourant 3 de ces tumeurs, 3 blocs de côlon humain sain ont été réalisés par le service de pathologie du CHU de Reims.

De plus, à partir des 4 xénogreffes décrites précédemment, 13 blocs de tumeurs coliques humaines xénogreffées chez la souris immunodéprimée (NMRI-Foxn1nu/Fox1nu, Janvier®) ont été réalisés par l'Inserm U113 à Strasbourg.

Ces échantillons ont été conservés soit par inclusion en paraffine, soit par congélation, afin de valider les développements méthodologiques réalisés dans ce travail de thèse.

II. B. 1. Coupes des blocs fixés et inclus en paraffine

Les 12 échantillons de tumeurs humaines métastasées ont été fixés au paraformaldéhyde 4% pendant une nuit afin de maintenir la structure cellulaire et les membranes, puis inclus en paraffine (FFPE).

Des coupes de tissus FFPE sériées de six micromètres (6 µm) d'épaisseur ont été réalisées à partir de ces blocs, à l'aide d'un microtome (Nictron Mictrotech France, HM335E). Une coupe déposée sur un support en CaF_2 (fluorure de calcium, Crystal, Dorset, Royaume-Uni) servira pour l'analyse spectrale IR, et la coupe adjacente déposée sur une lame en verre sera dédiée à l'analyse histologique classique par coloration à l'hématoxyline-éosine (H&E). Cette caractérisation sera la référence histologique (gold standard) pour l'analyse IR.

Plus précisément, les coupes de tissus sont déposées sur une goutte d'eau distillée déposée sur le support en CaF_2 et sur la lame de verre. Ensuite, ces supports sont posés sur une plaque chauffante à 30°C jusqu'à évaporation complète de l'eau pour permettre une bonne adhérence de la coupe (Figure 23). Les coupes déposées sur une lame de verre sont colorées à l'H&E et celles déposées sur un support en CaF_2 sont analysées par imagerie IR.

II. B. 2. Déparaffinage des coupes fixées et incluses en paraffine

A l’issue de cette analyse par imagerie IR, les coupes paraffinées et fixées sur un support de CaF_2, comme décrit dans le paragraphe précédent, sont ensuite déparaffinées chimiquement dans du xylène 2x5min, réhydratées dans des bains d'éthanol de 100%, 90% et 70%, et sont rincées dans de l'eau distillée. Ces coupes déparaffinées chimiquement font ensuite l’objet d’une 2ème analyse IR. Les images spectrales IR acquises sur ces coupes déparaffinées chimiquement nous ont permis de construire un gold standard afin de valider la méthode de détection automatique des pixels de tissu développée dans l’article #1.

Des supports en CaF_2 d'une épaisseur de 1 mm ont été choisis pour l'analyse IR, car ils assurent une excellente transmission de la lumière infrarouge et visible, une bonne qualité spectrale, et une réduction de l’aberration chromatique en raison de leur indice de réfraction relativement faible.

II. B. 3. Coupes des blocs congelés

Après prélèvement, 8 blocs de xénogreffes de tumeurs coliques humaines, 3 blocs de muqueuses saines humaines, et 5 blocs de muqueuses saines de souris sont inclus non fixés dans une colle histologique OCT (Optimum Cutting Temperature) et stockés à -80°C. Pour chaque bloc de tissu, deux cryosections adjacentes de 6 µm d'épaisseur sont réalisées dans un Cryostat LEICA (CM 3050 S) maintenu à une température moyenne de -22°C. Les coupes sont étalées grâce à un pinceau sur un support en CaF_2 refroidi au préalable. L’ensemble est ensuite stocké à -20°C jusqu'à l'analyse IR. Les coupes adjacentes sont déposées sur une lame de verre, fixées au paraformaldéhyde (4%), puis colorée à l’H&E pour l'analyse histologique.

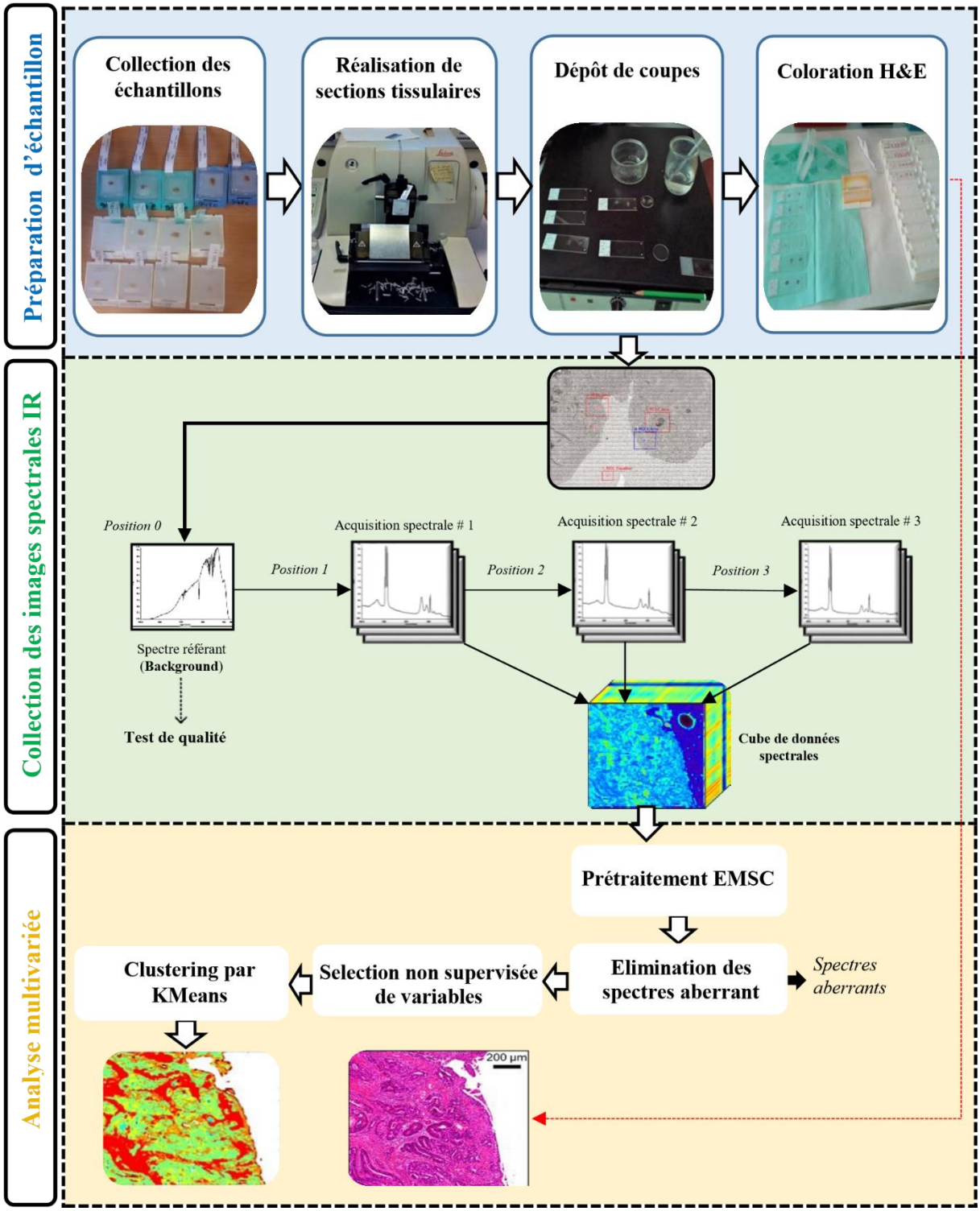


Figure 23. Pipeline général de l'étude : de la préparation d'échantillon FFPE en passant par la collection des images spectrales, jusqu'à l'analyse multivariée des données acquises.

II. C. Méthodologie et collecte des images spectrales FTIR

Les coupes tissulaires ont été analysées en spectroscopie IR à transformée de Fourrier (FTIR) en mode transmission à l'aide d'un spectromètre FTIR Spectrum Spotlight One de Perkin-Elmer couplé à un microscope infrarouge Spectrum Spotlight 300. Ce système est équipé d'un détecteur de type MCT (Tellurure de Mercure-Cadmium) composé de 16 éléments en ligne ; chaque élément mesure le signal correspondant à un pixel de 6.25×6.25µm^2 au niveau de l'échantillon. Ce système optique a été purgé à l'air sec pour réduire les contributions spectrales de la vapeur d'eau atmosphérique et du CO_2. Avant d'accéder au spectre caractérisant notre échantillon, une étape de collection du spectre de bruit de fond ou spectre de référence est nécessaire. Etant donné que les coupes de tissus sont déposées sur des supports en CaF_2, un spectre de ce support est acquis, avec 240 accumulations, de façon à obtenir la transmission spectrale sans échantillon. Ce spectre est ensuite soustrait automatiquement à chaque spectre de l'image FTIR de l'échantillon analysé. A l'aide d'un éclairage en lumière blanche par transmission, l'image visible de l'ensemble de la coupe tissulaire est obtenue. Ensuite, à partir de cette image visible, des zones spécifiques dites régions d'intérêts (ROIs) sont définies manuellement pour être imagées par le rayonnement IR (Figure 23). Ces images sont enregistrées avec une résolution spectrale de 4 cm^{-1}, et 16 accumulations par mesure, offrant un bon compromis entre le rapport signal/bruit et le temps d'acquisition (108). Le système d'imagerie IR est équipé d'un objectif Cassegrain 15X, d'ouverture numérique de 0,65. La taille des pixels de 6.25×6.25 µm^2 permet de collecter des images IR de résolution spatiale d'ordre cellulaire, du même ordre de grandeur que l'image de coloration histologique (127).

Pour chaque échantillon, jusqu'à 4 zones tissulaires différentes et une zone de la paraffine pure (en périphérie de l'échantillon) ont été imagées avec ces paramètres d'acquisition. Au total, 84 images spectrales FTIR ont été collectées, chacune composée de 1100 à 45 000 spectres, couvrant ainsi une partie substantielle de chaque section du tissu.

II. D. Prétraitement des données spectrales

Avant de procéder à l'analyse multivariée des spectres IR, un prétraitement est réalisé. Ce prétraitement a pour but d'extraire le signal originaire du tissu en éliminant les signaux indésirables pour permettre l'application ultérieure des analyses multivariées dans des conditions optimales.

II. D. 1. Prétraitements routiniers

Le prétraitement des données spectrales brutes débutait par la correction de la contribution atmosphérique qui se manifeste principalement dans la gamme 1300-2000 cm^{-1} à cause de la vapeur d'eau et dans la gamme 2300-2440 cm^{-1} à cause du dioxyde de carbone. Cette étape a été réalisée par le logiciel Spectrum IMAGE (Perkin-Elmer).

Les intensités des spectres, exprimées en transmittance, étaient ensuite converties en absorbance.

Puis, les spectres ont été réduits à la gamme spectrale 900-1800 cm^{-1} car cette région, appelée « empreinte digitale », est connue comme étant la plus informative pour des échantillons biologiques analysés par spectroscopie infrarouge.

II. D. 2. Modélisation des signatures spectrales IR des médias d'enrobage des tissus

Comme décrit précédemment, les échantillons analysés sont enrobés dans de la paraffine ou de l'OCT afin de garantir leur conservation et/ou de faciliter la réalisation de coupes fines. Or, ces substances chimiques génèrent une signature IR importante qui peut gêner l'analyse des images IR à plusieurs niveaux. Tout d'abord, ces signatures peuvent avoir des bandes spectrales communes avec le tissu sous-jacent analysé. Ensuite, la distribution de ces substances varie à la surface des échantillons, entraînant une variabilité importante de leur contribution dans les spectres acquis, empêchant ainsi les algorithmes de traitement décrits dans la section suivante de se focaliser uniquement sur les informations tissulaires. Enfin, l'hétérogénéité en termes de composition/conformation/structure de ces substances entraîne une variabilité des

caractéristiques de leurs bandes spectrales principales telles que la largeur à mi-hauteur ou encore les ratios de pics.

Il est donc important de modéliser les sources de variabilité dans les signatures spectrales de ces médias afin de pouvoir s'en affranchir par la suite dans les spectres tissulaires. L'Analyse en Composantes Principales (ACP) a ici été utilisée. Cette méthode linéaire et non supervisée de réduction de dimension transforme les variables originales en de nouvelles variables décorrélées, appelées composantes principales, qui expliquent le maximum de la variance dans les données originales (128). Avant d'appliquer l'ACP sur les spectres, une étape de centrage-réduction des données est souvent réalisée.

Que ce soit pour la paraffine ou bien pour l'OCT, la même procédure a été mise en œuvre pour modéliser les spectres IR du milieu d'enrobage de l'échantillon tissulaire. Afin de simplifier son explication, cette procédure sera décrite et illustrée pour la paraffine, sans perte de généralité pour son application à l'OCT.

Afin de modéliser le signal de la paraffine, une image spectrale a été acquise sur une zone composée uniquement de paraffine pure sur chaque échantillon étudié. Les spectres (environ 10000) de cette image sont ensuite soumis aux prétraitements routiniers décrits dans la section précédente et sont enfin normalisés par SNV. Cependant, la paraffine n'étant pas homogène sur les coupes fines étudiées, choisir un unique spectre de paraffine ou même un spectre moyen de la paraffine ne modéliserait pas correctement son signal au sein du tissu. Une ACP a ainsi été appliquée sur ces spectres centrés-réduits afin d'affiner la modélisation du signal IR de la paraffine. Ensuite, les 9 premières composantes principales, exprimant environ 98% de la variance de l'image FTIR de paraffine, et le spectre moyen de l'image FTIR de paraffine sont conservés pour modéliser le signal IR de la paraffine. Ce modèle servira de matrice d'interférence dans le modèle EMSC décrit dans la section suivante. Le taux de variance expliquée a été fixé à partir des applications du déparaffinage numérique précédemment réalisées dans l'équipe (90,101).

L'application de cette méthodologie à l'OCT étant nouvelle au laboratoire, l'impact du nombre de composantes principales nécessaires pour le modéliser a été étudiée dans le chapitre IV.

II. D. 3. Extended Multiplicative Signal Correction

Des interférences spectrales d'origine physique (telle que la diffusion) ou chimique (telle que la présence du médium d'enrobage) peuvent altérer la qualité des données collectées. Afin de corriger les spectres de ces interférences et ainsi ne conserver que les informations biochimiques relatives à la composition intrinsèque de l'échantillon tissulaire, le prétraitement EMSC a été appliqué dans ce travail de thèse puisqu'il réalise conjointement la correction des effets additifs de la ligne de base, la correction de la variabilité des composantes spectrales caractéristiques du médium d'enrobage, et la correction des effets multiplicatifs d'échelle en raison des différences de longueur de trajet optique et la normalisation des spectres.

Comme pour la section précédente, le modèle EMSC va être présenté dans le cadre du déparaffinage numérique, sans perte de généralité quant à son application à la correction du signal de l'OCT.

L'EMSC a été adapté et appliqué avec succès aux images acquises sur des tissus paraffinés (FFPE) IR (75,101,107,121). Brièvement, à partir d'une image spectrale $\mathbf{S} \in \mathbb{R}^{N_s \times N_\lambda}$ composée de N_s spectres, où chaque spectre $\mathbf{s_i} \in \mathbb{R}^{1 \times N_\lambda}$, avec $1 \leq i \leq N_s$, contient $N_\lambda = \{\lambda_1, \lambda_2, \ldots \ldots \lambda_{N_\lambda}\}$ nombres d'ondes, l'EMSC modélise chaque spectre $\mathbf{s_i}$ comme une combinaison linéaire d'un spectre de référence $\hat{\boldsymbol{s}}$, d'une matrice d'interférence $\mathbf{I}$ et d'une fonction polynômiale $\mathbf{c_i P}$, plus une erreur de modélisation $\mathbf{e_i}$. Ce modèle linéaire s'écrit pour chaque spectre $\boldsymbol{s_i}$:

$$\mathbf{s_i} = a_i \hat{\mathbf{s}} + \mathbf{b_i I} + \mathbf{c_i P} + \mathbf{e_i} \qquad \text{Éq (9)}$$

où :

- le spectre de référence $\hat{\mathbf{s}}$ a été choisi comme le spectre moyen de l'image spectrale considérée, c'est-à-dire : $\hat{\mathbf{s}} = \frac{1}{N_s} \sum_{i=1}^{N_s} \mathbf{s_i}$
- la matrice d'interférence $\mathbf{I} \in \mathbb{R}^{N_I \times N_\lambda}$ correspond au modèle du médium d'enrobage (paraffine ou OCT) composé de N_I composantes, et construit d'après la procédure décrite dans la section précédente afin de modéliser les sources de variabilités spectrales de ce médium.
- $\mathbf{c_i P}$ permet de modéliser la déformation de la ligne de base et l'effet de la diffusion de la lumière par un polynôme d'ordre p, où le vecteur $\mathbf{c_i} = [c_{i0}, c_{i1}, \ldots, c_{ip}] \in \mathbb{R}^{1 \times (p+1)}$

représente les coefficients du polynôme et où **P** correspond à la transposée de la matrice de Vandermonde des N_λ nombres d'onde composant les spectres :

$$\mathbf{P} = \begin{bmatrix} \lambda_1^0 & \cdots & \lambda_1^P \\ \vdots & \ddots & \vdots \\ \lambda_{N_\lambda}^0 & \cdots & \lambda_{N_\lambda}^P \end{bmatrix}^T \in \mathbb{R}^{(p+1)\times N_\lambda}$$

Dans ce travail, en se basant sur l'expérience de l'équipe, la ligne de base a été modélisée par un polynôme d'ordre 4 (129–131).

- a_i, $\mathbf{b_i} = [b_{i1}, b_{i2}, \ldots, b_{iN_I}] \in \mathbb{R}^{1\times N_I}$ et $\mathbf{c_i}$ sont les vecteurs des coefficients de régression de $\mathbf{s_i}$ par $\hat{\mathbf{s}}$, **I** et **P** respectivement. Ils sont estimés par la méthode des moindres carrés afin de minimiser la somme des erreurs de modélisation ou résidu défini par : $E_i = \sum_{k=1}^{N_\lambda} e_{ik}^2$

Chaque coefficient de régression donne une information significative unique. En effet, le coefficient a_i donne accès à l'information sur la contribution tissulaire dans le spectre. Le vecteur $\mathbf{b_i}$ permet d'avoir une indication sur la proportion de la paraffine dans chaque spectre donné. Le vecteur $\mathbf{c_i}$ informe sur la déformation de la ligne de base de chaque spectre de l'image. Quant à elle, l'erreur de modélisation $\mathbf{e_i}$ représente spectralement la différence relative de composition biomoléculaire entre le spectre moyen $\hat{\mathbf{s}}$ et le spectre $\mathbf{s_i}$.

Une fois les coefficients de régression estimés, les spectres sont corrigés par soustraction des contributions spectrales des effets indésirables, à savoir le médium d'enrobage et la ligne de base. Cette correction se traduit mathématiquement par :

$$\mathbf{s_i^{corr}} = \mathbf{s_i} - (\mathbf{b_i I} + \mathbf{c_i P}) = a_i\hat{\mathbf{s}} + \mathbf{e_i} \qquad \text{Éq (10)}$$

Ensuite, afin de tenir compte des facteurs de confusion, tels que l'épaisseur variable de l'échantillon, et aboutir à une intensité comparable entre tous les spectres, nous devons normaliser l'ensemble des données. Par division de chaque spectre corrigé par sa contribution a_i du spectre de référence, une normalisation autour du spectre moyen est réalisée :

$$\mathbf{s_i^{EMSC}} = \hat{\mathbf{s}} + \frac{\mathbf{e_i}}{a_i} \qquad \text{Éq (11)}$$

Après prétraitement EMSC, les spectres sont simultanément neutralisés des variabilités de la ligne de base et de la contribution du médium d'enrobage, et ils sont normalisés autour de leur spectre moyen.

Un exemple de prétraitement par EMSC de deux spectres bruts acquis sur un échantillon tissulaire paraffiné est présenté sur la Figure 24. Pour cette application, le modèle linéaire d'EMSC est illustré sur la Figure 24(a) pour deux spectres bruts. Ces spectres corrigés par EMSC et normalisés par le coefficient de régression a_i du spectre de référence sont présentés sur la Figure 24(b). Ce prétraitement fonctionne parfaitement puisque les sources de variabilités spectrales dues à la paraffine, à la diffusion de la lumière et à la différence de trajet optique ont été complètement corrigées. En effet, les deux spectres prétraités sont maintenant superposables et ne différent que par la différence de composition biomoléculaire existante entre les deux points d'acquisition.

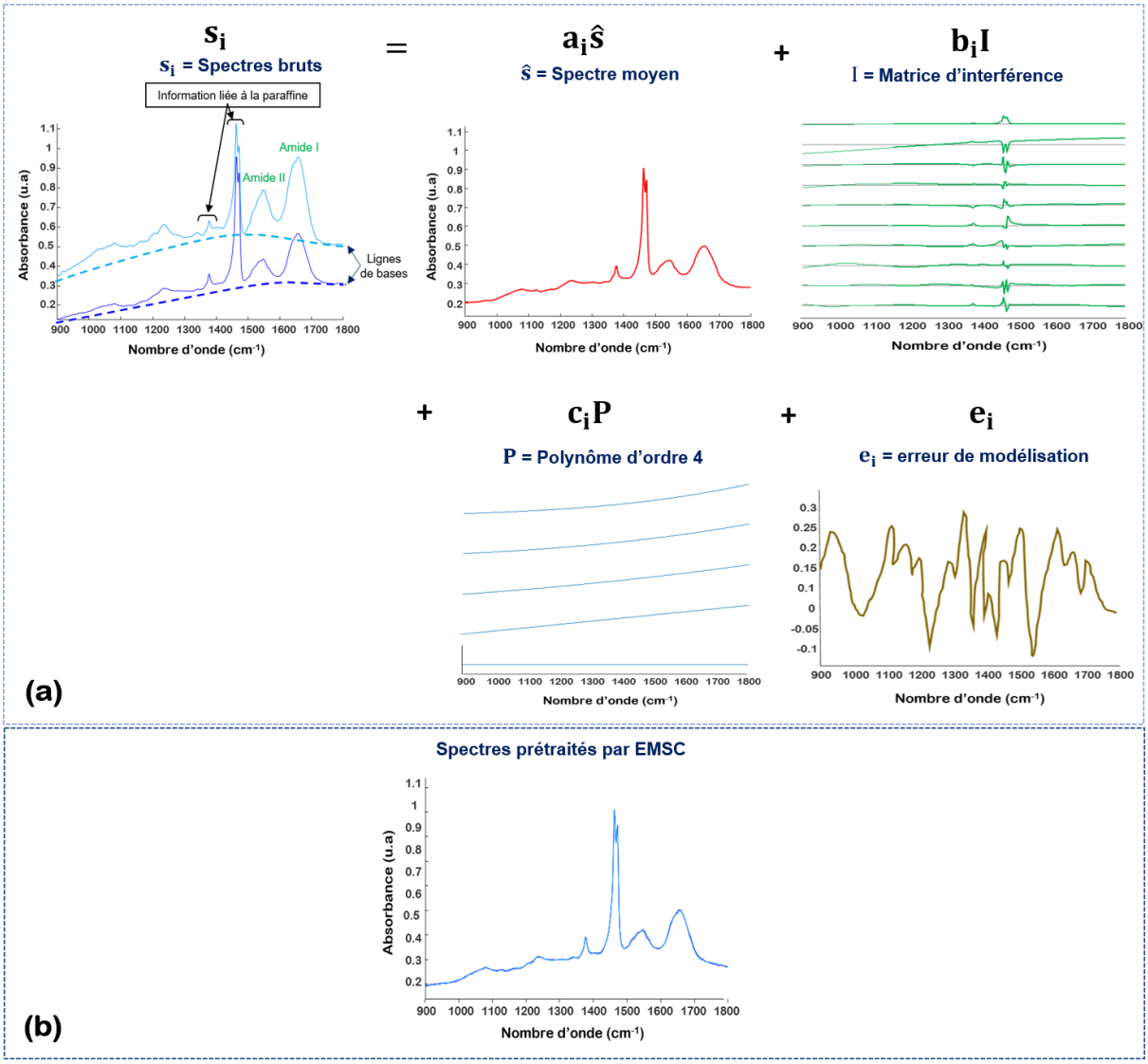


Figure 24. Exemple de prétraitement par EMSC de spectres IR acquis sur une coupe tissulaire fine paraffinée: a) Illustration du modèle linéaire appliqué sur deux spectres bruts. b) Spectres prétraités par EMSC.

II. E. Analyses et traitements multivariés des données spectrales

Une fois les spectres prétraités, leur structure doit être analysée afin d'en extraire une information utile pour répondre aux objectifs visés. Dans ce travail de thèse, ne disposant pas de gold standard précis des échantillons étudiés, des méthodes non-supervisées d'analyse multivariée des données ont été mises au point.

II. E. 1. t-Distributed Stochastic Neighbor Embedding (t-SNE)

De façon à obtenir des informations préliminaires sur la distribution globale des données collectées sur une coupe tissulaire paraffinée, la méthode de réduction de dimension non linéaire t-SNE a été employée.

t-SNE cherche à capturer, dans un espace de faible dimension, la structure locale des données de grande dimension, tout en conservant leur structure globale. Autrement dit, cette méthode cherche à ce que les distances entre points et leur distribution soient conservées entre l'espace final réduit et l'espace original de grande dimension.

Pour ce faire, t-SNE va convertir les distances euclidiennes entre points dans l'espace original en des distributions gaussiennes centrées sur chaque point et d'écart-type fixé par l'utilisateur à l'aide d'un paramètre appelé perplexité. Plus la perplexité est grande, et plus l'écart-type est grand.

Puis, ces données sont représentées dans un espace de faible dimension. Lors de la première itération de l'algorithme, les points sont simplement distribués aléatoirement dans cet espace de faible dimension. Ensuite, les distances euclidiennes entre points dans ce nouvel espace sont converties en des distributions de Student.

Les distributions des points dans les deux espaces sont ensuite comparées par le calcul de la divergence Kullback-Leibler (KL). En utilisant la méthode de descente du gradient de cette divergence de KL, les coordonnées des points dans le nouvel espace de faible dimension sont mises à jour.

En répétant ces étapes un grand nombre de fois, les nouvelles coordonnées sont ainsi optimisées pour représenter au mieux la distribution originale des données dans cet espace de faible dimension.

Dans la littérature, cette méthode est principalement appliquée pour faciliter la visualisation de la structure de données multivariées dans un espace de faible dimension, en général en deux dimensions (132–134).

II. E. 2. Clustering par KMeans

L'objectif de notre analyse multivariée des spectres IR est de chercher à déterminer les caractéristiques structurelles d'un ensemble de données représentant les structures histologiques du tissu en divisant les données en groupes (clusters). Une labellisation précise n'étant pas disponible pour les images spectrales IR acquises sur nos échantillons tissulaires, nous nous sommes tournés vers des méthodes de classification non supervisée. L'algorithme KMeans a été choisi car il est simple et rapide donc bien adapté à des données de grandes tailles. De plus, son efficacité a été prouvée dans de nombreuses études en histopathologie spectrale IR (107,112,121,135).

Cet algorithme itératif cherche à regrouper automatiquement les données en K classes de manière à minimiser la distance, en général euclidienne, entre les données et leur centre de classe (également appelé centroïde), c'est-à-dire qu'il cherche la partition $P = \{P_1, \dots, P_K\}$ qui minimise la fonction objectif suivante :

$$J_{KM} = \sum_{j=1}^{K} \sum_{i=1}^{N_j} \left\| \mathbf{s_{ij}} - \mathbf{c_j} \right\|^2 \qquad \text{Éq (12)}$$

où $\mathbf{s_{ij}}$ est le i^{ème} spectre appartenant à la classe j, $\mathbf{c_j}$ est le centroïde de la j^{ème} classe, et $\|.\|$ est la distance euclidienne. Chaque P_j est composé des N_j spectres $\mathbf{s_{ij}}$, avec $1 \leq i \leq N_j$, appartenant à la j^{ème} classe.

La première étape de cet algorithme est le choix aléatoire de K spectres du jeu de données pour représenter les K centroïdes initiaux des classes. La distance euclidienne est ensuite calculée entre chaque spectre et les centroïdes dans la deuxième étape. Dans la troisième étape, chaque spectre est affecté à la classe dont le centroïde lui est le plus proche. Dans la quatrième étape, les centroïdes sont mis à jour en calculant le spectre moyen de chaque classe, c'est-à-dire :

$$\mathbf{c_j} = \frac{1}{N_j}\sum_{i=1}^{N_j}\mathbf{s_{ij}} \qquad \text{Éq (13)}$$

Les étapes 2 à 4 sont répétées un nombre prédéfini de fois ou jusqu'à ce que plus aucun spectre ne soit réaffecté à une nouvelle classe.

Appliquée à l'histopathologie spectrale, KMeans regroupera des spectres avec des caractéristiques biochimiques similaires (donc appartenant à une même structure histologique) dans un même cluster. Différents clusters permettent donc de décrire les structures constituant un tissu. En attribuant une couleur à chaque cluster, il devient facile de générer des cartes en pseudo-couleurs, facilitant la visualisation des résultats et leur comparaison à une image colorée H&E réalisée sur une coupe adjacente. Cet algorithme permet donc d'élucider la structure histologique d'un échantillon. De plus, les centroïdes correspondant aux spectres moyens des différents clusters, leur analyse permet une interprétation biomoléculaire des différentes structures histologiques et peut conduire à la définition de marqueurs spectroscopiques spécifiques d'une structure tissulaire (90,121).

II. E. 3. Indices de validité

Un indice de validité est un critère quantitatif permettant d'évaluer ou mesurer la qualité d'une partition estimée par un algorithme de classification non-supervisée. D'une manière générale, il s'exprime comme le rapport entre la séparation entre les clusters (distance interclasses) et la compacité des clusters (distance intra-classes) d'une partition $P = \{P_1, \dots, P_K\}$ en K classes :

$$IV(P) = \frac{\text{Séparation entre clusters}}{\text{Compacité des clusters}}$$

Les notions de compacité et de séparation sont illustrées sur la Figure 25.

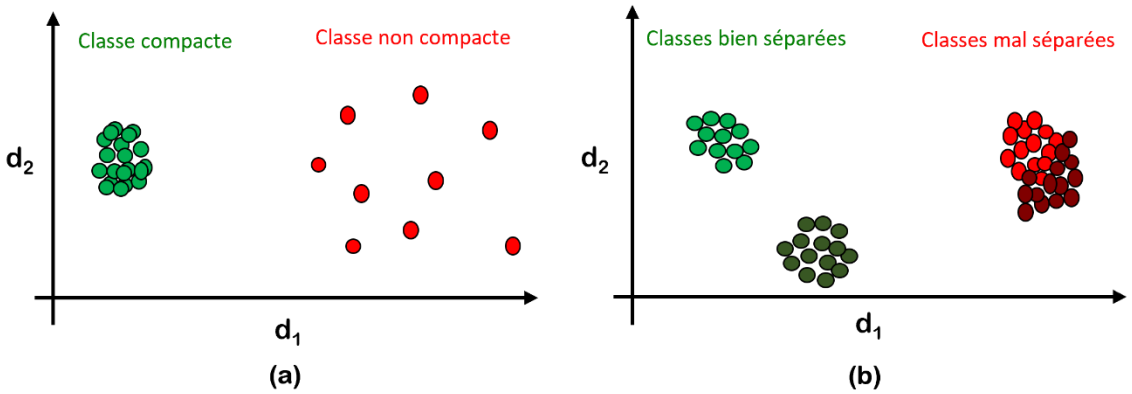


Figure 25. Notions de compacité (a) et séparation (b) de clusters illustrées sur un exemple de données à deux dimensions (d_1 et d_2).

Un indice de validité peut être utilisé pour aider à identifier la valeur optimale d'un paramètre impactant les résultats d'un algorithme de clustering. Dans la littérature, il a été largement appliqué à la détermination du nombre optimal de clusters (136–138). La Figure 26 illustre cette application sur un jeu de données simulé en deux dimensions et composé de trois classes distinctes (visible sur la Figure 26(a)). Pour cet exemple, KMeans est appliqué pour un nombre de clusters *k* variant de 2 à 5 (Figures 26(b-e)). Pour chaque valeur de k, l'indice de validité est calculé sur la partition estimée par KMeans (Figure 26(f)). La meilleure partition est alors obtenue pour le nombre de clusters $k = 3$ qui maximise IV(P). En effet, pour $k = 2$ clusters, KMeans réunit les pixels de deux classes différentes dans un même cluster. La compacité de ce cluster étant mauvaise, la valeur de l'indice de validité va être faible. Pour un nombre de classe plus grand que 3, alors KMeans scinde chaque classe en plusieurs clusters. Cette fois-ci, ce sont les séparations entre clusters qui sont faibles, et donc la valeur de l'indice de validité sera également faible. Seulement pour $k = 3$ clusters les séparations et compacités sont les meilleures.

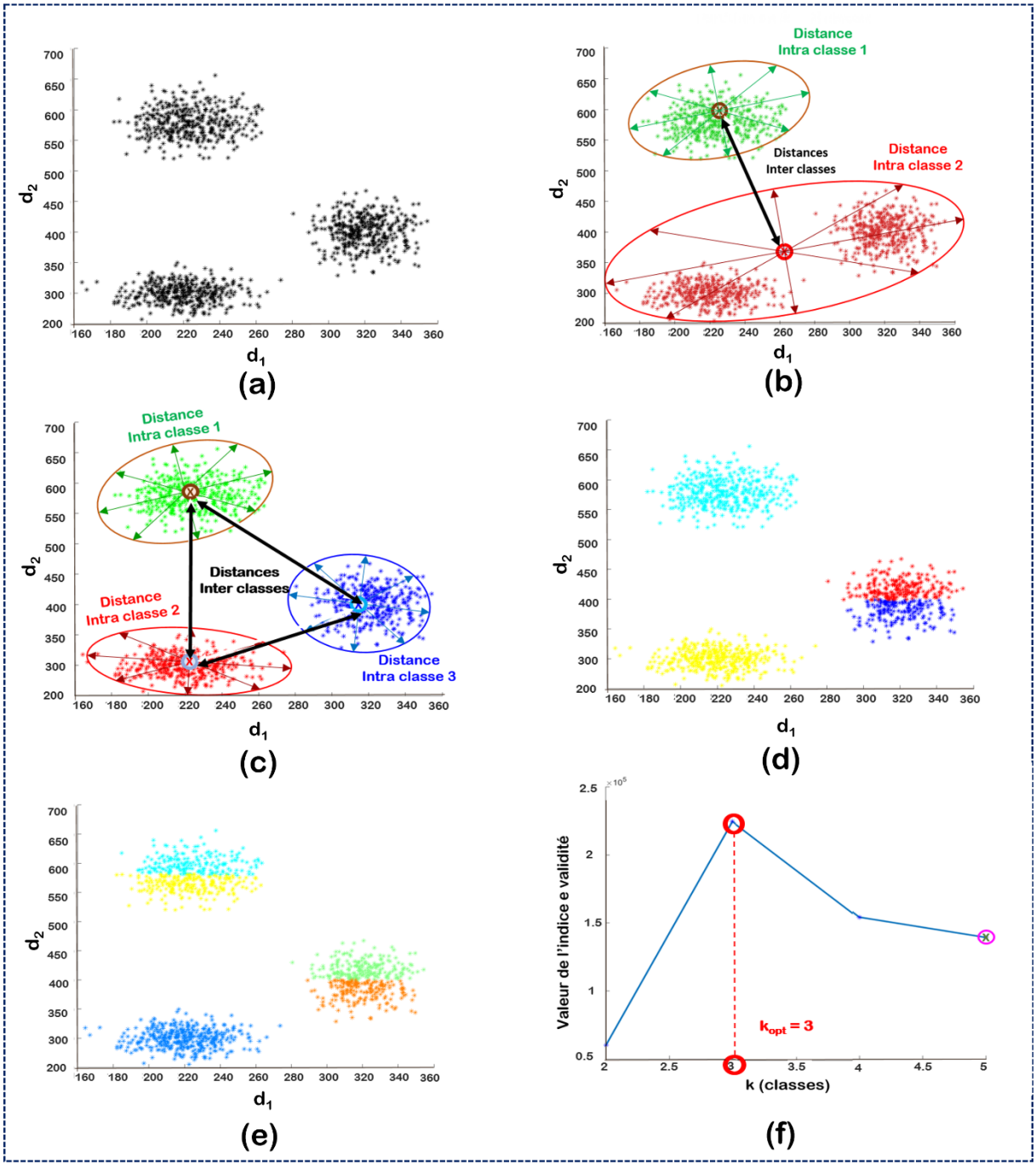


Figure 26. Exemple d'application de l'indice de validité pour optimiser le nombre de classes estimées par KMeans. a) Jeu de données simulé à deux dimensions (d_1 et d_2) composé de trois classes distinctes. b) Partition KMeans à 2 clusters. c) Partition KMeans à 3 clusters. d) Partition KMeans à 4 clusters. e) Partition KMeans à 5 clusters. f) Résultats du calcul d'un indice de validité pour chaque valeur de k.

En histopathologie spectrale, les indices de validité ont été appliqués avec succès pour déterminer le nombre optimal de classes nécessaire pour partitionner des images spectrales IR

acquises sur des coupes de tissu ganglionnaire (139)**,** de tumeurs cutanées (130) ou de côlon normal (101,140).

Dans ce travail de thèse, les indices de validité sont utilisés pour deux applications innovantes : (i) pour estimer les meilleures combinaisons de coefficients de régression EMSC permettant de distinguer par KMeans les pixels non tissulaires (principalement associés au médium d'enrobage) de ceux du tissu, (ii) pour réaliser une sélection non supervisée des nombres d'onde engendrant la meilleure partition KMeans de données IR.

De nombreux indices de validité ont été développés dans la littérature ces dernières décennies. Ces indices diffèrent principalement de par des définitions mathématiques différentes de la compacité et de la séparation de clusters. Dans ce travail de thèse, nous avons choisi d'exploiter les indices Xie−Beni (XB), Davies−Bouldin (DB), Pakhira−Bandyopadhyay−Maulik (PBM), et Silhouette Width Criterion (SWC) car ce sont les plus appliqués dans la littérature et leur efficacité a déjà été prouvée sur des données spectrales acquises sur des échantillons biologiques (140,141).

II. E. 3. 1. Davies-Bouldin (DB)

Il est défini comme le rapport moyen entre la compacité et la séparation de chaque cluster avec son cluster voisin le plus similaire :

$$\mathrm{DB}(\mathbf{P}) = \frac{1}{\mathrm{K}}\sum_{\mathrm{k}=1}^{\mathrm{K}} \max_{\mathrm{m}\in\{1,\ldots,\mathrm{K}\},\mathrm{m}\neq\mathrm{k}} \left\{ \frac{\left(\frac{1}{\mathrm{N_k}}\sum_{\mathbf{x_i}\in\mathbf{P_k}} \|\mathbf{x_i}-\mathbf{c_k}\|^2 + \frac{1}{\mathrm{N_m}}\sum_{\mathbf{x_i}\in\mathbf{P_m}} \|\mathbf{x_i}-\mathbf{c_m}\|^2\right)}{\|\mathbf{c_k}-\mathbf{c_m}\|} \right\} \quad \text{Éq (14)}$$

La meilleure partition est celle qui minimise la valeur de l'indice DB (142).

II. E. 3. 2. Pakhira-Bandyopadhyay-Maulik (PBM)

Cet indice est défini par le rapport quadratique entre la plus grande séparation normalisée $\mathrm{D_N}$ entre clusters et la cohésion totale normalisée $\mathrm{E_N}$ des clusters, et est représenté par l'équation :

$$\mathrm{PBM}(\mathbf{P}) = \left(\frac{\mathrm{D_N}}{\mathrm{E_N}}\right)^2 \quad \text{Éq (15)}$$

avec la plus grande séparation normalisée entre clusters calculée par

$$\mathrm{D_N} = \frac{\max_{\mathrm{l,m}=1,\ldots,\mathrm{K}} \|\mathbf{c_l}-\mathbf{c_m}\|}{\mathrm{K}}$$

et la cohésion totale normalisée des clusters définie par :

$$\mathrm{E_N} = \frac{\sum_{\mathrm{k}=1}^{\mathrm{K}}\sum_{\mathbf{x_i}\in\mathbf{P_k}} \|\mathbf{x_i}-\mathbf{c_k}\|}{\sum_{\mathrm{i}=1}^{\mathrm{N}} \|\mathbf{x_i}-\bar{\mathbf{x}}\|}$$

La meilleure partition est celle qui maximise la valeur de l'indice PBM (137).

II. E. 3. 3. Xie-Beni (XB)

Cet indice de validité est basé sur le rapport entre la compacité moyenne des clusters π et la séparation minimale entre les clusters s par :

$$\mathrm{XB}(\mathbf{P}) = \frac{\pi}{s} \qquad \text{Éq (16)}$$

avec

$$\pi = \frac{1}{N}\sum_{k=1}^{K}\sum_{\mathbf{x}_i \in \mathbf{P}_k} \|\mathbf{x}_i - \mathbf{c}_k\|^2$$

et

$$s = \min_{l\in\{1,\dots,K\},m\in\{1,\dots,K\},l\neq m}\{\|\mathbf{c}_l - \mathbf{c}_m\|^2\}$$

La meilleure partition est celle qui minimise la valeur de l'indice XB(143).

II. E. 3. 4. Alternative Simplified Silhouette Width Criterion (ASSWC)

Cet indice est une variante de l'indice (SWC) (144). Il est un indice de type somme normalisée. La cohésion est mesurée en fonction de la distance entre tous les points d'un même cluster et la séparation est basée sur la distance du plus proche voisin. Il est défini comme :

$$\mathrm{ASSWC}(\mathbf{P}) = \frac{1}{N}\sum_{i=1}^{N} s(\mathbf{x}_i) \qquad \text{Éq (17)}$$

avec

$$s(\mathbf{x}_i) = \frac{b_i}{a_i+\varepsilon},$$

a_i étant la distance entre le point de données $\boldsymbol{x}_i$ appartenant au cluster k et son centroïde $\mathbf{c}_k$:

$$a_i = \|\mathbf{x}_i - \mathbf{c}_k\|$$

b_i est la distance entre le point de données $\boldsymbol{x}_i$ et le centroïde du cluster voisin le plus proche :

$$b_i = \min_{m\in\{1,\dots,K\},m\neq k}\{\|\mathbf{x}_i - \mathbf{c}_m\|\}$$

Le terme ε est une petite constante (fixée à 10^{-6} dans ce travail) utilisée pour empêcher la division par zéro quand $a_i = 0$. La meilleure partition maximise la valeur de l'indice ASSWC (145).

II. E. 4. Algorithme Génétique (AG)

Les algorithmes génétiques (AG) ont été utilisés dans ce travail de thèse afin de développer une nouvelle méthode de sélection non supervisée de variables. La procédure complète que nous avons développée sera décrite dans le chapitre 5. Nous proposons ici une brève définition des algorithmes génétiques.

Un AG est un algorithme évolutif, qui s'inspire de la sélection naturelle et des processus génétiques de l'évolution biologique pour résoudre des problèmes d'optimisation complexes (146). C'est une méthode de recherche générale stochastique, capable d'explorer efficacement de grands espaces de recherche. Un AG débute par initialiser un groupe d'individus (ou chromosomes) formant une population. Chaque individu représente une solution potentielle au problème d'optimisation. Ensuite, en appliquant de façon successive des opérateurs appelés évaluation chromosomique, sélection des parents, croisement des parents et mutation, la population évolue pour générer de nouveaux individus appelés progéniture (ou enfants). Ensuite, le processus d'élitisme sauvegarde les meilleurs parents. Par conséquent, les enfants et les parents élites constituent la génération suivante. Les multiples répétitions de ces étapes garantissent la convergence de l'algorithme vers une solution sous-optimale au problème défini. Dans ce travail, cette procédure générale a été adaptée pour proposer une nouvelle méthode de sélection non supervisée de variables afin d'améliorer le clustering des données spectrales infrarouges.

II. F. Environnement numérique

Hormis la correction atmosphérique qui a été réalisée par le logiciel Spectrum IMAGE (Perkin-Elmer), le reste des prétraitements et analyses multivariées a été réalisé par le logiciel Matlab 2016a (Mathworks, Natick, MA) à l'aide de scripts entièrement développés par l'auteure de ce document et d'une station de travail avec un système d'exploitation de 64 bits, équipée d'un processeur Intel Core i7-4770 cadencé à 3,4 GHz, de 16 Go de RAM et de 4 cœurs.

Chapitre III : Développement d’une approche multivariée pour la détection automatique des pixels de la paraffine

III. A. Préambule

III. A. 1. Contexte et objectif

A l'heure actuelle, l'histologie conventionnelle est la méthode de référence pour le diagnostic et la prédiction de la réponse au traitement des cancers, reposant sur la coloration des tissus suivie de la reconnaissance de l'histologie par un pathologiste qualifié. Plus récemment, l'histopathologie spectrale IR (SHP) s'est révélée être une technique émergente qui peut compléter les modèles traditionnels de diagnostic basés sur l'histologie, en offrant une caractérisation automatique des structures histologiques et un diagnostic pertinent pour le cancer du côlon (104,106,124) réduisant ainsi la charge humaine. En couplant la signature biochimique obtenue par spectroscopie IR à l'information spatiale offerte par microscopie, cette technique peut analyser sélectivement la composition chimique de coupes de tissus non colorées.

Cependant, la fixation des échantillons au paraformaldéhyde et leur inclusion en paraffine est une des méthodes privilégiées de préparation et de conservation des échantillons en routine clinique dans les services d'anatomo-pathologie. Or la paraffine possède une signature spectrale IR caractérisée par des signaux dans les gammes 1360-1390 et 1420-1480 cm^{-1} (dues aux vibrations de déformation des liaisons CH) qui interfèrent fortement avec la signature spectrale IR du tissu sous-jacent. Ainsi, pour que cette technologie s'insère aisément dans le workflow opérationnel de la routine clinique, il est nécessaire de s'affranchir de cette contamination afin d'améliorer la précision des modèles de prédiction de la réponse au traitement thérapeutique qui est fortement affectée par la qualité des données spectrales (bruit, spectres aberrants), en particulier dans le cas où il convient d'identifier des biomarqueurs prédictifs de la réponse au traitement à partir de différences spectrales ténues entre les groupes de patients.

En vue d'éviter les interférences engendrées par la paraffine en imagerie IR, des chercheurs ont testé une approche consistant à éliminer la paraffine en réalisant un déparaffinage chimique avant de procéder à l'analyse IR ou Raman (147). Cependant, en imagerie Raman, aucun des solvants utilisés, dans les quatre protocoles de déparaffinage, à savoir le xylène, l'Histoclear, la récupération d'antigène par la chaleur (HMAR) à l'aide de xylène et de tampon citrate, et Trilogy (déparaffinage et démasquage combinés des antigènes) n'éliminait complètement la

contribution de la paraffine, même après 18 heures d'immersion de l'échantillon (148) Une autre étude en imagerie IR, comparant l'efficacité du déparaffinage par xylène, hexane et huile de paraffine, a prouvé que le protocole utilisant de l'huile de paraffine suivi d'un lavage à l'hexane était le plus efficace pour éliminer la paraffine sur des échantillons de côlon humain normal (149), sans pour autant être parfait. En effet, la rétention de la paraffine est différente d'une structure tissulaire à une autre. Le tissu conjonctif retient plus la paraffine que les régions glandulaires (149). D'autre part, la plupart des méthodes de déparaffinage chimique utilisent du xylène ou de l'hexane qui sont des composés chimiques toxiques, conduisant à l'extraction des composants biochimiques non polaires, tels que les lipides, qui pourraient être utilisés comme biomarqueurs.

Afin de neutraliser la contamination des spectres IR tissulaires par le signal de la paraffine de façon plus efficace et rapide, un protocole numérique a été défini en s'appuyant sur une modélisation des spectres IR de la paraffine par ACP et par la neutralisation de sa contribution dans les spectres tissulaires par EMSC (150). L'efficacité du déparaffinage numérique a été prouvée par la suite dans plusieurs études sur des échantillons de côlon normal ou des lésions cancéreuses cutanées ou coliques, fixées au formol et incluses en paraffine (90,121,131,149).

Cependant, dans les études sur l'histopathologie spectrale, les images IR sont acquises la plupart du temps sur des zones mixtes composées de tissu paraffiné et de paraffine pure entourant l'échantillon. Hors, le but de ces études étant de caractériser la composition biomoléculaire des tissus analysés, les pixels non-tissulaires doivent être éliminés.

Une solution à ce problème a été proposée dans (150) en s'appuyant sur une analyse des coefficients de régression du modèle EMSC de déparaffinage numérique. Cependant, cette approche repose sur un choix subjectif de seuils à partir d'une analyse univariée visuelle d'histogrammes. Cette approche est donc dépendante de l'utilisateur, nécessite plusieurs essais pour optimiser les seuils, et est difficilement applicable à l'analyse de nombreuses images IR acquises sur des échantillons paraffinés.

Par conséquent, notre objectif a été de développer une méthode d'analyse multivariée entièrement reproductible et automatisée pour éliminer correctement les pixels de la paraffine et ainsi ne conserver que les pixels tissulaires pour une analyse ultérieure de son histologie.

III. A. 2. Analyse multivariée des coefficients de régression d'EMSC pour une identification automatique des pixels de paraffine pure

Après application du modèle de déparaffinage numérique par EMSC, toutes les combinaisons possibles des coefficients de régression a_i, $\mathbf{b_i}$ et $\mathbf{c_i}$, et du logarithme népérien du résidu $\mathbf{r_i} = \ln(\mathbf{e_i})$ sont construites. Pour chaque combinaison, un KMeans à deux classes est appliqué. Puis, quatre indices de validité différents (XB, DB, SWC et PBM) (101) sont calculés sur cette partition pour en estimer la qualité. Les combinaisons de coefficients de régression optimisant les valeurs de ces indices de manière consensuelle sont retenues puisqu'elles permettent de séparer de façon optimale les pixels en deux classes avec, idéalement, une classe de pixels de paraffine pure et une classe de pixels de tissu.

III. A. 3. Résultats et discussion

Notre approche a été testée à deux niveaux : i) tout d'abord sur des données simulées afin de contrôler les différentes sources de variabilité parasite existantes dans des spectres infrarouges, à savoir la complexité de la ligne de base, et le rapport signal sur bruit ; ii) puis sur des images spectrales réelles acquises sur des coupes paraffinées de tumeurs coliques humaines.

Le bien fondé de notre hypothèse de travail, à savoir que les coefficients de régression de l'EMSC renferment l'information suffisante pour distinguer les pixels du milieu d'inclusion de ceux du tissu, a d'abord été vérifié par la mise en œuvre d'un outil statistique innovant de visualisation de données multidimensionnelles appelé t-distributed stochastic neighbor embedding (t-SNE). La validation de l'approche multivariée proposée, quant à elle, été mesurée par l'indice de Jaccard calculé entre la partition KMeans à deux classes, estimée à partir de la combinaison optimale des coefficients de régression de l'EMSC et une image gold-standard (partition fixée pour les données simulées ; et pour les données tissulaires, partition d'un KMeans à deux classes obtenue sur la même coupe tissulaire après déparaffinage chimique). En effet, un indice de Jaccard > 0.90 sur les données simulées, et > 0,84 sur les données réelles a toujours été obtenu, confirmant la validité de notre approche.

Nous avons montré que les meilleures combinaisons de coefficient de régression EMSC pour identifier les pixels de tissu sont dépendantes des caractéristiques des données analysées, en termes de bruit et de complexité de la ligne de base. En effet, pour des données ayant un fort

rapport signal sur bruit, quel que soit le type de données (simulées ou réelles), la meilleure combinaison de coefficients de régression de l'EMSC pour exclure les pixels de paraffine a toujours été la même, à savoir les coefficients de régression du spectre de référence (a_i), du spectre moyen du milieu d'inclusion (b_{i0}) et de la constante du polynôme utilisé pour modéliser la ligne de base (c_{i0}). Par contre, sur les données simulées, lorsque le rapport signal sur bruit est faible, alors le résidu devient un paramètre important dans l'identification des spectres de tissu.

Ce travail confirme donc la polyvalence du prétraitement par EMSC pour les images infrarouges acquises sur des échantillons FFPE. En plus de réaliser conjointement la neutralisation de la ligne de base et du signal de la paraffine, et de normaliser les spectres sur le spectre de référence, ces coefficients de régression peuvent être utilisés pour distinguer facilement les pixels de paraffine des pixels tissulaires.

En outre, nous avons également proposé des procédures pour accélérer notre approche afin d'économiser jusqu'à 99% du temps de calcul tout en préservant la précision de la détection des pixels tissulaires.

III. B. Article #1: “Automatic Identification of Paraffin Pixels on FTIR Images Acquired on FFPE Human Samples”

Cette partie du travail a donné lieu à un article publié dans Analytical Chemistry en 2021:

Boutegrabet, W., Guenot, D., Bouché, O., Boulagnon-Rombi, C., Marchal Bressenot, A., Piot, O., & Gobinet, C. (2021). Automatic Identification of Paraffin Pixels on FTIR Images Acquired on FFPE Human Samples. Analytical Chemistry, 93(8), 3750-3761.

analytical chemistry

pubs.acs.org/ac Article

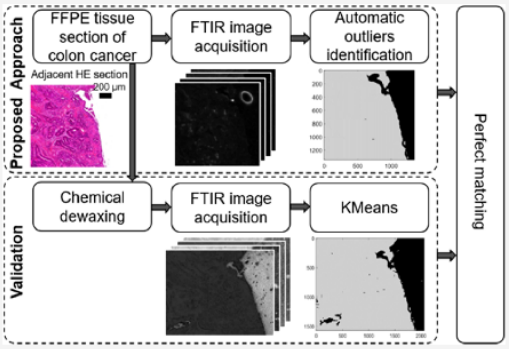
Automatic Identification of Paraffin Pixels on FTIR Images Acquired on FFPE Human Samples

Warda Boutegrabet, Dominique Guenot, Olivier Bouché, Camille Boulagnon-Rombi, Aude Marchal Bressenot, Olivier Piot, and Cyril Gobinet*

Cite This: *Anal. Chem.* 2021, 93, 3750−3761 Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: The transfer of mid-infrared spectral histopathology to the clinic will be possible provided that its application in clinical practice is simple. Rapid analysis of formalin-fixed paraffin-embedded (FFPE) tissue section is thus a prerequisite. The chemical dewaxing of these samples before image acquisition used by the majority of studies is in contradiction with this principle. Fortunately, the in silico analysis of the images acquired on FFPE samples is possible using extended multiplicative signal correction (EMSC). However, the removal of pure paraffin pixels is essential to perform a relevant classification of tissue spectra. So far, this task was possible only if using manual and subjective histogram analysis. In this article, we thus propose a new automatic and multivariate methodology based on the analysis of optimized combinations of EMSC regression coefficients by validity indices and KMeans clustering to separate paraffin and tissue pixels. The validation of our method is performed using simulated infrared spectral images by measuring the Jaccard index between our partitions and the image model, with values always over 0.90 for diverse baseline complexity and signal-to-noise ratio. These encouraging results were also validated on real images by comparing our method with classical ones and by computing the Jaccard index between our partitions and the KMeans partitions obtained on the infrared image acquired on the same samples but after chemical dewaxing, with values always over 0.84.



Tissue microscopic imaging by mid-infrared (IR) absorption spectroscopy appears as an emerging technique to help the pathologists in the molecular characterization of tissues. Combined with statistical data processing, the approach is efficient to evidence various histological structures and to differentiate between physiological and pathological states, without any labeling or staining agent.[1] The analytical capability of the technique relies on the multivariate nature of the recorded spectra, reflecting the overall biochemical composition of the sample and the molecular alterations associated with physiological changes or malignancy. In the scientific literature, this approach, named spectral histopathology (SHP), has proven to be effective in the identification of cancerous tissues in comparison to their nontumoral counterpart. Several types of lesions were studied, e.g., skin,[2] melanomas,[3] prostate,[4] lung,[5] cervical,[6] brain,[7] breast,[8,9] and colon[10−13] cancers. Besides the analysis of cancerous tissues, SHP was applied in other biomedical issues such as the characterization of inflammation[14] or age-related alterations in tissue such as skin.[15] These examples were proof-of-concept studies, but several assets of mid-IR imaging make it possible to consider the deployment of the technique for clinical diagnostic applications. For a routine use, this technique is less expensive once the equipment is acquired and allows us to map the tissue at a microscopic scale and in a short time.[1] Importantly, data processing can be automated so as to be totally independent of the operator contrary to conventional histopathological examination. In ambiguous issues, the interpretation of histology or immunostaining can be subject to a lack of consensus between several pathologists, each interpreting with his/her own expertise and subjectivity.

The clinical transfer of this biophotonic technology requires the confirmation of the proof-of-concept studies on large-scale retrospective cohorts coming from tumor biobanks. However, the majority of these biopsies are formalin-fixed paraffin-embedded (FFPE) for preservation, preparation, and long-term storage purposes, and paraffin has a significant infrared response disrupting the infrared image analysis. In previous studies,[2,12,16−21] we have demonstrated the possibility to analyze FFPE tissues without chemical dewaxing, which facilitates retrospective studies on reference tissues in tumor

Received: September 15, 2020
Accepted: February 3, 2021
Published: February 16, 2021

Downloaded via UNIV DE REIMS CHAMPAGNE-ARDENNE on February 3, 2022 at 19:39:23 (UTC).
See https://pubs.acs.org/sharingguidelines for options on how to legitimately share published articles.

© 2021 American Chemical Society

https://dx.doi.org/10.1021/acs.analchem.0c03910
Anal. Chem. 2021, 93, 3750−3761

banks. Indeed, spectral interferences of paraffin can be modeled and neutralized to keep only the spectral diversity associated with the molecular composition of the tissue structures. In addition, pixels corresponding to noisy spectra or pixels with a strong contribution of paraffin are considered as outliers and can be removed from the dataset. The identification of outliers' pixels is an important step for the construction of an efficient classification model.

The procedure to remove outliers was originally developed from the extended multiplicative signal correction (EMSC) preprocessing of the data.[22] In the current version of the algorithm,[16,20] the outlier's elimination proceeds by selecting manually thresholds on values of two coefficients, precisely of the reference spectrum fit and modeling error, respectively. However, this operator-dependent selection is a break on the automation of the SHP.

Therefore, in this study, we describe a novel methodology that automatically removes the outlier's spectra of mid-infrared spectral images collected from FFPE colon tissues. Original tools of data partitioning and simulated data were also used to demonstrate the performance of the separation between spectra informative of tissue composition and outliers' spectra. The performance of our automatic approach has been evaluated on simulated and real infrared images and compared with other existing semiautomatic methods for detection of paraffin and outliers' spectra.

EXPERIMENTAL SECTION

Samples. *Formalin-Fixed Paraffin-Embedded Tissue Sections.* FFPE blocks of metastatic colon carcinoma were obtained from the colon cancer surgery of three patients with T4N1M1 staging, at the pathology department of the Reims university hospital. Written informed patient consent was obtained according to the approved local ethics committees (no. AC-2019-3408).

Two FFPE blocks of xenografted human colon carcinoma were obtained from the INSERM U113 research group. This animal experiment was conducted in accordance with the French Ethical Approval Apafis#16125-2018030716202418 v2 according to the European guidelines.

For each block, two consecutive 6 μm thick sections were cut using a Microm HM 335 E microtome (Microm Microtech, Brignais, France). The first section was deposited on a calcium fluoride (CaF_2) window (Crystran, Dorset, U.K.) for mid-infrared spectral analysis. The adjacent section was mounted on a glass window and stained with hematoxylin and eosin (HE) for conventional histology, to serve as a reference for the spectral histopathology analysis.

Chemically Dewaxed FFPE Tissue Sections. In the majority of published studies, FFPE tissue sections are chemically dewaxed before mid-infrared spectral analysis to remove the paraffin,[5,6,8−10,13,23−26] which presents a parasitic signal superimposed to the tissue infrared signal.[16,20,27]

In this work, five FFPE tissue sections (two for mice and three for human patients) that were previously analyzed by mid-infrared spectral imaging were chemically dewaxed by immersion in several xylene baths. Then, mid-infrared spectral imaging was performed once again on these chemically dewaxed tissue sections, which will be thus considered as gold standard in this study.

Thus, to objectively evaluate the performances of the different investigated approaches, the distinction between tissue and non-tissue spectra on the FFPE sections (presented in the Identification Methods of Pure Paraffin Spectra section) will be compared with the clustering outcomes obtained from the corresponding dewaxed sections.

Frozen Tissue Section. To generate simulated FFPE IR spectral images, a xenografted human colon carcinoma sample has been embedded in the Tissue-Tek optimum cutting temperature (O.C.T.) formulation to slice 6 μm thin cryo-cross sections using a LEICA (CM 3050 S) at −20 °C. From this sample, a tissue area has been selected for FTIR spectral imaging using the same procedure as for FFPE tissue sections.

Data Collection. Fourier transform infrared (FTIR) images were collected in transmission mode using a Spectrum Spotlight 300 FTIR imaging system coupled to a Spectrum One FTIR spectrometer (PerkinElmer, Courtaboeuf, France), equipped with a liquid nitrogen-cooled mercury−cadmium−telluride (MCT) detector. Prior to spectral image acquisition, a visible image of the sample was collected to select up to four different tissue zones and one pure paraffin zone to be analyzed. For each pixel of these selected areas, 16 scans were averaged on the spectral range 750−4000 cm^{-1}, using a spectral resolution of 4 cm^{-1} and a pixel size of 6.25 × 6.25 μm^2. A background spectrum from the CaF_2 window was recorded using 240 accumulations and subtracted automatically from each collected image by the Spectrum Image software (PerkinElmer).

In total, 12 FTIR spectral images were collected on the FFPE tissue sections (three for mice and nine for human patients), and 12 on the chemically dewaxed tissue sections (three for mice and nine for human patients), with 15 000 pixels per image in average. In addition, five spectral images of paraffin composed approximately of 12000 pixels were recorded to model the spectral interference signal in the EMSC model defined in the Data Preprocessing section.

Simulated Spectral Images. The validation of a method on real-world datasets is the final and most valuable step. However, simulated datasets are usually constructed for the following reasons.[28] First, the ground truth is perfectly known. Second, the main variability sources that can be observed in the real datasets are completely under control. The influence of these variability sources on the results can thus be easily studied. Third, a new method is often dependent on internal parameters (such as the number of latent variables for partial least squares) that can be easily tuned on a simulated dataset. The behavior of a newly proposed method can thus be fully understood using a simulated dataset.

In this sense, we constructed simulated FTIR spectral images of FFPE tissue sections using the model described in the Supporting Information. This model allows us to modulate the simulated spectral images according to the following parameters. First, the ratio between the tissue and pure paraffin areas can be adjusted by specifying their respective number of pixels. Second, the baseline complexity can be modulated by specifying the order of the polynomial function used to model the baseline. Third, the signal-to-noise ratio (SNR) can be controlled by the standard deviation σ of the Gaussian noise added to the model.

Data Preprocessing. First, for the real FTIR images, raw data were corrected from the water vapor and carbon dioxide atmospheric absorptions by the Spectrum IMAGE software (PerkinElmer).

Second, these spectra were cut in the 900−1800 cm^{-1} fingerprint region since it is known to be the most informative spectral range for this type of samples.[17]

https://dx.doi.org/10.1021/acs.analchem.0c03910
Anal. Chem. 2021, 93, 3750 3761

Third, EMSC[22] was applied to the real and simulated IR images using the following linear model for each spectrum s_i:

$$s_i = a_i\hat{s} + b_i I + c_i P + e_i \quad (1)$$

In this work, $\hat{s}$ is a reference spectrum chosen as the average spectrum of the dataset. On an FTIR image acquired on a pure paraffin area, the mean spectrum was computed and a principal component analysis was performed to find the main sources of spectral variability due to paraffin. The mean spectrum and the $N_I = 9$ first principal components (expressing 98% of variance) were pooled in a matrix I, named interference matrix, to model the paraffin variability into the EMSC model. P is a fourth-order Vandermonde matrix of wavenumbers used to model the baseline and light scattering effect. e_i is the modeling error vector. a_i, $b_i = [b_{i0}, b_{i1}, ..., b_{iN_I}]$ and $c_i = [c_{i0}, c_{i1}, ..., c_{i4}]$ are the regression coefficients of $\hat{s}$, I, and P, respectively, and are estimated by ordinary least squares. Then, each spectrum is corrected using the following equation

$$s_i^c = \hat{s} + e_i/a_i \quad (2)$$

Thus, the EMSC preprocessing allows us to perform simultaneously (i) the neutralization of the variabilities of the baseline and of infrared contribution of the paraffin embedding medium and (ii) the normalization of the data spectra around the mean dataset spectrum. The reader can refer to refs 20, 22, and 27 for more details about the EMSC model and its application to spectral FTIR images acquired on FFPE tissue sections.

To assess the model performance, the modeling residue $\sum_{k=1}^{N_\lambda} e_{ik}^2$ is usually computed, where N_λ is the number of wavenumbers composing each spectrum.[16,17] However, this expression being quadratic, its visualization using histogram or estimated density function is difficult. To avoid this problem, we will consider the natural logarithm of the modeling residue $r_i = \ln(\sum_{k=1}^{N_\lambda} e_{ik}^2)$.

Identification Methods of Pure Paraffin Spectra. *Spectral Band Ratio (BR).* Spectral band ratio is a routinely used method to detect spectra contaminated with an unwanted compound contribution, such as water vapor, substrate, noise, or preservation medium.[29] This method is based on the computation of the ratio between the integrated intensities of two bands and on the definition of a decision threshold.

In our study, this method has been applied to detect the pure paraffin pixels on the recorded FTIR tissue images by computing the ratio between the 1600−1700 cm^{-1} Amide I band associated with the tissue and the 1430−1490 cm^{-1} paraffin band. Previously to the ratio computation, each spectral band was corrected from its baseline computed as the straight line passing through the two band extreme wavenumbers,[30] i.e., 1600 and 1700 cm^{-1} for the Amid I band, and 1430 and 1490 cm^{-1} for the paraffin band. An example of estimated baselines is shown in Figure S-1a.

Spectra were considered as pure paraffin pixels if their band ratio was less than a threshold that was manually and differently chosen for each FTIR image.

Univariate Analysis (UA) of EMSC a Regression Coefficient and r Modeling Residue. Univariate analysis of EMSC regression coefficient a and modeling residue r has been developed specifically to detect pure paraffin spectra[16] and applied in the majority of studies[2,17,20,27,31] on FTIR images acquired on FFPE tissue sections. This method is based on the manual selection by the operator of two different thresholds, one for the a fitting coefficient and another for the natural logarithm of the modeling residue r estimated by EMSC. These two thresholds, named τ_a and τ_r in the following, are selected independently of each other. Spectra for which $a_i > \tau_a$ and $r_i < \tau_r$ are considered as good-quality tissue spectra.

This method has also been adapted to analyze sample in tissue microarray.[12,19]

Multivariate Analysis (MA) of EMSC Fitting Coefficients. In this paper, we propose a new method to identify the nontissue spectra. After EMSC, a set of fitting coefficients belonging to $\{a,b_0,b_1,...,b_{N_I},c_0,c_1,...,c_4,r\}$ is selected and processed by KMeans clustering[32] to decompose the dataset into two groups, i.e., one group for the nontissue spectra and one for the tissue spectra. Contrary to the two previously presented methods, the proposed approach is multivariate by nature since the fitting coefficients are simultaneously exploited. Furthermore, this method is automatic since it is based on a clustering algorithm and objective since it is not based on a manual threshold selection.

To automatize the selection of the set of fitting coefficients, validity indices were applied. A validity index objectively measures the quality of a partition, usually based on within-cluster compactness and between-cluster separation measures.[18,33] When several partitions are estimated for different values of the parameters of a clustering algorithm, e.g., the number of clusters, a validity index is useful to determine the best partition, by optimizing the values of the parameters.

In our study, the validity indices were applied on two-cluster KMeans partitions estimated on each possible combination of EMSC coefficients. The best set of fitting coefficients is thus the one resulting in the optimal validity index value, leading to the most distinct clusters of tissue and preservation medium.

Numerous validity indices have been developed in the literature.[18,33] To have robust results, we decided to apply four different validity indices, i.e., Xie−Beni (XB), Davies−Bouldin (DB), Pakhira−Bandyopadhyay−Maulik (PBM), and Silhouette width criterion (SWC).[18,33] The optimal value is the smallest for XB and DB, and the highest for PBM and SWC.

Representation of Pixel Memberships. At the end of each of the above-presented procedure, the pixel memberships can be represented by a binary image where 0 and 1 denoted pixels identified as pure paraffin and tissue, respectively.

Image Registration. For a given tissue section, the FTIR spectral images on the chemically dewaxed tissue section were not acquired exactly at the same orientation and position as that on the FFPE tissue section, due to slight morphological changes during the dewaxing procedure. Furthermore, the size of the spectral image acquired on the chemically dewaxed sample was chosen higher to include the area scanned on the FFPE tissue section.

To precisely compare these acquired areas, intensity-based image registration was applied between a moving image and a fixed image using a rigid transformation consisting of translation and rotation.[34] In our case, the moving image is the binary image resulting from one of the previously presented identification methods of pure paraffin spectra, while the fixed image is the partition obtained by applying a

https://dx.doi.org/10.1021/acs.analchem.0c03910
Anal. Chem. 2021, 93, 3750 3761

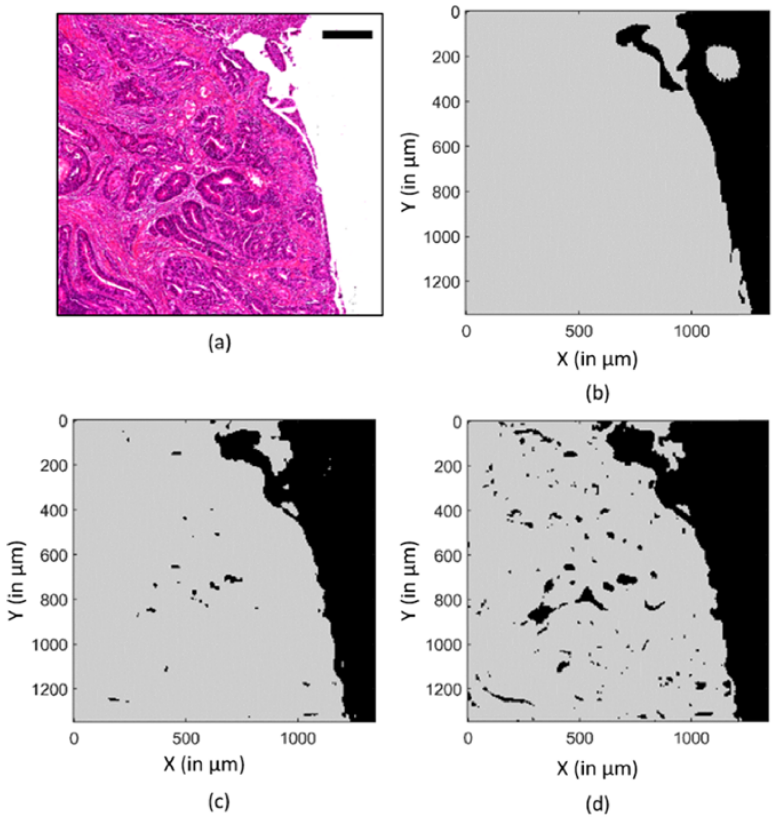


Figure 1. Identification of paraffin and tissue pixels by the spectral band ratio approach. (a) Human colon cancer FFPE tissue section stained with HE. The scale bar indicates 200 μm. (b–d) Binary images resulting from the thresholding of the estimated probability density using thresholds equal to 0.3, 1.6, and 2.3, respectively. The black and gray pixels correspond to estimated paraffin and tissue pixels.

two-cluster KMeans on the FTIR image acquired on the chemically dewaxed sample.

t-Distributed Stochastic Neighbor Embedding (t-SNE). In this study, t-SNE was used as a tool facilitating the visualization of multivariate datasets.[35] From a high-dimensional space, this nonlinear dimensionality reduction technique aims to find the best low-dimensional (usually two-dimensional) mapping preserving the local neighborhood structure of data.[36] With t-SNE, very similar data points close to each other in the original high-dimensional space are kept close in the new low-dimensional space.

Gold Standard and Jaccard Index. To evaluate their performance, the identification methods of pure paraffin spectra presented above were compared to a gold standard using the Jaccard index.[37] For simulated data, the gold standard was directly accessible from the model since paraffin and tissue pixels are known. For the real data, it was defined by applying a two-cluster KMeans on the FTIR image acquired on the same tissue section after chemical dewaxing.

The Jaccard index measures the similarity between two sample sets A and B, and is defined as the ratio between the intersection of A and B and the union of A and B

$$J(A, B) = \frac{|A \cap B|}{|A \cup B|} \tag{3}$$

The Jaccard index is between 0 for disjoint sets and 1 for identical sets.

In this work, the set A is defined as the pixels identified as pure paraffin on a simulated FTIR image or on a real FTIR image before chemical dewaxing by one of the pure paraffin identification methods presented above. The set B corresponds to the pixels identified as paraffin on the simulated image, or as CaF_2 by a two-cluster KMeans applied on the real FTIR image acquired on the same tissue section after chemical dewaxing.

Programming Environment. All of the data processing presented in this study was carried out using in-house scripts written in Matlab (MathWorks, Natick, MA).

■ RESULTS AND DISCUSSION

To compare their performance, the identification methods of pure paraffin spectra were applied on the FTIR images acquired on the FFPE tissue sections. However, to ease the reading, the results will be illustrated using a representative FTIR spectral image acquired on a human colon carcinoma FFPE sample whose HE stained section is presented in Figure 1a.

Furthermore, the classical UA and the proposed MA identification methods of pure paraffin spectra being based on preprocessing by EMSC, the model components and examples of application on this spectral image are presented in Figures S-2 and S-3. The reference spectrum $\hat{s}$, the components of the interference matrix $\boldsymbol{I}$, and the polynomial functions comprising the Vandermonde matrix $\boldsymbol{P}$ are shown in Figure S-2. Examples of raw spectra acquired on this image on paraffin and tissue pixels, and their EMSC preprocessed versions are shown in Figure S-3. The efficiency of EMSC is visible since the preprocessed paraffin spectra are mainly composed of noise, in

https://dx.doi.org/10.1021/acs.analchem.0c03910
Anal. Chem. 2021, 93, 3750−3761

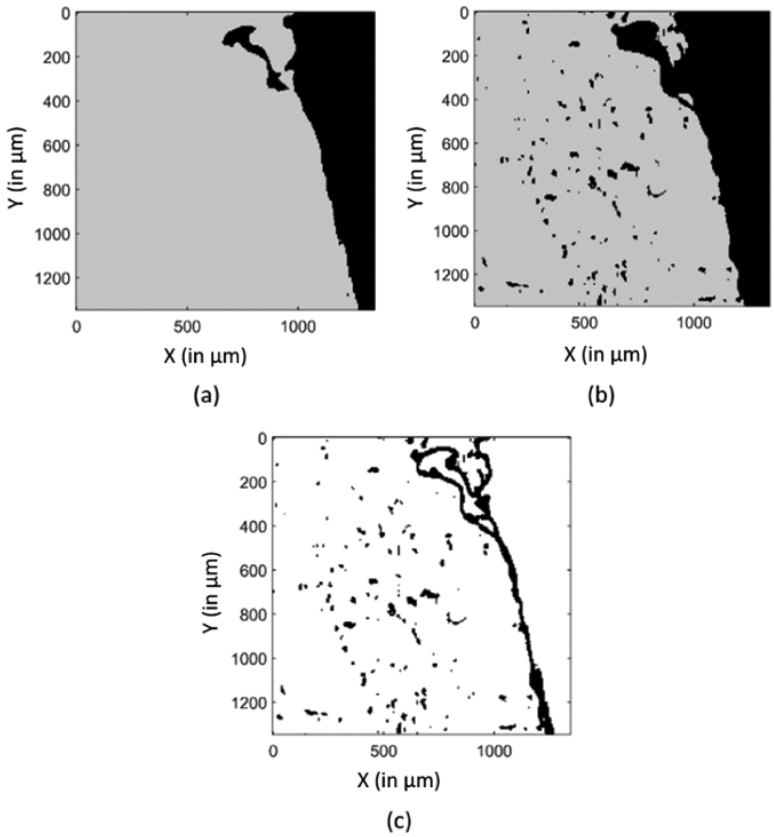


Figure 2. Identification of paraffin and tissue pixels by the univariate analysis of EMSC a regression coefficient and r natural logarithm of modeling residue. Binary images resulting from the thresholding of the estimated probability densities using $\tau_a = 2$ and $\tau_r = 0$ (a) and $\tau_a = 9$ and $\tau_r = -3$ (b), respectively. The black and gray pixels correspond to estimated paraffin and tissue pixels. (c) Image in black represents the pixels differently identified between (a) and (b).

contrast to the preprocessed tissue spectra on which the neutralization of the paraffin signature is evident while preserving a tissue signature variability revealing the subtle biomolecular differences between tissue pixels.

Limitations of Conventional Identification Methods of Pure Paraffin Spectra. First, the spectral band ratio method was applied on the spectral images acquired on the FFPE tissue sections. The ratio was computed between the paraffin and amide I bands corrected from their baseline (Figure S-1a) and can be represented as a ratio intensity image (Figure S-1b). Then, these computed ratio values were summarized by their smoothed probability density function estimated by a normal kernel function[38] (Figure S-1c). By visual inspection of this distribution and as a function of its shape, the threshold value is selected by the operator. This density has the shape specific of a bimodal probability density function. The threshold value must thus be determined between these two modes by a visual analysis.

As an illustrative example, three different operators analyzed independently this ratio distribution. The first one selected a threshold equal to 0.3 to be tissue-conservative, while the second chose an intermediate threshold equal to 1.6, and the third fixed the threshold to 2.3 to remove all of the paraffin pixels. Figure 1b−d shows the binary images obtained after thresholding by these three operators. An underspecified threshold erroneously identifies nontissue parts of the sample as tissue (Figure 1b). On the contrary, an overspecified threshold confuses tissue parts with paraffin (Figure 1d).

Second, the univariate analysis of EMSC regression coefficient a and natural logarithm of modeling residue r was tested on the same spectral image to detect the pure paraffin pixels. After EMSC preprocessing, the distributions of a and r are estimated by a normal kernel function[38] (Figure S-4). The a and r distributions usually have shapes specific of bimodal probability density functions in the context of spectral image acquired on FFPE tissue sections.[16] For each distribution, a threshold value is thus determined between the two modes by a visual analysis. Pure paraffin spectra are characterized by low regression coefficient a, while tissue spectra have a high regression coefficient a. A low natural logarithm of modeling residue r is typical of paraffin or tissue spectra well fitted by the EMSC model, whereas a high value is characteristic of noisy or outlier spectra.

To illustrate the sensitivity of this method to threshold selection, the threshold values were selected by two different operators. The first one selected $\tau_a = 2$ and $\tau_r = 0$ to surely conserve all of the tissue pixels. The second one chose $\tau_a = 9$ and $\tau_r = -3$ to completely remove the paraffin pixels. The binary images resulting from these two thresholdings are visible in Figure 2a,b. The comparison of these figures with the image of unstained and adjacent HE tissue section (Figure 1a) reveals a small correlation, indicating that the paraffin and tissue pixels are badly identified. Indeed, on the one hand, a significant portion of paraffin pixels are misidentified (Figure 2a). On the other hand, a modification of these thresholds resulted in the inverse behavior, i.e., an overidentification of

https://dx.doi.org/10.1021/acs.analchem.0c03910
Anal. Chem. 2021, 93, 3750 3761

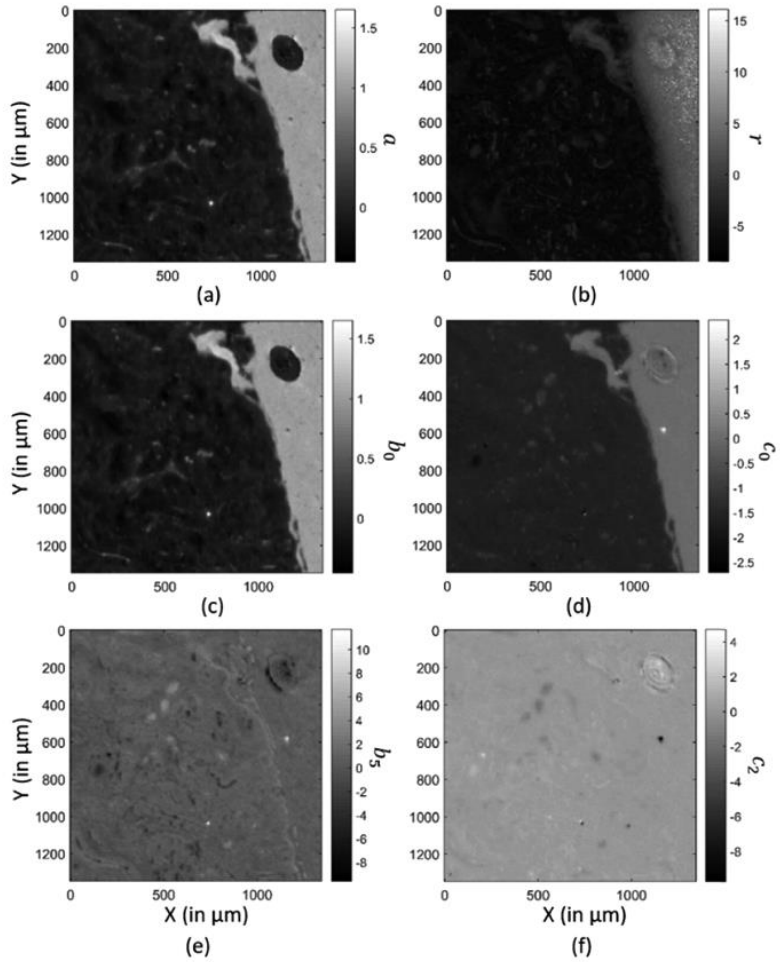


Figure 3. Grayscale images reconstructed from: (a) the regression coefficient a of the reference spectrum, (b) the natural logarithm of the modeling residue r, (c) the regression coefficient b_0 of the mean pure paraffin spectrum, (d) the regression coefficient c_0 corresponding the zero-order polynomial coefficient, (e) the regression coefficient b_5 of the fifth paraffin principal component, and (f) the regression coefficient c_2 corresponding to the second-order polynomial coefficient, estimated by the EMSC model.

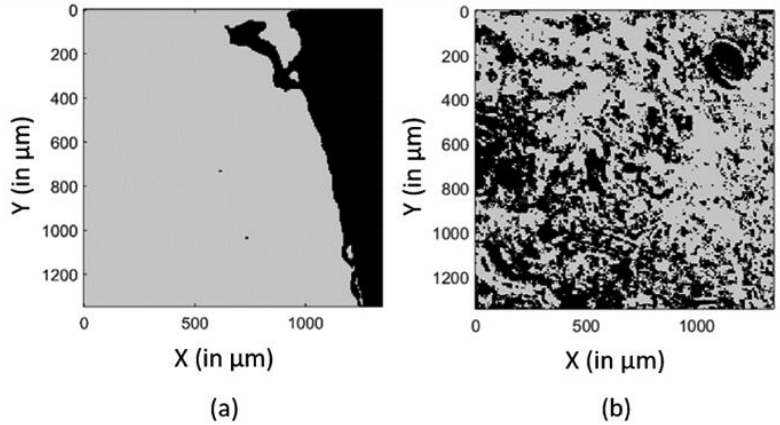


Figure 4. Partitions obtained by applying a two-cluster KMeans on the (a) $\{a, r, b_0, c_0\}$ and (b) $\{b_5, c_2\}$ combinations of EMSC regression coefficients. Black and gray pixels correspond to paraffin and tissue pixels in (a) and to unidentified clusters in (b).

paraffin pixels (Figure 2b). The difference image (Figure 2c) represents the pixels that are differently identified between these two binary images. The threshold is a very sensitive parameter since for this example 6% of pixels are changing of identification. Taken together, all of these results prove that the performance of this type of pure paraffin pixel identification methods based on thresholding is very sensitive to the chosen threshold values. Furthermore, identifying the

https://dx.doi.org/10.1021/acs.analchem.0c03910
Anal. Chem. 2021, 93, 3750 3761

optimal thresholds to completely remove the paraffin pixels is impractical in real situations because it is sample-, image-, and user-dependent.

Toward a Multivariate and Automatic Approach. The two previously presented thresholding methods are based on the analysis of one or two parameters. However, the EMSC model provides many supplementary information about the physical and chemical compositions of the sample at each studied pixel, which can be helpful for the discrimination between paraffin and tissue pixels.

A two-dimensional t-SNE applied on the EMSC regression coefficients of data acquired on a human colon carcinoma FFPE section reveals two clearly visible data groups, suggesting the presence of two different spectral patterns, which can be identified as paraffin and tissue areas, respectively (Figure S-5). Paraffin and tissue areas can thus be separated from the analysis of the EMSC regression coefficients.

To confirm this intuition, images reconstructed from some EMSC regression coefficients are illustrated in Figure 3. A clear contrast between paraffin and tissue areas can be observed using a, r, b_0, and c_0 (Figure 3a−d). This observation is confirmed by the two-cluster KMeans partition estimated on the $\{a, r, b_0, c_0\}$ EMSC regression coefficients (Figure 4a). The use of a and r in the classical univariate analysis method[16] is thus justified, but this method does not exploit all of the available information such as b_0 and c_0, which provide complementary information about paraffin-tissue edges.

Altogether, these results justify the multivariate exploitation of the EMSC regression coefficients, which is the core of our proposed approach.

Furthermore, being based on a simple application of a two-cluster KMeans to separate the pixels into two clusters, one for paraffin and one for tissue, our proposed method is automatic since it does not require the setting of parameters by the operator, in contrast to the previously presented thresholding methods.

However, all of the EMSC regression coefficients are not useful. Distinguishing between paraffin and tissue pixels is difficult, if not impossible on the images reconstructed using b_5 and c_2 (Figure 3e,f). This result is confirmed by the two-cluster KMeans partition estimated using the EMSC regression coefficient combination $\{b_5,c_2\}$ (Figure 4b), which is clearly not correlated to the sample structure (Figure 1a).

The relevant question is thus which combination of EMSC fitting coefficients is optimal for the distinction between tissue and paraffin spectra. In this work, validity indices were used as an objective and quantitative measure to answer this question and propose a fully automatic method.

The development of our proposed method based on the multivariate analysis of the best estimated combination of EMSC regression coefficients determined by validity indices is thus completely justified by these previous results. To objectively study its efficiency, we tested our method on simulated spectral images.

Evaluation of the MA of EMSC Fitting Coefficients on Simulated Spectral Images. To evaluate the performance of our proposed method, a total of 30 simulated spectral images were generated according to the procedure described in the Supporting Information and the parameter setting detailed in the Supporting Information. An example of a simulated spectral image is given in Figure S-6.

The first part of this evaluation consisted of testing the ability of our multivariate approach coupled to validity indices to automatically estimate the optimal EMSC regression coefficient combination, leading to the most compact and separated paraffin and tissue clusters. For this purpose, a noise-free (SNR = 32 dB) spectral image simulated with a first-order polynomial function was used. To give the same weight to all of the EMSC regression coefficients, each one was normalized using the standard normal variate (SNV) method before the application of our procedure. Then, each validity index was applied on the two-cluster partition estimated by the KMeans algorithm applied on each possible combination of EMSC coefficients. For a given validity index, the EMSC coefficient combinations were then ranked according to their computed validity index values (in ascending order for XB and DB, and in descending order for PBM and SWC). For each validity index, the top 5 ranked combinations of EMSC regression coefficients, i.e., giving the best compact and separated clusters, were retained and are given in Table S-1. The four used validity indices give very similar results. However, this information was summarized by considering only the EMSC coefficient combinations identified as belonging to these top 5 ranked combinations by the four validity indices simultaneously. These combinations are named consensual combinations in the remaining of the manuscript. From Table S-1, three consensual combinations are identified, namely, $\{a,b_0\}$, $\{b_0,c_0\}$, $\{a,b_0,c_0\}$. The complete workflow of the proposed method is provided in Figure S-7.

These results are confirmed by the application of t-SNE to the regression coefficients $\{a,b_0,c_0\}$. Indeed, this combination is efficient to distinguish between paraffin and tissue spectra (Figure S-8a). In addition, the correlation of each of these three coefficients to the ground truth classes can also be separately visualized (Figures S-8b−d). Indeed, paraffin spectra are characterized by insignificant contribution of the reference spectrum (a), and by high contribution of paraffin (b_0) and baseline (c_0), while the opposite is observed for the tissue spectra.

These results are in accordance with the way that the model has been generated since the data are composed of three main sources of variability, i.e., the tissue, the paraffin, and the baseline, whose contributions are mainly estimated in the EMSC model by the a, b_0, and c_0 coefficients, respectively. It has to be noted that the identification of the paraffin and tissue pixels by our unsupervised methodology using the $\{a,b_0\}$, $\{b_0,c_0\}$, $\{a,b_0,c_0\}$ combinations is almost perfect since resulting in a Jaccard index around 0.9925. The misidentified pixels are the three with an almost zero contribution of tissue, as explained in the Supporting Information and visible in Figure S-6b,e. On the contrary, the worst combinations identified by the validity indices, namely, $\{c_1\}$, $\{b_1\}$, and $\{b_3\}$, also give the worst supervised Jaccard index of 0.3166, 0.3082, and 0.3098, respectively.

However, it has to be noted that the modeling error residue has not been identified by the validity indices as a parameter relevant for the separation of data into two distinct clusters, which can be justified by the use of a noise-free simulated spectral image, as explained above. Furthermore, a simple first-order polynomial function has been used to model the baseline, while usually a higher-order polynomial function is necessary to estimate the baseline effect on FTIR images acquired on FFPE tissue sections.[12,16−19,21,31,39]

https://dx.doi.org/10.1021/acs.analchem.0c03910
Anal. Chem. 2021, 93, 3750 3761

To complete the characterization of our proposed method, the evaluation of its robustness to perturbations is important. It is thus interesting to study the impact of the SNR and baseline complexity on our complete methodology including the estimation of the optimal EMSC regression coefficient combination by validity indices and Jaccard index. To achieve this goal, different simulated FTIR spectral images were thus generated using a baseline polynomial order varying from 0 to 4, and an SNR varying from 2^0 to 2^5 (see the Supporting Information for more information). For each simulated image and for each validity index, the top 5 combinations of EMSC regression coefficients leading to the best validity index values were determined (data not shown). To summarize these results, Table 1 presents the consensual EMSC regression coefficient combinations estimated by the four validity indices for each polynomial order and SNR couple. To confront these results with the gold standard for each polynomial order and SNR couple, the Jaccard index was computed for each of these consensual combinations. To summarize this information for each polynomial order and SNR couple without overestimating the results of our approach, only the worst consensual coefficient combination, i.e., the one giving the smallest Jaccard index value, was considered. Figure 5a presents these results for all of the studied polynomial order and SNR couples as a map.

Table 1. Consensual EMSC Regression Coefficient Combinations Estimated by the Four Validity Indices on the Simulated Spectral Images as a Function of the Baseline Polynomial Order and the Signal-to-Noise Ratio (SNR) Expressed in Decibels (dB)

polynomial order	SNR					
	32	16	8	4	2	1
0	a, b_0	a, r	a, r	a, r	a, r	a, r
	a, c_0	a, b_0	a, b_0	a, b_0	a, r, b_0	
	a, b_0, c_0	a, r, b_0	a, r, b_0	a, r, b_0		
1	a, b_0	a, b_0	a, r	a, r	a, r	a, r
	b_0, c_0		a, b_0	a, r, b_0		
	a, b_0, c_0		a, r, b_0			
2	a, b_0	a, c_0	a, r	a, r	a, r	a, r
	b_0, c_0		a, c_0	a, c_0	a, c_0	a, c_0
	a, b_0, c_0		a, r, c_0	a, r, c_0	a, r, c_0	a, r, c_0
3	a, b_0	a, b_0	a, r	a, r	a, r	a, r
	b_0, c_0		a, r, c_0		a, r, c_0	a, r, c_0
	a, b_0, c_0					
4	a, b_0	a, c_0	a, c_0	a, c_0	a, c_0	a, c_0
	b_0, c_0		a, r	a, r	a, r	a, r
	a, b_0, c_0					

Globally, the Jaccard index decreases as a function of the baseline polynomial order and of the inverse of the SNR. This result was expected since the identification of paraffin from tissue pixels becomes more difficult when the perturbation sources increase, especially noise. However, regardless of the couples of SNR and baseline polynomial order, the Jaccard index remains very high (over 0.9). Our unsupervised method based on validity indices is thus objectively estimating efficient combinations of EMSC regression coefficients for the separation of tissue and paraffin pixels.

From Table 1, for a high SNR (16 or 32 dB) and regardless of the baseline complexity, the best combinations remain $\{a,b_0,c_0\}$ and $\{a,b_0\}$ as for the previous results obtained for a noise-free model with a first-order polynomial baseline. For increasing noise level (evaluated precisely on unshown results to SNR < 12 dB on the simulated images), the natural logarithm of modeling residue r becomes a relevant parameter, to separate paraffin from tissue pixels, in place of c_0 for low baseline polynomial order (0 or 1), and b_0 for higher baseline polynomial order. These results thus show that the set of EMSC regression coefficients must be adapted as a function of the dataset characteristics. To confirm this result, the Jaccard index was computed for each couple of SNR and baseline polynomial order using only the $\{a,b_0,c_0\}$ combination (Figure 5b). As explained above, the $\{a,b_0,c_0\}$ combination was previously identified as optimal for noise-free spectral image.

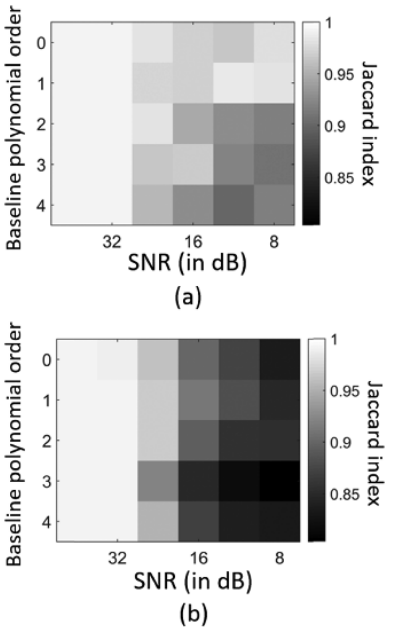


Figure 5. Minimum Jaccard index estimated on simulated spectral images as a function of the baseline polynomial order and the signal-to-noise ratio (SNR) expressed in decibels (dB) (a) for the best combinations of EMSC regression coefficients estimated by validity indices and (b) for the $\{a,b_0,c_0\}$ combination.

By comparison of Figure 5a,b, optimization of the combination of EMSC regression coefficients is thus recommended since it gives better results (Jaccard index over 0.9) than the $\{a,b_0,c_0\}$ combination (Jaccard index over 0.8) regardless of SNR and baseline complexity.

These results are thus in contradiction with the classical method, which is based on the immutable exploitation of the a and r coefficients. The classical method could thus be optimal only for highly noisy datasets. Furthermore, regardless of the baseline complexity and noise level, the EMSC regression coefficient a of the reference spectrum is a necessary parameter to distinguish efficiently the paraffin from tissue pixels. The use of this parameter by the classical method is thus justified. Furthermore, this result showing the importance of this EMSC regression coefficient is in accordance with other studies,[40] where it has been exploited for FTIR image registration.

Altogether, these results prove that our methodology is flexible and adaptive to the main sources of distinction between paraffin and tissue spectra.

https://dx.doi.org/10.1021/acs.analchem.0c03910
Anal. Chem. 2021, 93, 3750 3761

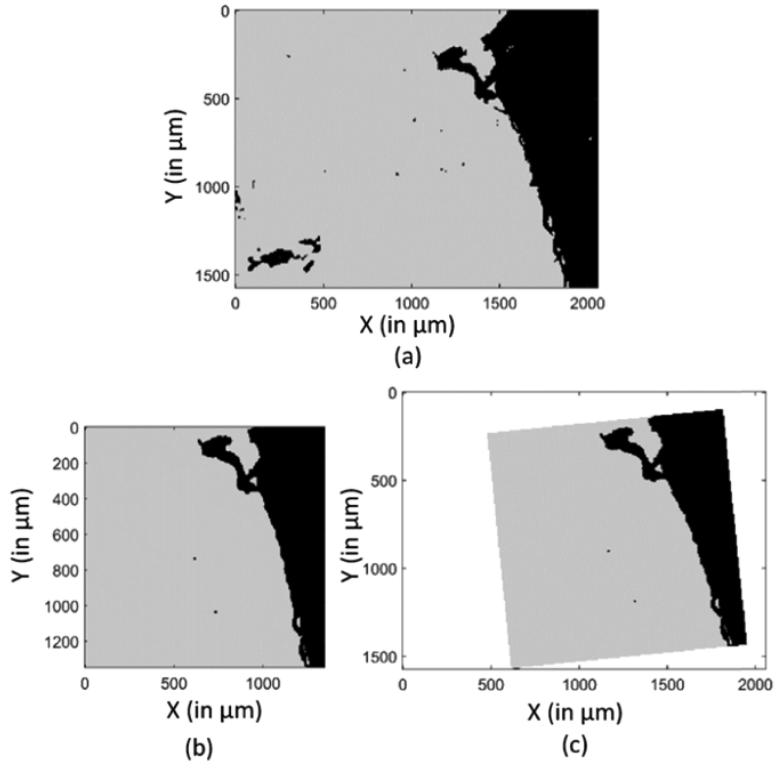


Figure 6. (a) Two-cluster KMeans partition estimated on the raw FTIR image acquired on a chemically dewaxed FFPE human colon carcinoma section. Black and gray pixels correspond to substrate and tissue pixels. (b) Partition obtained by our methodology on the same FFPE sample, before chemical dewaxing. Black and gray pixels correspond to paraffin and tissue pixels. (c) Image (b) after registration using image (a) as the fixed image.

Validation of the MA of EMSC Fitting Coefficients on Real FTIR Spectral Images. The performances of our multivariate approach for the identification of paraffin and tissue pixels have been successfully evaluated on simulated images. The last step consists of the validation of our methodology on real FTIR images acquired on two samples, i.e., human and xenografted FFPE colon carcinoma sections.

As for the simulated data, the first objective of this real data analysis was to determine the best combination of EMSC regression coefficients for the separation of paraffin and tissue pixels. For each image and for each possible combination of coefficients, a two-cluster KMeans partition was estimated and fed the four validity indices presented above. For each validity index, the top 5 combinations of EMSC regression coefficients leading to the best validity index values were determined. The obtained results are consistent with those obtained on the simulated images since the consensual combinations are $\{a,b_0\}$, $\{a,c_0\}$, and $\{a,b_0,c_0\}$. As an example, the complete results estimated on one FTIR image acquired on an FFPE human colon carcinoma sample are given in Table S-2. Extrapolating the interpretation of results obtained on simulated datasets, the obtention of these combinations necessary induces a high SNR of these real data, which is true considering the acquisition setting and the nature of the analyzed samples, and which was confirmed by the SNR of 40 dB measured on this spectral image. The same results have been obtained independently on the other acquired FTIR images (data not shown).

As for the simulated data, these results were confirmed by the application of t-SNE to the regression coefficients $\{a,b_0,c_0\}$. Indeed, the efficiency of our methodology to distinguish between paraffin and tissue pixels is demonstrated using a two-dimensional t-SNE by the estimation of two distinct compact and separated clusters, which are completely correlated with the pixel labels estimated by our method (Figure S-9a). Furthermore, the influence of each of these three coefficients on the paraffin and tissue clusters can also be separately visualized (Figure S-9b–d). As for the simulated images, paraffin spectra are characterized by high contributions of paraffin (b_0) and baseline (c_0), while tissue spectra have a high contribution of the reference spectrum (a). Taken together, these results demonstrate the efficiency of our method on real datasets and also the realism of our simulated data generative model, where the main sources of variability of real FTIR spectral images have been efficiently incorporated.

To properly validate our methodology on real spectral datasets, it is necessary to compare our results with ground truth defined by the chemically dewaxed FFPE tissue sections. The FTIR spectral images were acquired on chemically dewaxed samples and analyzed by a two-cluster KMeans clustering to separate substrate from tissue spectra. An example of such a gold standard clustering image is provided in Figure 6a.

The application of our methodology on the same section before chemical dewaxing determined $\{a,b_0,c_0\}$ as the best EMSC coefficient combination for the separation of paraffin and tissue pixels (Figure 6b). To compensate the orientation and position differences between panels (a) and (b) in Figure 6, the registered version of Figure 6b was computed as stated above. The resulting image (Figure 6c) illustrates the

https://dx.doi.org/10.1021/acs.analchem.0c03910
Anal. Chem. 2021, 93, 3750 3761

efficiency of the used rigid registration algorithm since this image perfectly matches Figure 6a. The paraffin and tissue areas of the studied FFPE tissue section thus visually seem to have been perfectly recognized by our proposed multivariate approach.

To objectively quantify the matching between the gold standard and estimated registered results, the Jaccard index was computed for three patients (representing nine different FTIR images) and two different mice (representing three different FTIR images). As shown in Table 2, the majority of samples have a high Jaccard index (over 0.84) with values slightly inferior to those obtained on the simulated images (around 0.99). This decrease of Jaccard index on the real data can be explained by: (i) the imperfect chemical dewaxing process which is known to be incomplete, aggressive with possible sample deterioration, sensitive to chemical reagents, bath time, and histology of the analyzed tissue region[31] and (ii) the imperfect image registration due to the use of a simple rigid model while chemical dewaxing is well known to induce nonrigid alterations of the tissue topology. The same process, i.e., image registration and Jaccard index computation, has been applied on the same FTIR images using the BR and UA methods. For the vast majority of the considered samples, our proposed method is better than the two classical methods (third and fourth columns of Table 2). Even if the Jaccard index difference between the three methods is small, our method is automatic, contrary to the two classical ones, which require time-consuming manual optimization and an experienced user. Taken together, these results demonstrate the efficiency of our method, as well as its simplicity, flexibility, automation, and potential implementation in clinical routine compared to chemical dewaxing and the classical approaches.

Table 2. Jaccard Index Computed between the Paraffin Pixels Identified by a Two-Cluster KMeans Partition Obtained from an FTIR Image Acquired on a Chemically Dewaxed Section and by One of the Following Identification Method of Pure Paraffin Spectra Applied on an FTIR Image Acquired on an FFPE Section: MA, UA, and BR[a]

	method		
patient and ROI	MA	UA	BR
patient no. 1, ROI no. 1	**0.9855**	0.9796	0.9651
patient no. 1, ROI no. 2	**0.9598**	0.9492	0.9533
patient no. 1, ROI no. 3	**0.9797**	0.9775	0.9706
patient no. 1, ROI no. 4	**0.9564**	0.8268	0.9011
patient no. 2, ROI no. 1	0.8804	0.8668	**0.8988**
patient no. 2, ROI no. 2	**0.9192**	0.9183	0.9160
patient no. 3, ROI no. 1	0.8462	**0.8466**	0.8349
patient no. 3, ROI no. 2	**0.9618**	0.9295	0.9038
patient no. 3, ROI no. 3	**0.9635**	0.8518	0.8566
mouse no. 1, ROI no. 1	**0.8454**	0.7944	0.7760
mouse no. 1, ROI no. 2	**0.8733**	0.8613	0.7281
mouse no. 2, ROI no. 1	**0.9417**	0.9323	0.9389

[a]Several regions of interest (ROI) were studied for three patients and two mice. For each line, the bold value indicates the best result, thus the best method.

Methodological Discussion. Our methodology could also be applied to construct a specific database dedicated to the training of a supervised classification model for distinguishing paraffin and tissue pixels. The EMSC regression coefficients combination estimated as the best by our method to identify tissue from nontissue pixels could allow a robust calibration of the classification model.

It has to be noted that the registration algorithm has been applied on binary images because this algorithm requires grayscale images in inputs and these binary images are directly the results of the investigated pure paraffin pixel identification methods. However, using this strategy, the pixel identification errors may influence the registration results. In our case, this property could be in fact an advantage since the errors made by the different pure pixel identification methods will be amplified. Thus, a method making more identification errors than another will give a worse registered binary image and thus a worse Jaccard index. However, the high Jaccard indices presented in Table 2, and their small differences between methods prove that these errors have little influence on the registration results. However, it should be interesting to deeply study the impact of the data representation (other than binary images), of propagation of identification errors, and of model complexity on the registration of FTIR images acquired on FFPE tissue sections, which is beyond the scope of this article.

It must be noted that optimization of the combination of EMSC regression coefficients can be time-consuming for numerous and high-dimensional FTIR spectral images. For example, for a real FTIR spectral image composed of 46 656 spectra with 451 wavenumbers per spectrum, the exhaustive optimization over all EMSC coefficient combinations and using four validity indices has taken 76 h using a computer equipped with a 3.4 GHz Intel Core i7-4770 CPU, 16 Go RAM, and 4 cores. However, 87% of this computational time is due to DB and PBM. Considering only XB and SWC or using other fast-to-compute validity indices results in a drastic reduction of this computational time.

Furthermore, the performance of our unsupervised methodology remains relatively stable as a function of the SNR and baseline polynomial order on the simulated images for a fixed combination of EMSC regression coefficients since the Jaccard index is over 0.8 (Figure 5b). Consequently, for the processing of FTIR spectral images acquired almost in the same conditions, i.e., on tissue sections prepared according to the same standardized protocol of paraffin embedding using the same instrument with the same acquisition parameters, a good practice should thus be to optimize the combination on a unique image or on simulated data mimicking real data, and then to use the same combination on all of the other images.

Furthermore, our results on real and simulated FTIR images show that the optimized EMSC coefficient combinations are always a subset of $\{a,b_0,c_0,r\}$. In real applications, another good practice should be to perform the optimization considering only the 15 possible subsets of $\{a,b_0,c_0,r\}$, i.e., $\{a\}$, $\{b_0\}$, $\{c_0\}$, $\{r\}$, $\{a,b_0\}$, $\{a,c_0\}$, $\{a,r\}$, $\{b_0,c_0\}$, $\{b_0,r\}$, $\{c_0,r\}$, $\{a,b_0,c_0\}$, $\{a,b_0,r\}$, $\{a,c_0,r\}$, $\{b_0,c_0,r\}$, $\{a,b_0,c_0,r\}$. For example, for the same real FTIR image (46 656 spectra), using the four validity indices, this optimization has drastically reduced the computational time to 45 s.

The performance of KMeans algorithm is well known to be initialization-dependent. A common practice is thus to apply KMeans several times on the same dataset to maximize the chances to converge to the global minimum of the KMeans objective function. The more the searched clusters and the higher the complexity of the dataset (i.e., the number of dimensions and cluster overlapping), the more the necessity

https://dx.doi.org/10.1021/acs.analchem.0c03910
Anal. Chem. 2021, 93, 3750 3761

to repeat KMeans clustering. However, in our proposed methodology, the number of clusters is as small as possible, i.e., two, with one for paraffin pixels and one for the tissue pixels. Furthermore, the use of only few EMSC coefficients drastically reduces the complexity of data processed by KMeans, as revealed by the scatter plots of two-dimensional t-SNE applied on EMSC coefficients estimated from simulated and real datasets (Figures S-5a, S-8a, and S-9a). In the proposed methodology, repetition of KMeans algorithm was thus not considered as necessary and was thus avoided to reduce the computational time as much as possible. This choice was validated by the similarity of the results obtained on the simulated data and on the 12 different real images.

CONCLUSIONS

For complete histopathological characterization of tissue samples by FTIR imaging, the tissue area must be identified as precisely as possible. However, on FFPE tissue sections, the strong infrared signature of paraffin complicates this task by blurring the frontier between tissue and paraffin. So far, the solutions proposed in the literature were based on the subjective and manual choice of thresholds from univariate histogram analysis of various quantities measured from the recorded spectra, leading to highly variable results between different operators. In this article, we proposed a new simple, objective, and automatic methodology based on the multivariate exploitation of EMSC fitting coefficients. Using t-SNE, validity indices, and Jaccard index on simulated and real datasets, we demonstrated the efficiency of our methodology to automatically determine the best EMSC fitting coefficients for the separation of paraffin and tissue pixels on infrared images simulated or acquired on metastatic and xenografted human colon cancer FFPE tissues. Mainly, the high similarity between FFPE tissue sections and their chemically dewaxed versions validates and confirms the efficiency of our approach. Thus, this work confirms the efficiency and versatility of EMSC for infrared images acquired on FFPE samples since it can neutralize and normalize paraffin and baseline on tissue spectra, and combinations of its estimated regression coefficients enhance information necessary to easily distinguish paraffin from tissue pixels.

ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.analchem.0c03910.

Description of the generative model of simulated spectral images and its setting; spectral band ratio computation (Figure S1); examples of components of the EMSC model (Figure S2); examples of application of the EMSC model (Figure S3); estimated probability densities of two EMSC regression coefficients (Figure S4); results of t-SNE applied on the EMSC regression coefficients of data acquired on a human colon carcinoma FFPE section (Figure S5); example of a simulated FTIR image (Figure S6); workflow of the proposed method (Figure S7); results of t-SNE applied to the optimal EMSC regression coefficients of simulated (Figure S8) and real (Figure S9) FTIR images; and top 5 combinations of EMSC regression coefficients of simulated (Table S1) and real (Table S2) FTIR images (PDF)

AUTHOR INFORMATION

Corresponding Author

Cyril Gobinet − *BioSpec'I' EA 7506, Université de Reims Champagne Ardenne, 51097 Reims, France;* Email: cyril.gobinet@univ-reims.fr

Authors

Warda Boutegrabet − *Institut National de la Santé et de la Recherche Médicale, IRFAC Inserm U1113, Université de Strasbourg (Unistra), 67200 Strasbourg, France; BioSpec'I' EA 7506, Université de Reims Champagne Ardenne, 51097 Reims, France;* orcid.org/0000-0002-6340-5967

Dominique Guenot − *Institut National de la Santé et de la Recherche Médicale, IRFAC Inserm U1113, Université de Strasbourg (Unistra), 67200 Strasbourg, France*

Olivier Bouché − *BioSpecT EA 7506, Université de Reims Champagne Ardenne, 51097 Reims, France; Hepato-Gastroenterology Department, CHU de Reims, 51092 Reims, France*

Camille Boulagnon-Rombi − *MEDyC CNRS UMR 7369, Université de Reims Champagne Ardenne, 51097 Reims, France; Biopathology Laboratory, CHU de Reims, 51092 Reims, France*

Aude Marchal Bressenot − *BioSpecT EA 7506, Université de Reims Champagne Ardenne, 51097 Reims, France; Biopathology Laboratory, CHU de Reims, 51092 Reims, France*

Olivier Piot − *BioSpecT EA 7506, Université de Reims Champagne Ardenne, 51097 Reims, France; Platform of Cellular and Tissular Imaging (PICT), 51097 Reims, France*

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.analchem.0c03910

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

The authors thank the University of Strasbourg and the Région Grand Est for financial support, the Platform of Cellular and Tissular Imaging (PICT) at University of Reims Champagne-Ardenne, as well as Nicole Bouland and Elisabeth Martin for technical support.

REFERENCES

(1) Hermes, M.; Morrish, R. B.; Huot, L.; Meng, L.; Junaid, S.; Tomko, J.; Lloyd, G. R.; Masselink, W. T.; Tidemand-Lichtenberg, P.; Pedersen, C.; Palombo, F.; Stone, N. *J. Opt.* **2018**, *20*, No. 023002.

(2) Ly, E.; Piot, O.; Durlach, A.; Bernard, P.; Manfait, M. *Analyst* **2009**, *134*, 1208−1214.

(3) Ly, E.; Cardot-Leccia, N.; Ortonne, J.-P.; Benchetrit, M.; Michiels, J.-F.; Manfait, M.; Piot, O. *Br. J. Dermatol.* **2010**, *162*, 1316−1323.

(4) Pezzei, C.; Pallua, J. D.; Schaefer, G.; Seifarth, C.; Huck-Pezzei, V.; Bittner, L. K.; Klocker, H.; Bartsch, G.; Bonn, G. K.; Huck, C. W. *Mol. BioSyst.* **2010**, *6*, 2287−2295.

(5) Akalin, A.; Mu, X.; Kon, M. A.; Ergin, A.; Remiszewski, S. H.; Thompson, C. M.; Raz, D. J.; Diem, M. *Lab. Invest.* **2015**, *95*, 406−421.

(6) Steller, W.; Einenkel, J.; Horn, L.-C.; Braumann, U.-D.; Binder, H.; Salzer, R.; Krafft, C. *Anal. Bioanal. Chem.* **2006**, *384*, 145−154.

(7) Krafft, C.; Thümmler, K.; Sobottka, S. B.; Schackert, G.; Salzer, R. *Biopolymers* **2006**, *82*, 301−305.

(8) Benard, A.; Desmedt, C.; Smolina, M.; Szternfeld, P.; Verdonck, M.; Rouas, G.; Kheddoumi, N.; Rothé, F.; Larsimont, D.; Sotiriou, C.; Goormaghtigh, E. *Analyst* **2014**, *139*, 1044−1056.
(9) Pounder, F. N.; Reddy, R. K.; Bhargava, R. *Faraday Discuss.* **2016**, *187*, 43−68.
(10) Khanmohammadi, M.; Bagheri Garmarudi, A.; Samani, S.; Ghasemi, K.; Ashuri, A. *Pathol. Oncol. Res.* **2011**, *17*, 435−441.
(11) Lasch, P.; Haensch, W.; Naumann, D.; Diem, M. *Biochim. Biophys. Acta, Mol. Basis Dis.* **2004**, *1688*, 176−186.
(12) Nallala, J.; Diebold, M.-D.; Gobinet, C.; Bouché, O.; Sockalingum, G. D.; Piot, O.; Manfait, M. *Analyst* **2014**, *139*, 4005−4015.
(13) Kallenbach-Thieltges, A.; Großerüschkamp, F.; Mosig, A.; Diem, M.; Tannapfel, A.; Gerwert, K. *J. Biophotonics* **2013**, *6*, 88−100.
(14) Vuiblet, V.; Fere, M.; Gobinet, C.; Birembaut, P.; Piot, O.; Rieu, P. *J. Am. Soc. Nephrol.* **2016**, *27*, 2382−2391.
(15) Eklouh-Molinier, C.; Happillon, T.; Bouland, N.; Fichel, C.; Diébold, M.-D.; Angiboust, J.-F.; Manfait, M.; Brassart-Pasco, S.; Piot, O. *Analyst* **2015**, *140*, 6260−6268.
(16) Ly, E.; Piot, O.; Wolthuis, R.; Durlach, A.; Bernard, P.; Manfait, M. *Analyst* **2008**, *133*, 197−205.
(17) Nguyen, T. N. Q.; Jeannesson, P.; Groh, A.; Piot, O.; Guenot, D.; Gobinet, C. *J. Biophotonics* **2016**, *9*, 521−532.
(18) Nguyen, T. N. Q.; Jeannesson, P.; Groh, A.; Guenot, D.; Gobinet, C. *Analyst* **2015**, *140*, 2439−2448.
(19) Nallala, J.; Gobinet, C.; Diebold, M. D.; Untereiner, V.; Bouché, O.; Manfait, M.; Sockalingum, G. D.; Piot, O. *J. Biomed. Opt.* **2012**, *17*, No. 116013.
(20) Wolthuis, R.; Travo, A.; Nicolet, C.; Neuville, A.; Gaub, M.-P.; Guenot, D.; Ly, E.; Manfait, M.; Jeannesson, P.; Piot, O. *Anal. Chem.* **2008**, *80*, 8461−8469.
(21) de Lima, F. A.; Gobinet, C.; Sockalingum, G. D.; Garcia, S. B.; Manfait, M.; Untereiner, V.; Piot, O.; Bachmann, L. *Analyst.* **2017**, *142*, 1358−1370.
(22) Afseth, N. K.; Kohler, A. *Chemom. Intell. Lab. Syst.* **2012**, *117*, 92−99.
(23) Patel, I. I.; Trevisan, J.; Singh, P. B.; Nicholson, C. M.; Krishnan, R. K. G.; Matanhelia, S. S.; Martin, F. L. *Anal. Bioanal. Chem.* **2011**, *401*, 969−982.
(24) Ali, S. M.; Bonnier, F.; Lambkin, H.; Flynn, K.; McDonagh, V.; Healy, C.; Lee, T. C.; Lyng, F. M.; Byrne, H. J. *Anal. Methods* **2013**, *5*, 2281−2291.
(25) Fernandez, D. C.; Bhargava, R.; Hewitt, S. M.; Levin, I. W. *Nat. Biotechnol.* **2005**, *23*, 469−474.
(26) Pallua, J. D.; Pezzei, C.; Zelger, B.; Schaefer, G.; Bittner, L. K.; Huck-Pezzei, V. A.; Schoenbichler, S. A.; Hahn, H.; Kloss-Brandstaetter, A.; Kloss, F.; Bonn, G. K.; Huck, C. W. *Analyst* **2012**, *137*, 3965−3974.
(27) Travo, A.; Piot, O.; Wolthuis, R.; Gobinet, C.; Manfait, M.; Bara, J.; Forgue-Laffite, M.-E.; Jeannesson, P. *Histopathology* **2010**, *56*, 921−931.
(28) Kéry, M.; Royle, J. A. *Appl. Hierarchical Model. Ecol.* **2016**, *1*, 123−143.
(29) Lasch, P. *Chemom. Intell. Lab. Syst.* **2012**, *117*, 100−114.
(30) Lieber, C. A.; Mahadevan-Jansen, A. *Appl. Spectrosc.* **2003**, *57*, 1363−1367.
(31) Nallala, J.; Lloyd, G. R.; Stone, N. *Analyst* **2015**, *140*, 2369−2375.
(32) MacQueen, J. In *Some Methods for Classification and Analysis of Multivariate Observations*, Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, 1967; pp 281−297.
(33) Arbelaitz, O.; Gurrutxaga, I.; Muguerza, J.; Pérez, J. M.; Perona, I. *Pattern Recognit.* **2013**, *46*, 243−256.
(34) Oliveira, F. P. M.; Tavares, J. M. R. S. *Comput. Methods Biomech. Biomed. Eng.* **2014**, *17*, 73−93.
(35) van der Maaten, L.; Hinton, G. *J. Mach. Learn. Res.* **2008**, *9*, 2579−2605.
(36) Cao, Y.; Wang, L. Automatic Selection of t-SNE Perplexity. 2017, arXiv:1708.03229. arXiv.org e-Print archive. https://arxiv.org/abs/1708.03229.
(37) Vorontsov, I. E.; Kulakovskiy, I. V.; Makeev, V. J. *Algorithms Mol. Biol.* **2013**, *8*, No. 23.
(38) Hill, P. D. *Commun. Stat. − Theory Methods* **1985**, *14*, 605−620.
(39) Nallala, J.; Lloyd, G. R.; Hermes, M.; Shepherd, N.; Stone, N. *Vib. Spectrosc.* **2017**, *91*, 83−91.
(40) Trukhan, S.; Tafintseva, V.; Tøndel, K.; Großerueschkamp, F.; Mosig, A.; Kovalev, V.; Gerwert, K.; Kohler, A. *J. Biophotonics* **2020**, No. e201960223.

https://dx.doi.org/10.1021/acs.analchem.0c03910
Anal. Chem. 2021, 93, 3750 3761

Supporting Information

Automatic identification of paraffin pixels on FTIR images acquired on FFPE human samples.

Warda Boutegrabet[1,2], Dominique Guenot[1], Olivier Bouché[2,3], Camille Boulagnon-Rombi[4,5], Aude Marchal Bressenot[2,5], Olivier Piot[2,6], Cyril Gobinet[2,*]

[1]Université de Strasbourg (Unistra), Institut National de la Santé et de la Recherche Médicale, IRFAC Inserm U1113, 3 avenue Molière, 67200 Strasbourg, France

[2]Université de Reims Champagne Ardenne, BioSpecT EA 7506, 51 rue Cognacq-Jay, 51097 Reims, France

[3]CHU de Reims, Hepato-Gastroenterology Department, rue du Général Koenig, 51092 Reims, France

[4]Université de Reims Champagne Ardenne, CNRS, MEDyC UMR 7369, 51 rue Cognacq-Jay, 51097 Reims, France

[5]CHU de Reims, Biopathology Laboratory, rue du Général Koenig, 51092 Reims, France

[6]Platform of Cellular and Tissular Imaging (PICT), 51 rue Cognacq-Jay, 51097 Reims, France

*corresponding author: cyril.gobinet@univ-reims.fr

Contents

1) Description of the generative model of simulated spectral images

In order to validate our approach, we constructed simulated FTIR spectral images of FFPE tissue sections where the i[th] spectrum $\boldsymbol{s_i}$ is modelled using the following linear model:

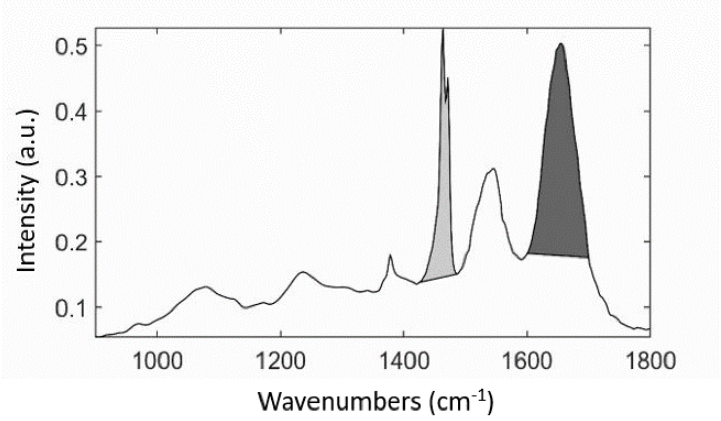
$$\boldsymbol{s_i} = \alpha_i \boldsymbol{t_i} + \beta_i \boldsymbol{p_i} + \boldsymbol{l_i} + \sigma \boldsymbol{n_i} \quad (1)$$

$\boldsymbol{t_i}$ is a spectrum randomly selected in a database composed of 2878 real FTIR spectra acquired in the tissue area of a frozen section from a xenografted human colon carcinoma. In this database, the spectra were previously corrected from the baseline using the Lieber-Mahadevan-Jansen polynomial method[30] using a 4th order polynomial function to model the baseline and normalized by Min-Max normalization. $\boldsymbol{p_i}$ is a spectrum randomly selected in a database composed of 12544 real FTIR spectra acquired in a pure paraffin area at the periphery of a human colon carcinoma FFPE section. Pure paraffin FTIR spectra being almost not distorted by a baseline, the spectra composing this database were not previously corrected from it. $\boldsymbol{l_i}$ is a baseline randomly selected in a database composed of 10905 baselines estimated by the Lieber-Mahadevan-Jansen polynomial method[30] on real FTIR spectra acquired in the tissue area of a human colon carcinoma FFPE section. $\boldsymbol{n_i}$ is a random noise vector generated using a standard normal distribution, i.e. with zero mean and unit standard deviation. α_i and β_i are the contributions of tissue and paraffin spectra respectively, which were generated by respecting the following constraints in order to approximate the topography of a real FFPE tissue section. First, $0 \leq \alpha_i \leq 1$, $0.5 \leq \beta_i \leq 1$ and $\beta_i = 1 - \alpha_i/2$ in order to limit the range of spectral intensities and to ensure the presence of a paraffin signal in each simulated spectrum. Second, the simulated FFPE tissue area is surrounded by a simulated pure paraffin area for which $\beta_i = 1$, and thus $\alpha_i = 0$ according to the second constraint. Third, the α_i coefficients in the tissue part of the simulated images were generated using the distribution of the fitting coefficient of the mean spectrum of a real human colon carcinoma FFPE section as estimated by the EMSC method (section entitled "Data pre-processing"). Fourth, these simulated α_i coefficients were sorted in ascending order, with the smallest values on the outer part of the tissue area and the highest on the inner part, in order to create a gradient of tissue. σ is the standard deviation of the Gaussian noise used to control the signal to noise ratio.

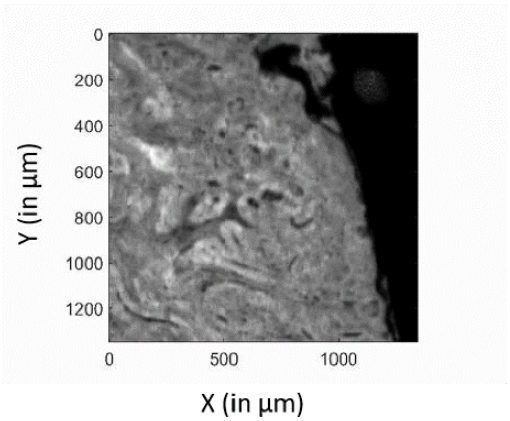
2) Parameter setting of the generative model of simulated spectral images

In order to evaluate the performance of our method described in section entitled "Multivariate analysis of EMSC fitting coefficients", a total of 30 simulated spectral images were generated according to the procedure described in the previous section and using the following parameter setting. The simulated FFPE tissue section acquisition area corresponds to a square with 30-pixel sides, including a tissue area represented by an internal square with 20-pixel sides. The remaining outer pixels are thus paraffin pixels. This simulated sample topology is considered as the ground truth labels for the simulated spectra (Figure S-6(a)). Following this sample topology, the contributions of tissue and paraffin spectra were generated respecting the four constraints (Figure S-6(b-c)). More precisely, the distribution of the EMSC regression coefficient α (Figure S-6(d)) originating from a real FFPE tissue section was used to simulate the tissue contribution respecting the third constraint. Depending on the experiment, the order of the polynomial function varied between 0 and 4 to simulate the baseline, and the SNR varied between 1 and 1000. As examples, Figures S-6(e-g) show some simulated spectra at different locations in the tissue area using a noise-free simulation model with a first-order polynomial function for the baseline. A wide variability in paraffin, tissue and baseline contributions is clearly visible on these spectra. In particular, Figure S-6(e) shows the spectra simulated for the three pixels labelled as tissue pixels (rectangle A on Figure S-6(b)), but with a very weak contribution of tissue ($\alpha < 0.1$ and $\beta > 0.95$ in the simulated model). All the simulated spectra are depicted on Figure S-6(h).

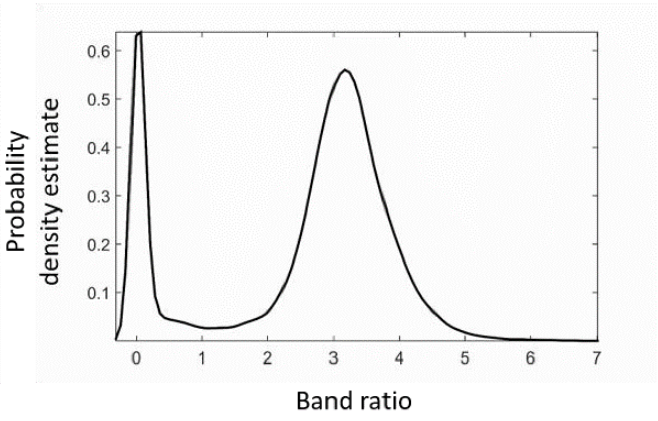
3) Supplementary figures



(a)



(b)



(c)

Figure S-1: Spectral band ratio computation. (a) Mean spectrum of a spectral image acquired on an FFPE tissue section. The light and dark gray areas represent the integrated paraffin (1430–1490 cm^{-1}) and Amide I (1600–1700 cm^{-1}) bands, respectively, used to compute the spectral band ratio. (b) Grayscale image reconstructed from the computed spectral band ratio values. (c) Estimated probability density of spectral band ratio.

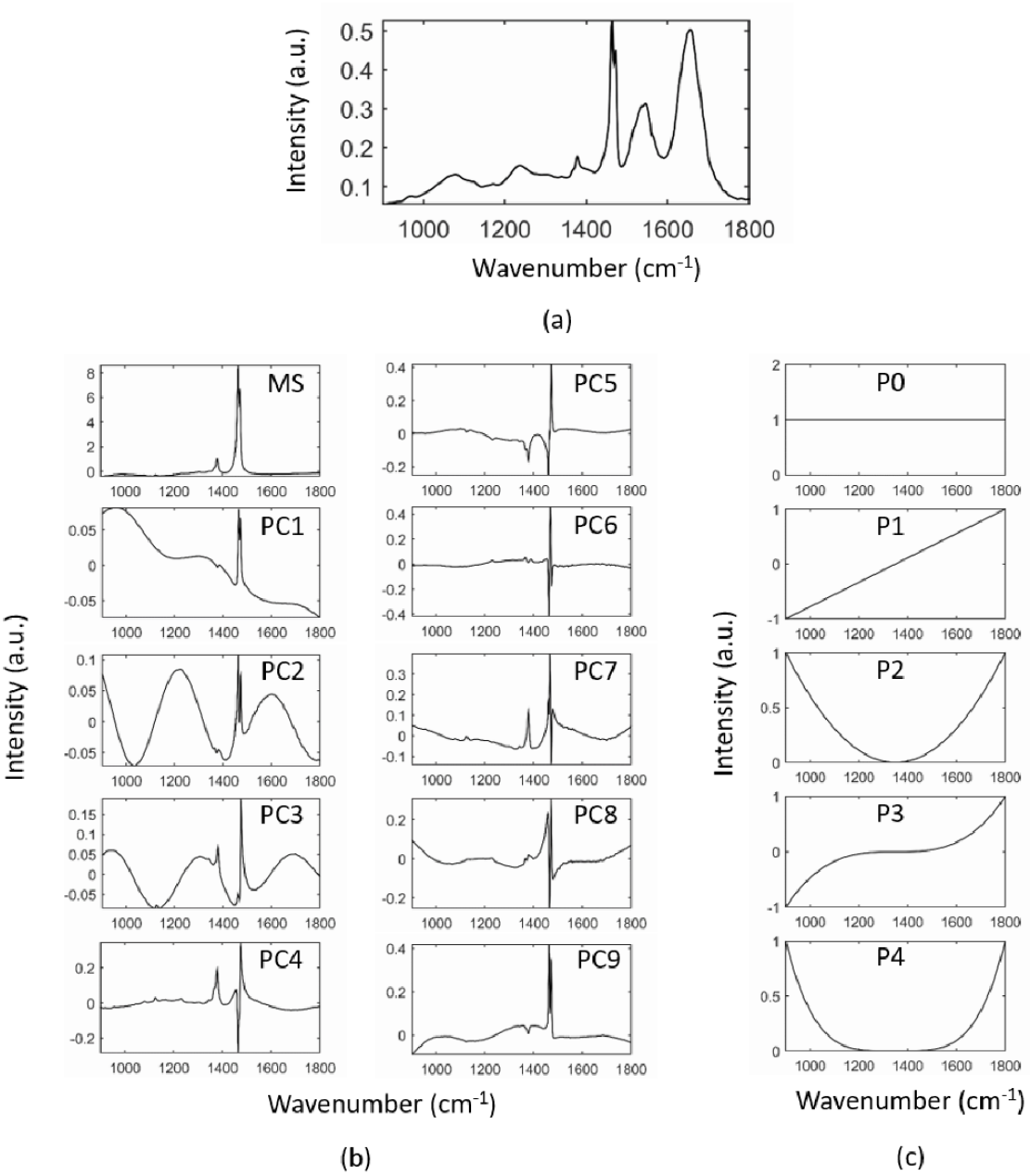


Figure S-2: Components of the EMSC model applied on a FTIR spectral image acquired on a human colon carcinoma FFPE sample. (a) The reference spectrum $\hat{\boldsymbol{s}}$ computed as the mean image spectrum. (b) The interference matrix $\boldsymbol{I}$ composed of the mean spectrum (MS) of a FTIR spectral image acquired on a pure paraffin area, and of the 9 first principal components (PC1 to PC9) computed on this pure paraffin spectral image in order to model the spectral variabilities of paraffin, such as maximum peak position, peak width, etc. (c) The Vandermonde matrix $\boldsymbol{P}$ composed of polynomial functions of order 0 to 4 (P0 to P4).

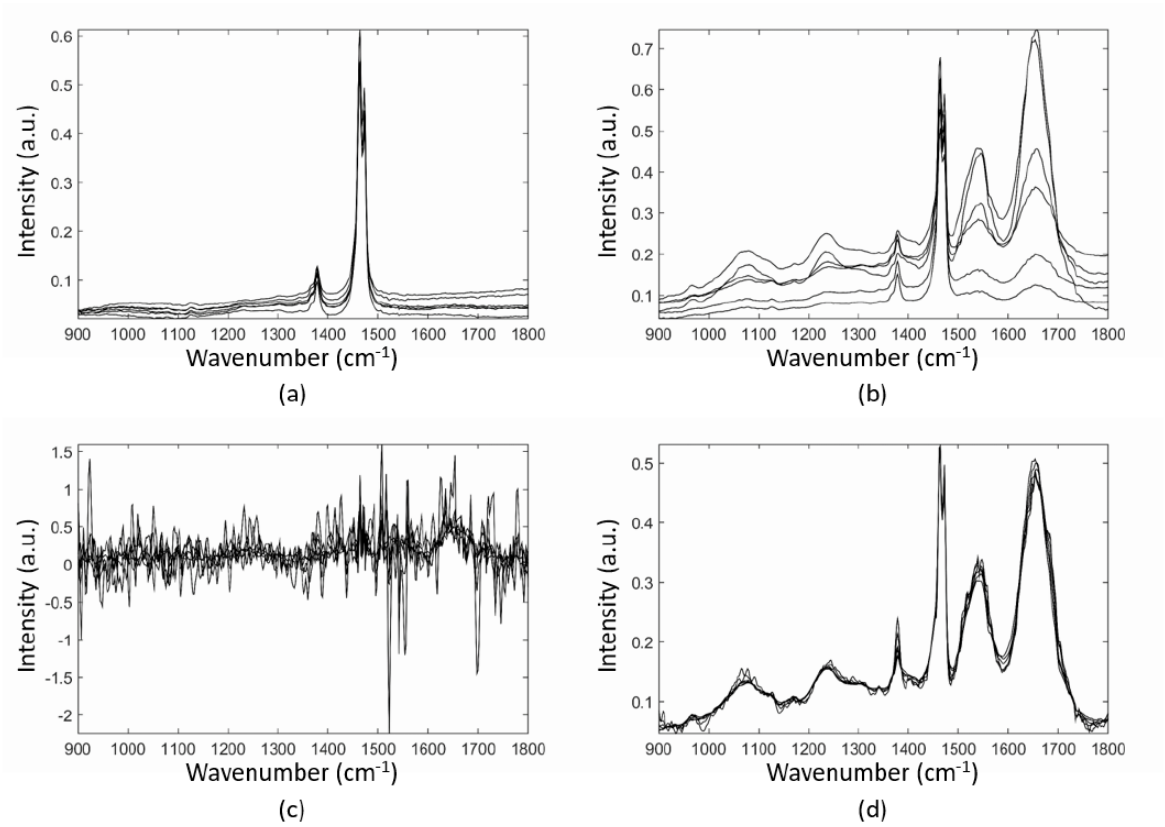
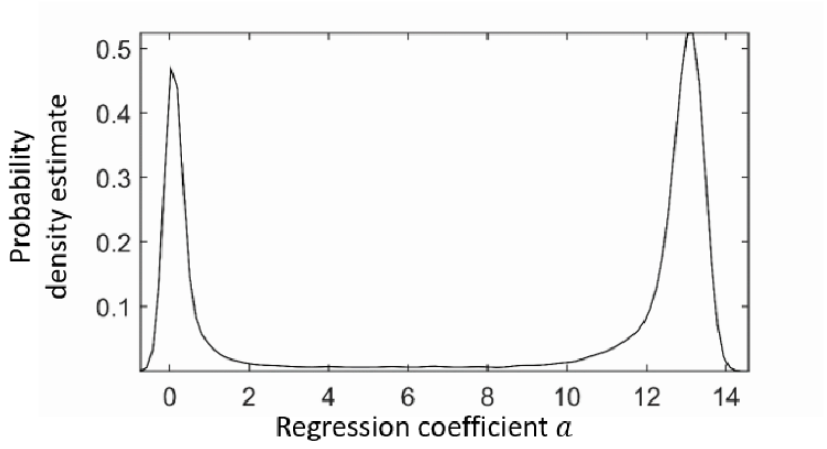
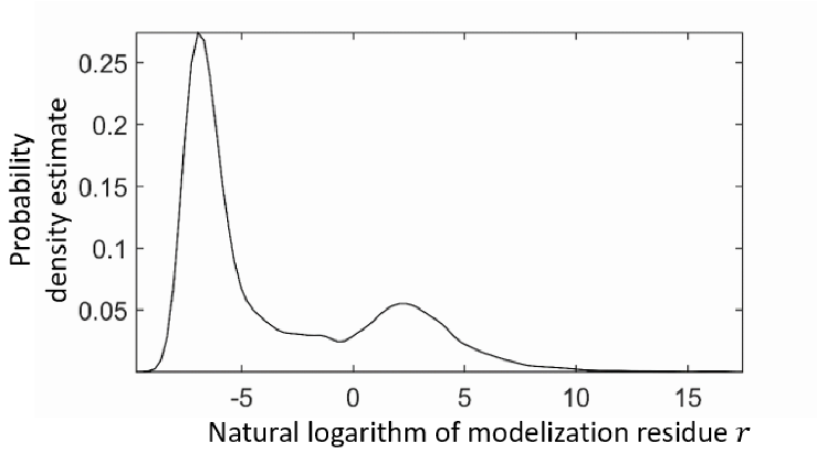


Figure S-3: Examples of application of the EMSC model to the FTIR spectral image acquired on a human colon carcinoma FFPE sample used throughout the paper as an illustrative example. Examples of raw spectra acquired on the paraffin (a) and FFPE tissue (b) areas of this sample. The same spectra after EMSC pre-processing, i.e. EMSC pre-processed spectra of paraffin (c) and FFPE tissue (d).

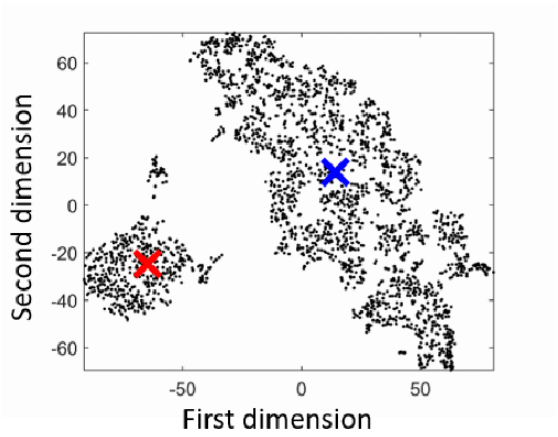


(a)

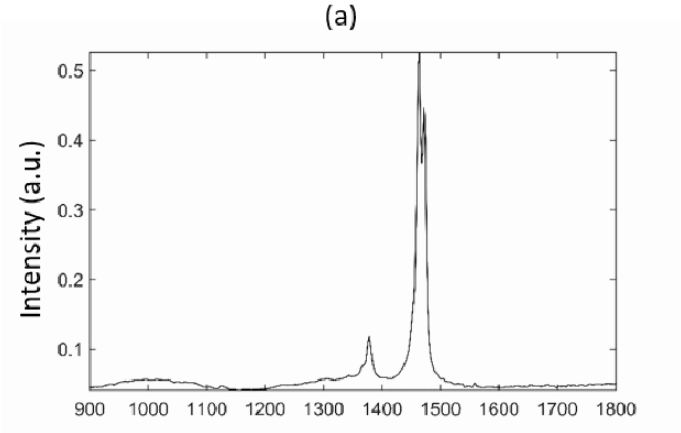


(b)

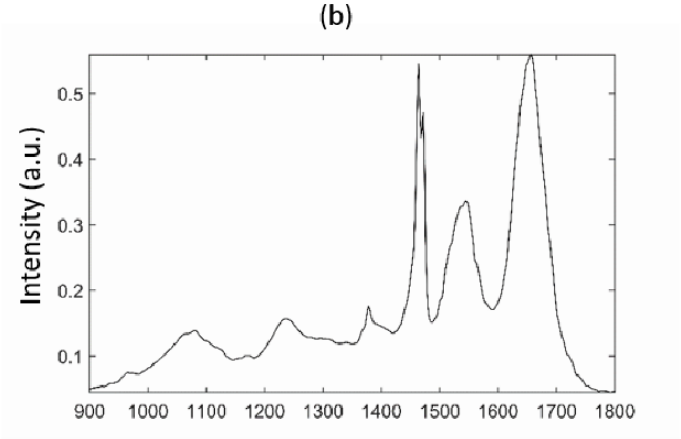
Figure S-4: The estimated probability densities of the regression coefficient a (a) and the natural logarithm of modelization residue r (b).



(a)



(b)



(c)

Figure S-5: Two data groups as revealed by t-SNE applied on the EMSC regression coefficients of data acquired on a human colon carcinoma FFPE section. (a) Scatter plot of a two-dimensional t-SNE. The cluster visible at the bottom left of the figure corresponds to the pure paraffin spectra, while the one at the top and at the right is typical of the tissue spectra. (b) The paraffin spectrum corresponding to the t-SNE point as marked by a red cross on (a). (c) The FFPE tissue spectrum corresponding to the t-SNE point as marked by a blue cross on (a).

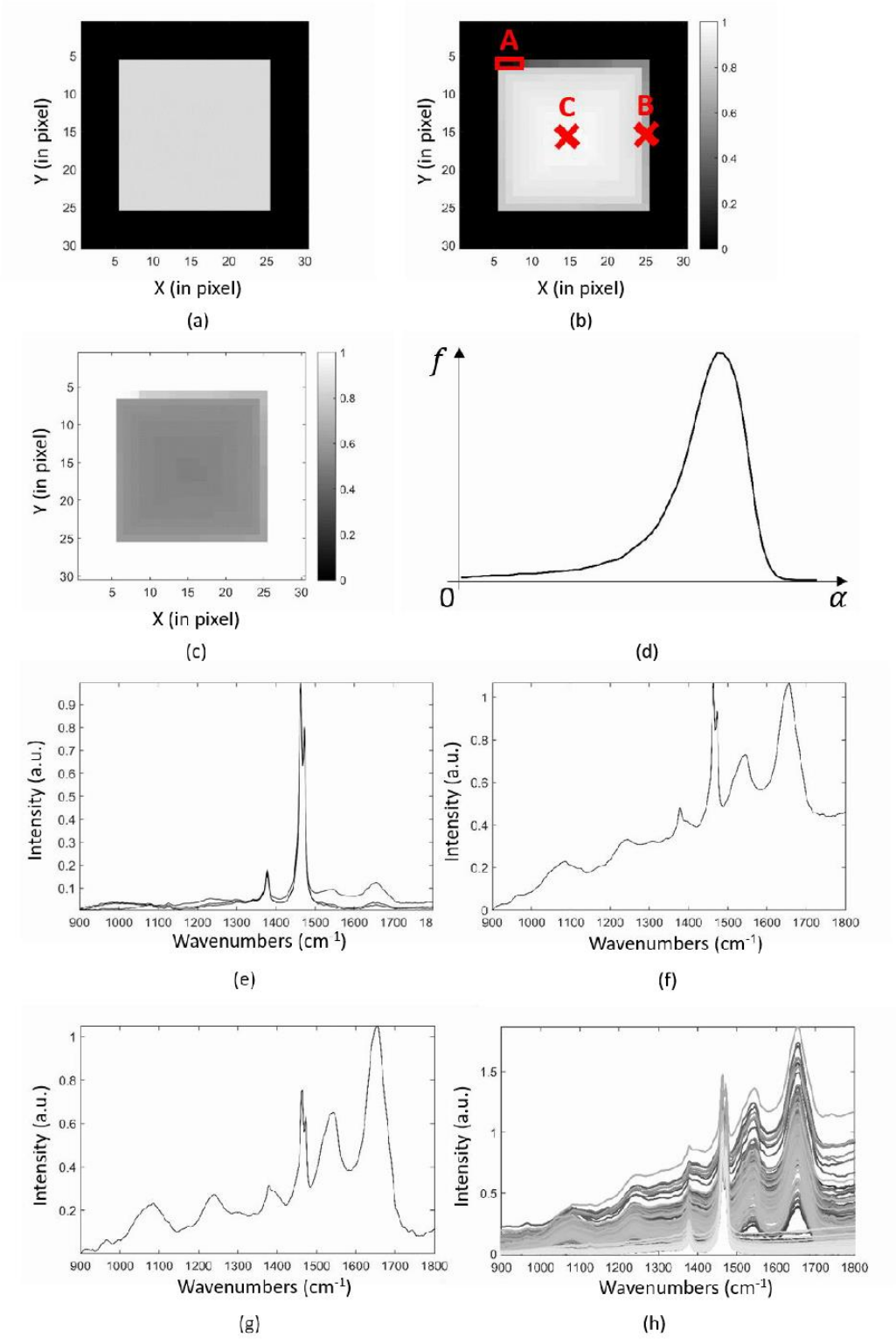


Figure S-6: An example of a simulated FTIR image acquired on a FFPE tissue section. (a) Topology of the simulated sample where black and gray pixels represent paraffin and tissue pixels respectively. (b) The spatial map of the simulated tissue concentration coefficients. (c) The spatial map of the simulated paraffin concentration coefficients. Note that on panels (a-c) a pure paraffin area is surrounding a FFPE tissue part. (d) A real distribution used to simulate the tissue concentration coefficients. (e-g) Examples of simulated spectra of pixels annotated A, B and C on panel (b). (h) The complete simulated dataset.

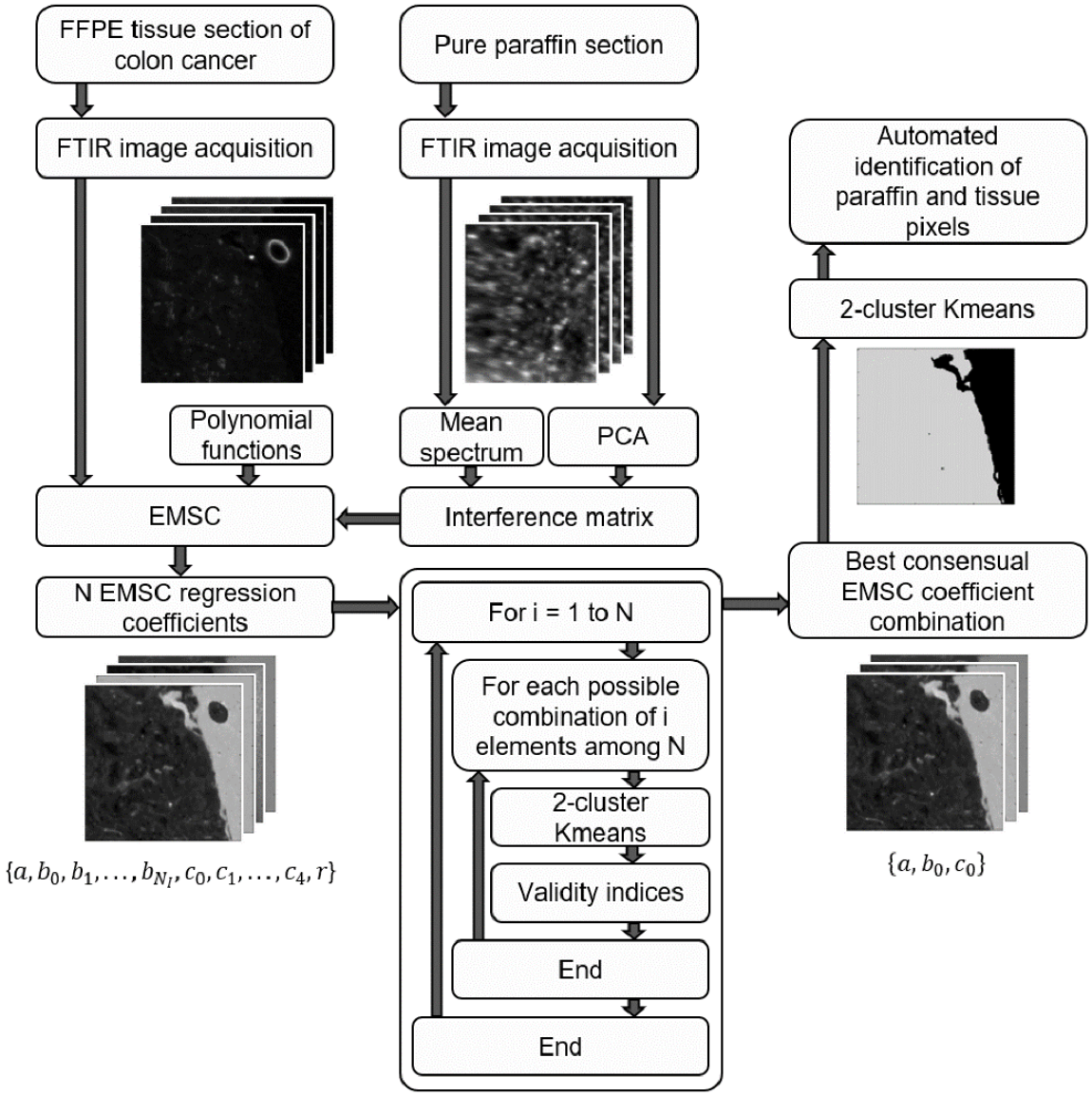


Figure S-7: Workflow of the proposed method. A FTIR image is acquired on a pure paraffin section in order to compute the mean paraffin spectrum and the principal components loadings which describe the most important sources of variation in the paraffin spectra. These components are injected into the interference matrix of the EMSC model. An example of these components if provided on Figure S-2(b). Then, a FTIR image is acquired on a FFPE tissue section which is preprocessed by EMSC as described in the experimental section. The EMSC regression coefficients are injected into an algorithm in order to determine, for each possible combination of these coefficients, the values of validity indices applied on 2-cluster KMeans partitions. The combination leading to the best consensual value of validity indices is used to run a 2-cluster KMeans in order to automatically identify the paraffin and tissue pixels.

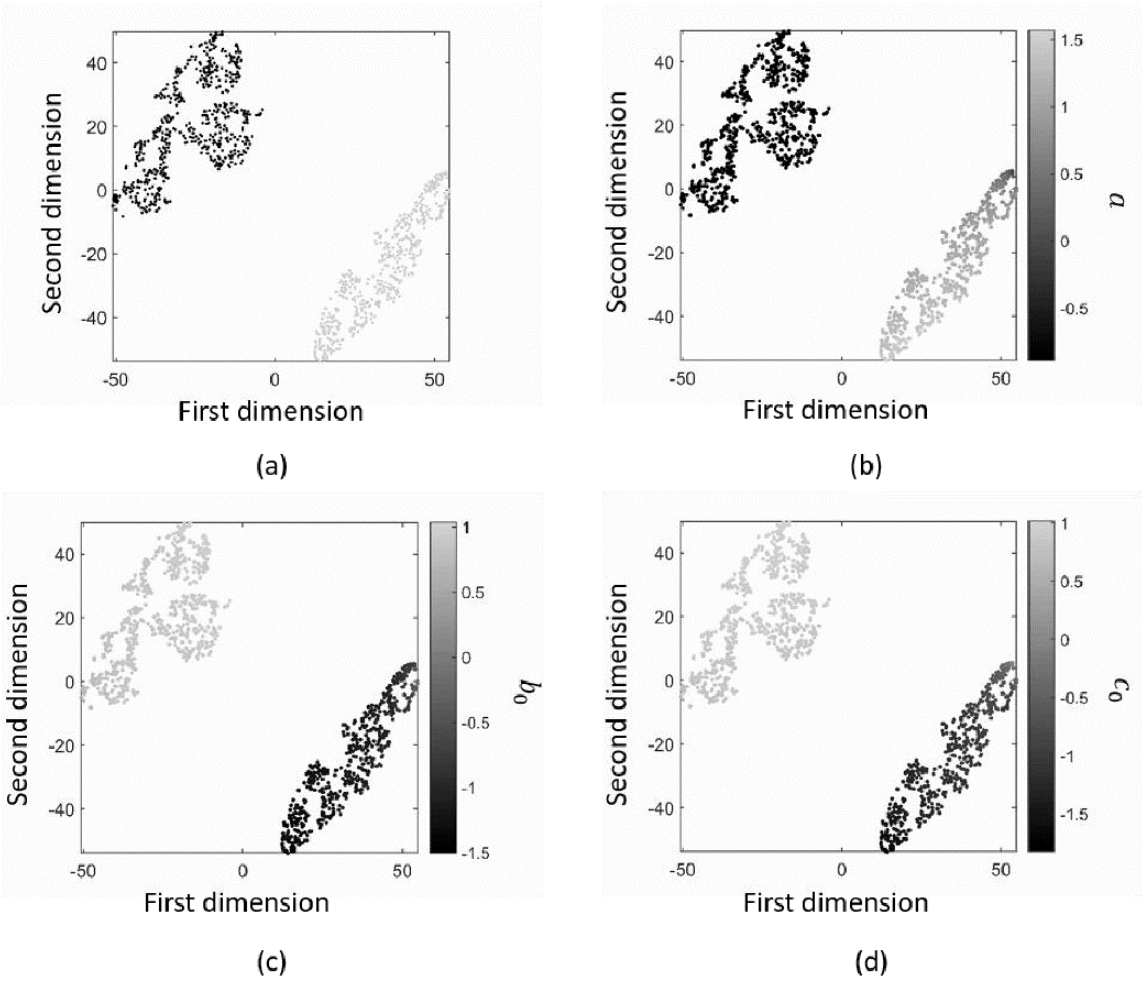


Figure S-8: Scatter plots of a two-dimensional t-SNE applied to the SNV-normalized EMSC regression coefficients $\{a, b_0, c_0\}$ estimated from simulated data with a first-order polynomial function, with point colors defined by (a) the pixel true labels (black and gray pixels correspond to paraffin and tissue pixels, respectively), (b) the a regression coefficient value, (c) the b_0 regression coefficient value, (d) the c_0 regression coefficient value.

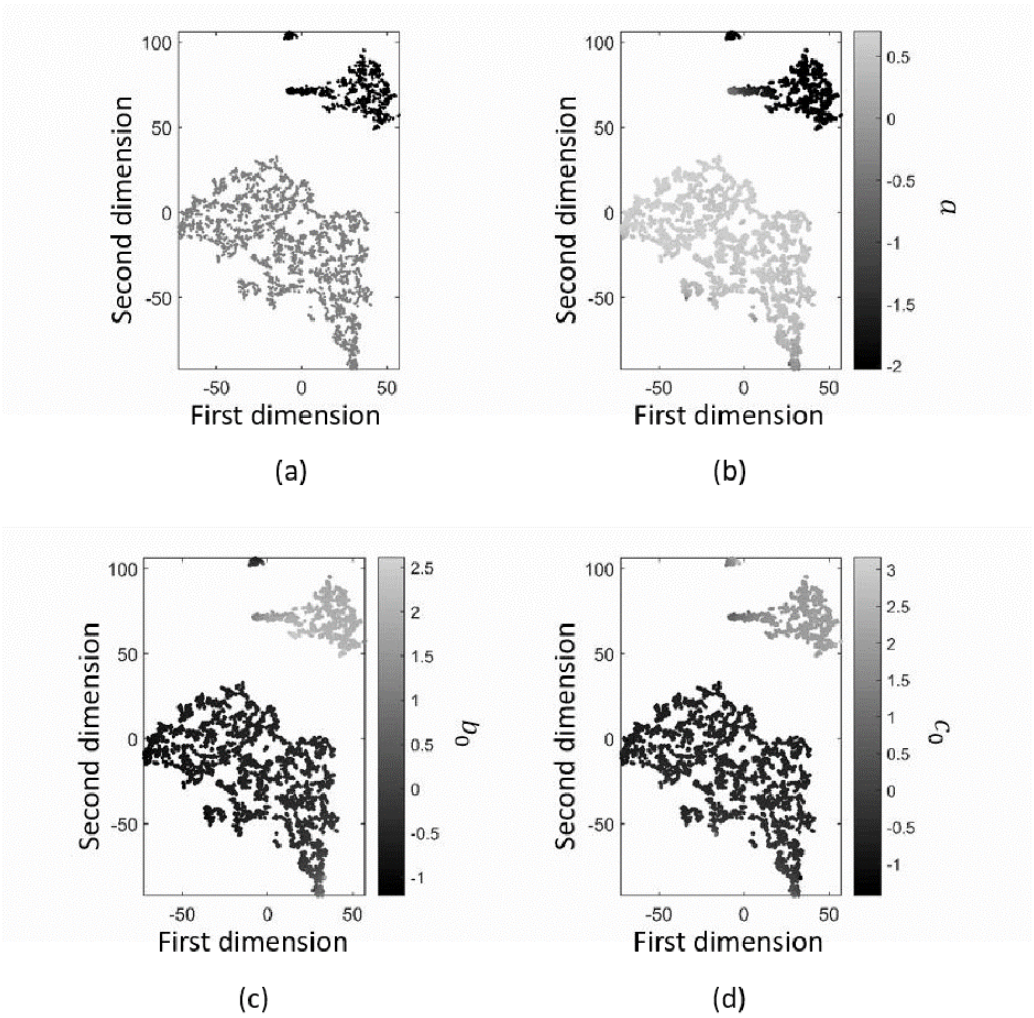


Figure S-9: Scatter plots of a two-dimensional t-SNE applied on the SNV-normalized EMSC regression coefficients $\{a,\ b_0,\ c_0\}$ estimated from a human colon cancer FFPE tissue section with point colors defined by (a) the pixel labels estimated by our proposed multivariate approach (black and gray pixels correspond to paraffin and tissue pixels, respectively), (b) the a regression coefficient value, (c) the b_0 regression coefficient value, (d) the c_0 regression coefficient value.

4) Supplementary tables

Table S-1: The top 5 combinations of SNV-normalized EMSC regression coefficients estimated on a simulated spectral image with a first-order polynomial function for four different validity indices (XB, DB, PBM, SWC) computed on their two-cluster KMeans partitions.

Rank\Validity index	XB	DB	PBM	SWC
1st	b_0	b_0	a, b_0, c_0	a, b_0
2nd	a, b_0	a	a, b_0	a, b_0, c_0
3rd	a	a, b_0	b_0, c_0	b_0, c_0
4th	b_0, c_0	b_0, c_0	a, c_0	a, c_0
5th	a, b_0, c_0	a, b_0, c_0	a, r, b_0, c_0	a, r, b_0

Table S-2: The top 5 combinations of SNV-normalized EMSC regression coefficients estimated on a FTIR image acquired on a human colon cancer sample for four different validity indices (XB, DB, PBM and SWC) computed on their two-cluster KMeans partitions.

Rank\Validity index	XB	DB	PBM	SWC
1st	a	a	a, c_0	a, c_0
2nd	a, c_0	a, c_0	a, b_0, c_0	a, b_0, c_0
3rd	a, b_0, c_0	a, b_0, c_0	a, r, b_0, c_0	a, r, b_0, c_0
4th	a, b_0	a, b_0	a, b_0	a, b_0
5th	a, r, b_0, c_0	a, r, b_0, c_0	a, r, c_0	a, c_3

III. C. Perspectives

Une première perspective intéressante de ce travail serait d'étudier la transférabilité de notre approche à d'autres milieux d'enrobage tels que l'OCT par exemple. Cette perspective a en fait été étudiée dans le cadre de ce travail de thèse et fait l'objet du chapitre suivant.

Une autre perspective intéressante serait d'étudier l'intérêt de notre approche pour distinguer différentes pathologies (par exemple normal versus cancéreux) ou différents stades d'une pathologie, c'est-à-dire d'utiliser les coefficients de régression de l'EMSC pour faire du diagnostic. En effet, les compositions biomoléculaires de ces types d'échantillons étant différentes, les effets de diffusion et de rétention du milieu d'enrobage peuvent être différents. Ainsi, les coefficients de régression de l'EMSC pourraient être significativement différents en fonction du type d'échantillon.

Il serait également intéressant d'appliquer notre méthodologie à d'autres types de méthodes d'extraction de variables ou de séparation de sources telles que la MCR-ALS ou l'Analyse en Composantes Indépendantes (ICA) qui ont déjà été appliquées à l'histopathologie spectrale (151–153). En effet, ces méthodes ne corrigent pas le signal de la paraffine, mais permettent d'identifier les différentes sources de variation contenues dans les données spectrales, y compris celle du milieu d'enrobage telle que la paraffine, et de quantifier leurs contributions. Les coefficients de concentration ainsi estimés par ces méthodes peuvent donc également permettre de distinguer les pixels de paraffine de ceux de tissu.

L'approche proposée dans ce travail utilise quatre indices de validité pour déterminer par consensus la meilleure combinaison des coefficients de régression de l'EMSC. L'application d'une classification non supervisée sur cette combinaison permet alors d'aboutir à une unique partition distinguant pixels de paraffine et pixels de tissu. A l'inverse, il serait intéressant d'estimer une partition à partir de chaque combinaison optimale déterminée par chaque indice de validité. En combinant ces quatre partitions, la robustesse de l'identification des pixels de tissu serait améliorée. La même idée pourrait être appliquée en combinant différentes méthodes d'extractions de variables ou de séparation de sources, comme expliqué plus haut. Le but de cette perspective serait d'aboutir à un consensus sur l'identification des pixels de tissus par différentes approches.

Chapitre IV : Identification des spectres non tissulaires sur des images spectrales IR acquises sur des coupes de tissus congelés

IV. A. Contexte et objectif

La micro spectroscopie infrarouge à transformée de Fourier (FTIR) est une technique émergente pour l'analyse biochimique des tissus et des matériaux cellulaires. L'histopathologie spectrale, définie comme la combinaison de l'imagerie infrarouge (IR) et de l'analyse multivariée, fournit des informations objectives sur la biochimie d'échantillons cellulaires ou tissulaires, et a été appliquée dans de nombreux domaines de la recherche biomédicale.

En fonction de leur origine, les échantillons sont conservés sous plusieurs formes. Or, il est devenu évident que la collecte, la préparation et le stockage des échantillons ont le potentiel d'influencer la composition biochimique des tissus biologiques, donc leur signature spectrale. Dans la mesure du possible, une attention particulière doit être apportée à la modélisation de ces effets dans le prétraitement des spectres IR.

En ce qui concerne les méthodes de conservation des échantillons, la fixation en paraffine et l'enrobage par OCT sont les principales utilisées.

En imagerie IR, la modélisation du signal de la paraffine et sa correction dans les spectres de tissu ont été développées avec succès ces dernières années à partir de l'EMSC (90,93,101,121). De plus, une nouvelle méthode multivariée pour sélectionner la meilleure combinaison des coefficients de régression de l'EMSC a été développée pour l'histopathologie spectrale IR dans le chapitre précédent du présent manuscrit afin d'identifier automatiquement les pixels de paraffine et ainsi ne conserver que les pixels tissulaires (154).

Cependant, il est recommandé de réaliser l'histopathologie spectrale sur des échantillons congelés puisque ce mode de conservation conserve mieux l'architecture et la biochimie des tissus, en particulier leurs informations lipidiques (155). Ces échantillons peuvent être incorporés dans des réactifs d'enrobage tel que l'OCT (153,156), un type de fixation tissulaire moins invasif que la fixation en paraffine (157). Ainsi, l'utilisation du milieu d'enrobage OCT maintient les échantillons humides, assure la stabilité des structures fines des tissus et permet de les stocker pendant de longues périodes à -80 °C, tout en évitant la formation de cristaux qui peuvent endommager les tissus. De plus, ce matériau d'enrobage reste à la surface et ne pénètre pas dans les tissus de l'échantillon à analyser.

Cependant, en histopathologie spectrale IR, il est recommandé de réaliser des cryosections avec un appareillage affuté et d'isoler très soigneusement les coupes fraîchement réalisées du bloc tissulaire afin d'éviter leur contamination par l'OCT (78). D'autre part, cette contribution de l'OCT n'est pas éliminée expérimentalement avant l'imagerie IR, et l'OCT présente une forte signature IR dont certaines bandes chevauchent des bandes IR spécifiques des échantillons tissulaires. En effet, le spectre IR de l'OCT est composé principalement de bandes à 1733, 1246 et 1110 cm^{-1} masquant certaines informations biologiques du tissu (157). De plus, la signature IR de l'OCT est suffisamment variable pour biaiser une analyse multivariée appliquée sur une image FTIR acquise sur une région d'intérêt d'un tissu incluant une partie d'OCT pur. En effet, une étude réalisée par imagerie IR sur des coupes transversales de follicules pileux enrobés dans de l'OCT a montré qu'un KMeans à 10 classes devait être appliqué à ce type d'échantillon pour obtenir une partition réaliste du cheveu en cinq classes (158). Ainsi, cinq clusters étaient associés à la présence d'un composé OCT, confirmant la variabilité du signal de l'OCT, conduisant à une confusion dans la répartition des contributions biologiques et rendant l'interprétation des résultats de l'histopathologie spectrale plus complexe.

Ainsi, afin d'appliquer l'histopathologie spectrale IR dans les meilleures conditions sur des échantillons de tissus biologiques, la correction du signal de l'OCT et l'exclusion des pixels associés à l'OCT sont nécessaires. Le but de la présente étude est donc de transférer et d'évaluer l'approche d'analyse multivariée (MA) des coefficients de régression EMSC développée dans le chapitre précédent pour des échantillons tissulaires de côlon fixés en paraffine à des échantillons du même organe mais enrobés dans de l'OCT.

IV. B. Matériels et méthodes

IV. B. 1. Coupes de tissu congelé

Au total, 8 blocs de tissus congelés de xénogreffes de carcinome du côlon (souris Nude NMRI-Foxn1nu/Fox1nu, Janvier®) et 5 blocs de muqueuses saines de souris, obtenus auprès de l'unité de recherche IRFAC-INSERM U113 de Strasbourg, sont inclus non fixés dans une colle histologique OCT. Pour chaque échantillon, deux sections adjacentes de 6 µm d'épaisseur ont été coupées dans un Cryostat LEICA (CM 3050 S) maintenu à une température moyenne de -22°C. Pour l'analyse IR, une première section a été placée sur un substrat de fluorure de calcium (CaF_2) (Crystran, Dorset, UK) transparent au rayonnement infrarouge. La deuxième section a été montée sur une lame de verre pour l'analyse histologique après coloration à l'H&E. Afin de maintenir la structure cellulaire du tissu congelé, les coupes tissulaires déposées sur les lames en verre sont fixées au paraformaldéhyde (4%) avant la coloration H&E.

IV. B. 2. Collecte des données spectrales

Les images FTIR en mode transmission ont été collectées en utilisant un système d'imagerie FTIR Perkin Elmer Spectrum Spotlight 300 équipé d'un détecteur de tellure de mercure et de cadmium (MCT) refroidi à l'azote liquide, avec 16 accumulations par spectre, une résolution spectrale de 4 cm^{-1} et une résolution spatiale de 6,25×6,25 μm^2 sur la gamme spectrale 750-4000 cm^{-1}. Un spectre du bruit de fond généré par le support CaF_2 a été enregistré avec 240 accumulations et soustrait automatiquement de chaque spectre de l'image collectée via le logiciel Spectrum Image (Perkin Elmer). Afin de faciliter la sélection des régions d'intérêt (ROI) pour l'analyse IR, une image visible en utilisant la lumière blanche a été collectée pour chaque section tissulaire avant chaque acquisition. Pour chaque échantillon, jusqu'à 3 zones tissulaires différentes et 1 zone d'OCT pur ont été imagées. Au total, 39 images spectrales FTIR et 13 d'OCT pur ont été collectées.

IV. B. 3. Analyse des données spectrales

Les analyses de données spectrales ont été effectuées à l'aide de scripts internes écrits en Matlab (The Mathworks, Natick, MA).

IV. B. 3. 1. Prétraitements standards

Comme décrit dans la section D.1 du chapitre II, les spectres bruts de chaque image FTIR ont été prétraités pour améliorer la qualité des données, suivant les trois étapes suivantes : une correction des absorptions atmosphériques de vapeur d'eau et de CO_2 effectuées via le logiciel Spectrum IMAGE (Perkin-Elmer), une conversion transmittance-absorbance des intensités des spectres, et une limitation de la gamme spectrale à la région d'empreinte digitale 900-1800 cm^{-1}, cette région étant la plus informative pour les échantillons de côlon (106,131).

IV. B. 3. 2. Modèle d'OCT

Comme décrit dans la partie « Contexte et objectif » de ce chapitre, l'OCT interfère avec le signal tissulaire et possède une signature spectrale variable dans la gamme 900-1800 cm^{-1}.

La procédure décrite dans la section D.2 du chapitre II pour modéliser la signature spectrale IR du médium d'enrobage a été déclinée ici pour modéliser la variabilité de la contribution spectrale de l'OCT. Après normalisation par SNV, les spectres d'une image FTIR acquise sur une zone d'OCT pur (sans tissu) sont modélisés par l'ACP. Afin de capturer la variance maximale dans l'ensemble des spectres IR d'OCT tout en réduisant la quantité de données, une matrice d'interférence **I** est construite à partir du spectre moyen de l'OCT pur et des 9 premières composantes principales (CPs) de l'OCT. Cette matrice d'interférence a ensuite été introduite dans le modèle d'EMSC.

IV. B. 3. 3. Extended Multiplicative Signal Correction (EMSC)

Puis comme décrit dans la section D.3 du chapitre II, un prétraitement par EMSC a été appliqué aux images spectrales IR acquises sur les coupes tissulaires afin de réaliser conjointement une normalisation des données sur le spectre moyen, une correction de la ligne de base par un polynôme du quatrième ordre, et une neutralisation de la variabilité du signal de l'OCT en utilisant le modèle d'interférence décrit dans la section précédente.

IV. B. 3. 4. Analyse multivariée des coefficients de régression de l'EMSC (MA)

Dans ce travail, nous avons évalué le potentiel de transfert de la méthode d'analyse multivariée des coefficients de régression de l'EMSC (ou MA pour Multivariate Analysis), développée principalement pour l'identification non supervisée et automatique des spectres de la paraffine (154), à la détection des spectres non tissulaires sur des coupes tissulaires congelées. Il est à noter que, contrairement à la paraffine, l'OCT n'infiltre pas le tissu et reste à sa périphérie sous

la forme d'une couche fine. Ainsi, à l'extérieur de la coupe congelée (au-delà de l'OCT) et à l'intérieur de l'échantillon tissulaire (si des cavités existent), des pixels du support en CaF_2 sont visibles. Ainsi, dans la suite, les pixels non tissulaires regroupent à la fois les pixels d'OCT mais également du support en CaF_2.

La même méthodologie que celle développée dans le chapitre précédent (154) a été appliquée pour l'OCT. En effet, la méthode MA a été appliquée sur des images IR acquises sur des sections tissulaires congelées, non fixées, incluses dans la colle histologique OCT. Les différentes combinaisons des coefficients de régression EMSC ont été évaluées par des indices de validité (XB, DB, PBM et SWC) appliqués à des partitions estimées par KMeans afin de décomposer l'ensemble de données spectrales en deux clusters, un cluster des spectres non tissulaires (OCT et CaF_2) et un autre des spectres tissulaires. En vue d'identifier efficacement tous les spectres non tissulaires et afin d'optimiser l'application de la méthode MA sur des images FTIR de tissu congelé, notre étude s'est focalisée sur trois points principaux, à savoir : i) l'intérêt d'intégrer une matrice d'interférence **I** de l'OCT au modèle d'EMSC, ii) l'impact du nombre de CPs considérées dans la matrice d'interférence **I** pour capturer la variabilité du signal de l'OCT, iii) l'impact du choix du spectre de référence inclus dans l'EMSC.

IV. C. Résultats et discussion

IV. C. 1. Limite de la méthode MA sans modélisation du signal d'OCT

La méthode d'analyse multivariée des coefficients de régression de l'EMSC (MA) a été appliquée sur des images spectrales acquises sur des coupes tissulaires congelées sans modéliser le signal spectral de l'OCT (pas de matrice d'interférence **I** incluse dans le modèle d'EMSC). Ainsi, les quatre indices de validité XB, DB, PBM et SWC ont été appliqués sur des partitions KMeans à deux clusters estimées pour chaque combinaison possible des coefficients EMSC $\{a, r, c_0, c_1, c_2, c_3, c_4\}$. En fonction des valeurs calculées par chaque indice de validité et ordonnées par ordre croissant pour XB et DB, et par ordre décroissant pour PBM et SWC, les 5 meilleures combinaisons des coefficients EMSC donnant les deux clusters les plus compacts et séparés ont été retenues. Un exemple illustratif de ces combinaisons obtenues sur une image FTIR acquise sur une section tissulaire de carcinome de côlon xénogreffé est présenté dans la Table 6. La coupe analysée par imagerie IR et la coupe adjacente colorée à l'H&E sont visibles sur les Figures 27(a-b).

Rang\Indice de validité	XB	DB	PBM	SWC
1	c_3	c_3	c_1, c_3	c_1, c_3
2	c_1, c_3	c_1, c_3	r, c_1, c_3	c_4
3	c_1	c_1	a, r, c_1, c_3	c_2
4	r, c_1, c_3	r, c_1, c_3	r, c_0, c_1, c_3	a, c_0
5	c_0	r, c_3	a, r, c_0, c_1, c_3	a, r

Table 6. Les 5 meilleures combinaisons de coefficients de régression EMSC estimées pour chacun des quatre indices de validité (XB, DB, PBM et SWC) appliqués sur une image spectrale FTIR acquise sur une section tissulaire de carcinome de côlon xénogreffé.

D'après la Table 6, deux combinaisons de coefficients de régression EMSC font consensus, à savoir $\{c_1, c_3\}$ et $\{c_3\}$, pour séparer au mieux les pixels tissulaires de ceux non tissulaires lorsque le signal de l'OCT n'est pas modélisé dans l'EMSC. Sans modèle d'EMSC, ce sont donc les formes différentes de ligne de base qui permettent de distinguer les pixels de tissus de ceux de l'OCT et du support en CaF_2.

Afin de vérifier l'efficacité de ces combinaisons à séparer les pixels tissulaires des non tissulaires, la partition KMeans à deux classes sur la combinaison $\{c_1, c_3\}$ a été estimée et est représentée sur la Figure 27(c). Comparée à l'histologie de cet échantillon (Figure 27(b)), cette partition est en faible corrélation avec la structure de l'échantillon analysé, indiquant que les pixels non tissulaires et les pixels tissulaires sont mal identifiés. En effet, des spectres typiques du substrat en CaF_2 (Figure 27(d)) et des spectres avec une forte contamination par le signal de l'OCT (Figure 27(e)) ont été identifiés comme étant des pixels tissulaires.

Ce résultat confirme ainsi qu'une modélisation du signal de l'OCT est nécessaire pour détecter correctement les pixels non tissulaires (OCT, substrat).

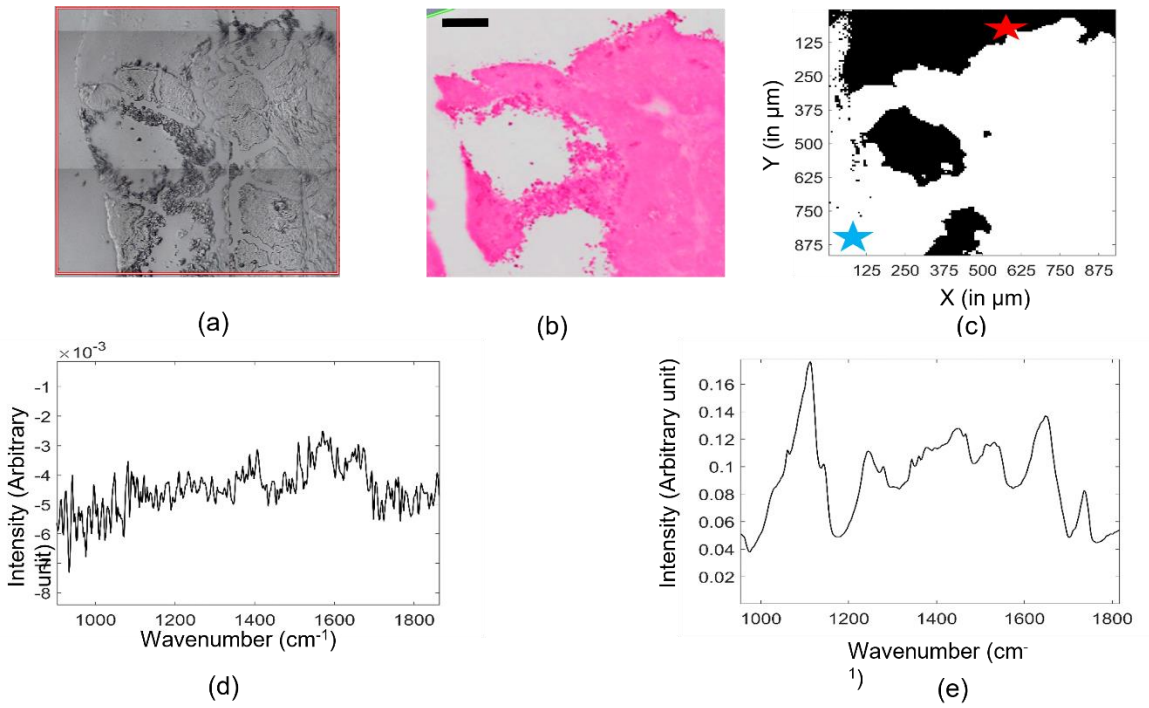


Figure 27. Exemple d'application de la méthode d'analyse multivariée des coefficients de régression d'EMSC sans matrice d'interférence sur une image FTIR acquise sur un échantillon congelé de carcinome de côlon xénogreffé. (a) Image visible de la région analysée par imagerie IR sur une coupe de cet échantillon. (b) Image de la coupe adjacente colorée à l'H&E. La barre d'échelle indique 100 µm. (c) La partition KMeans à deux classes obtenue sur la combinaison $\{c_1, c_3\}$. Les pixels noirs et blancs correspondent respectivement aux pixels estimés d'OCT-CaF_2 et de tissu. (d, e) Exemples de spectres bruts correspondants aux pixels identifiés comme tissulaires par application d'un KMeans à deux classes sur la combinaison $\{c_1, c_3\}$ et marqués sur (c) par une croix bleu (d), et une croix rouge (e).

IV. C. 2. Succès de la méthode MA avec un modèle d'interférence de l'OCT

Les résultats présentés précédemment montrent que la contribution de l'OCT doit être la plus précisément possible modélisée car elle interfère avec le signal tissulaire. Une matrice d'interférence **I** modélisant la variabilité de signal d'OCT est alors introduite dans le modèle d'EMSC.

Le premier objectif de cette partie d'analyse est donc de déterminer la meilleure combinaison de coefficients de régression EMSC pour la séparation des pixels OCT-CaF_2 de ceux de tissu après une modélisation du signal OCT. La méthode MA a été appliquée à des images FTIR acquises sur des sections tissulaires de carcinome de côlon ou de côlon sain de plusieurs échantillons murins ou humains. Pour chaque image FTIR acquise sur ces coupes congelées, les informations d'entrée initiales de la méthode MA ont été les coefficients de régression EMSC $\{a, r, b_0, b_1, \dots, b_9, c_0, c_1, c_2, c_3, c_4\}$. Les meilleures combinaisons consensuelles de coefficients EMSC estimées pour chaque image par la méthode MA sont résumées dans la Table 7.

À partir de la Table 7, les trois combinaisons consensuelles les plus souvent identifiées sont $\{a, b_0\}, \{b_0, c_0\}, \{a, b_0, c_0\}$. Seules deux images engendrent des combinaisons consensuelles différentes, à savoir $\{a, c_0\}$ et $\{a, r, c_0\}$, pour séparer les pixels non tissulaires de ceux tissulaires.

Un exemple de la partition KMeans à deux classes estimée sur la combinaison optimale des coefficients de régression EMSC $\{a, r, c_0\}$ obtenue sur l'image FTIR de l'échantillon noté Souris #1, ROI #1 dans la Table 7 est illustré sur la Figure 28(a).

Une contribution a non significative du spectre de référence, une erreur de modélisation r élevée et une forte ligne de base c_0 suffisent à discriminer les spectres d'OCT-CaF_2 des spectres tissulaires (Figures 28(b-d)). Cependant, le fait que l'erreur de modélisation r ait été identifiée pour deux images IR par MA comme un paramètre pertinent pour la séparation des données en deux clusters distincts OCT-CaF_2 et tissu peut se justifier par deux hypothèses :

- Ces deux images possèdent un rapport signal sur bruit plus faible que les autres images.
- La modélisation du signal de l'OCT est imparfaite comme on peut le voir sur la Figure 28(d) où l'erreur de modélisation r est forte et très hétérogène dans les zones d'OCT pur.

Échantillon et ROI	Type d'échantillon	Etat physio-pathologique	Combinaisons optimales de coefficients
Souris #1, ROI #1	Xénogreffe	Cancéreux	a, c_0 a, r, c_0
Souris #2, ROI #1	Xénogreffe	Cancéreux	a, b_0 b_0, c_0 a, b_0, c_0
Souris#3, ROI #1	Xénogreffe	Cancéreux	a, b_0 a, b_0, c_0 a, c_0
Souris #4, ROI #1	Xénogreffe	Cancéreux	a, b_0 b_0, c_0 a, b_0, c_0
Souris #4, ROI #2	Xénogreffe	Cancéreux	a, b_0 b_0, c_0 a, b_0, c_0
Souris #5, ROI #1	Xénogreffe	Cancéreux	b_0, c_0 a, b_0, c_0 a, c_0
Souris#6, ROI #1	Xénogreffe	Cancéreux	a, c_0 a, r, c_0
Souris #7, ROI #2	Xénogreffe	Sain	a, b_0 b_0, c_0 a, b_0, c_0
Souris#8, ROI #2	Xénogreffe	Sain	a, b_0 b_0, c_0 a, b_0, c_0
Patient #1, ROI #1	Patient	Sain	b_0, c_0 a, b_0, c_0 a, c_0

Table 7. Combinaisons consensuelles de coefficients de régression EMSC estimées par la méthode MA sur des images spectrales FTIR acquises sur des coupes tissulaires congelées de carcinome de côlon xénogreffé ou de muqueuse saine de souris ou de patients humains.

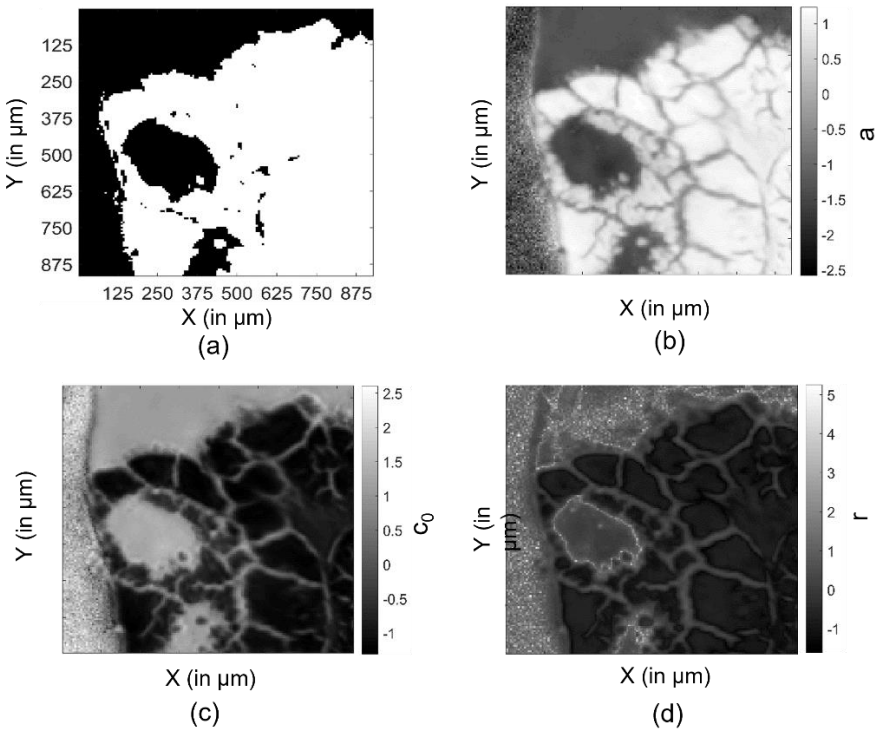


Figure 28. Exemple d'application de l'approche multivariée sur une image FTIR acquise sur un échantillon congelé de carcinome de côlon xénogreffé (Souris #1, ROI #1, Table 7). (a) La partition KMeans à deux classes obtenue sur la combinaison $\{a, r, c_0\}$. Les pixels noirs et blancs correspondent respectivement aux pixels estimés d'OCT-CaF_2 et de tissu. (b) Image en niveaux de gris reconstruite à partir du coefficient de régression a. (c) Image en niveaux de gris reconstruite à partir du coefficient de régression c_0. (d) Image en niveaux de gris reconstruite à partir des valeurs d'erreur de modélisation r.

Un deuxième exemple illustratif de la partition KMeans à deux classes estimée sur la combinaison optimale des coefficients de régression EMSC $\{a, b_0, c_0\}$ obtenue sur l'image FTIR de l'échantillon congelé de carcinome du côlon xénogreffé noté Souris #2, ROI #1 dans la Table 7 est illustré sur la Figure 29(c).

La comparaison de cette figure à l'image visible de la coupe (Figure 29(a)) et à sa coupe adjacente colorée à l'H&E (Figure 29(b)) révèle l'efficacité de cette combinaison $\{a, b_0, c_0\}$ à distinguer les contributions non tissulaires de celles tissulaires. En effet, l'image KMeans estimée sur la combinaison $\{a, b_0, c_0\}$ obtenue par MA est fortement corrélée aux deux images de références (Figures 29(a-b)).

Ces résultats sont confirmés sur les images reconstruites à partir des coefficients de régression de la combinaison optimale, à savoir les coefficients a, b_0 et c_0, et représentées sur les Figures 29(d-f). En effet, un contraste clair entre les zones non tissulaires (OCT et support) et les zones

tissulaires peut être visualisé sur ces figures. L'utilisation de la combinaison $\{a, b_0, c_0\}$ confirme donc son efficacité à discriminer les pixels tissulaires.

En outre, dans cet exemple d'application, les spectres d'OCT sont caractérisés par une forte contribution du signal d'OCT via le coefficient b_0 (Figure 29(e)), une faible contribution du spectre de référence via le coefficient a (Figure 29(d)) et une forte ligne de base via le coefficient c_0 (Figure 29(f)). L'inverse est observé pour les spectres tissulaires. Quant aux spectres du support en CaF_2, ils sont caractérisés par des valeurs faibles de ces 3 coefficients.

Les spectres bruts d'OCT présentent une forte hétérogénéité et une forte corrélation avec le spectre du tissu. Ce n'est qu'après prise en compte de cette variabilité dans le modèle d'EMSC que les différences caractéristiques entre le tissu et l'OCT deviennent plus prononcées. Appliquées sur les coefficients de régression estimées par EMSC, notre méthode MA est alors efficace pour estimer la combinaison optimale permettant de détecter tous les pixels non tissulaires (OCT, substrat) d'une image spectrale IR.

Dans l'ensemble, ces résultats sont en accord avec ceux obtenus sur le même type d'échantillons de tumeur colique humaine mais fixés et inclus en paraffine et présentés dans le chapitre précédent. En effet, la même combinaison optimale $\{a, b_0, c_0\}$ avait été obtenus, prouvant ainsi l'efficacité de l'EMSC à modéliser les interférences spectrales des milieux d'enrobage et le caractère hautement informatif de ses coefficients de régression (154).

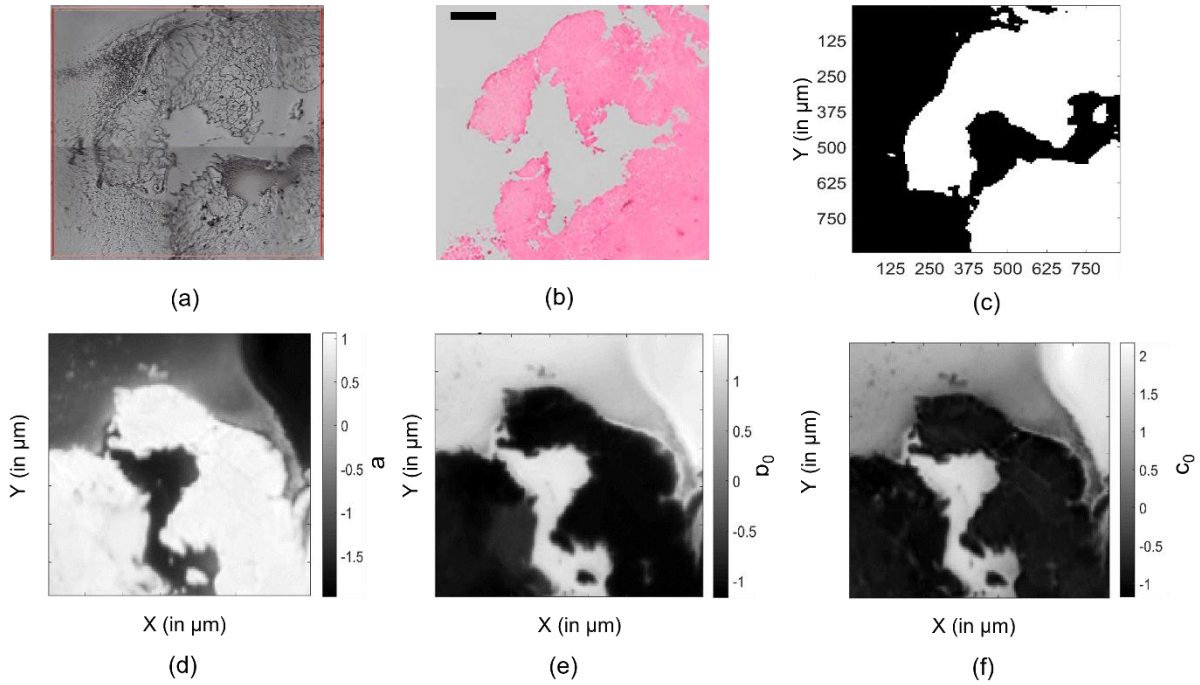


Figure 29. Exemple d'application de l'approche multivariée sur une image FTIR acquise sur un échantillon congelé de carcinome de côlon xénogreffé (Souris #2, ROI #1, Table 7). (a) Image visible de la région analysée par imagerie IR sur une coupe de cet échantillon. (b) Image de la coupe adjacente colorée à l'H&E. La barre d'échelle indique 100 µm. (c) La partition KMeans à deux classes obtenue sur la combinaison $\{a, b_0, c_0\}$. Les pixels noirs et blancs correspondent respectivement aux pixels estimés d'OCT-CaF_2 et de tissu. (d) Image en niveaux de gris reconstruite à partir du coefficient de régression a. (d) Image en niveaux de gris reconstruite à partir du coefficient de régression b_0. (e) Image en niveaux de gris reconstruite à partir du coefficient de régression c_0.

IV. C. 3. Impact du modèle d'interférence de l'OCT sur la méthode multivariée

Pour compléter la caractérisation de notre méthode multivariée de détection non supervisée des pixels non tissulaires, l'évaluation de sa robustesse en fonction de la complexité du modèle d'interférence du signal IR de l'OCT a été réalisée. Dans cette partie, nous avons donc évalué l'influence sur la méthode MA du nombre de (CPs) capturant la variabilité du signal de l'OCT, estimées sur une image FTIR d'OCT pure et imposées dans la matrice d'interférence **I** du modèle EMSC.

Ainsi, pour chaque nombre de CPs appartenant à l'ensemble $\{2, 3, 4, 7, 14, 26\}$ et exprimant entre 75% et 97% de la variance totale des spectres IR acquis sur cette image d'OCT pur, la matrice d'interférence I a été construite, puis une image spectrale IR acquise sur un échantillon

tissulaire enrobé dans de l'OCT a été prétraitée par le modèle d'EMSC incorporant cette matrice d'interférence I. Enfin, la méthode MA a été appliquée, toujours pour les quatre indices de validité XB, PBM, DB et SWC, afin d'estimer les meilleures combinaisons de coefficients de régression EMSC pour séparer les données en deux classes compactes et séparables.

Un exemple de résultat obtenu sur l'échantillon (Souris #1, ROI #1) de la Table 7 est présenté dans la Table 8 en donnant les 2 meilleures combinaisons de coefficients de régression EMSC estimés par les quatre indices de validité pour chaque nombre de CPs inclus dans la matrice d'interférence I du modèle d'EMSC.

Globalement, quel que soit le nombre de CPs considéré dans la matrice d'interférence I, notre méthode MA utilisant les indices de validité appliquée sur une image FTIR de tissu congelé estime objectivement les mêmes combinaisons de coefficients de régression EMSC efficaces à la discrimination des pixels non tissulaires (Table 8). En terme d'estimation des meilleures combinaisons de coefficients de régression d'EMSC, notre méthode MA est donc très stable en fonction du nombre de CPs inclus dans la matrice d'interférence I du modèle d'EMSC.

Variance expliquée (%)	**75**	**92**	**93**	**95**	**96**	**97**
Nombre de CPs	2	3	4	7	14	26
Meilleures combinaisons	a, c_0 a, r, c_0	a, c_0 a, r, c_0	a, c_0 a, r, c_0	a, c_0 a, r, c_0	a, c_0 a, r, c_0	a, c_0 a, r, c_0

Table 8. Combinaisons consensuelles des coefficients de régression EMSC estimées par les quatre indices de validité sur l'image spectrale FTIR acquise sur la coupe tissulaire de l'échantillon Souris #1, ROI #1 de la Table 7.

Afin de visualiser et confirmer concrètement l'influence de nombre de CPs inclus dans la matrice d'interférence I sur la stabilité de détection des spectres non tissulaires par MA, les partitions KMeans à deux classes sur la combinaison optimale estimée $\{a, r, c_0\}$ sont illustrées sur les Figures 30(a-b) pour 2 CPs et 14 CPs expliquant respectivement 75% et 96% de variance du signal de l'OCT. Il est évident que le nombre de CPs inclus dans I influence énormément la discrimination entre pixels tissulaires et non tissulaires. Plus le nombre de CPs augmente, et plus le nombre de pixels identifiés comme non tissulaires augmente.

Afin de comprendre ce phénomène, des spectres bruts de pixels identifiés comme non tissulaires avec un modèle à 14 CPs et comme tissulaires avec un modèle à 2 CPs ont été analysés

visuellement. Deux exemples de ces spectres bruts sont présentés sur les Figures 30(c-d). Ces spectres supplémentaires identifiés comme non tissulaires pour un grand nombre de CPs sont caractéristiques d'une faible contribution du tissu ou d'un faible rapport signal sur bruit. Quelques spectres présentent également les signes d'une légère contamination du signal tissulaire par l'OCT.

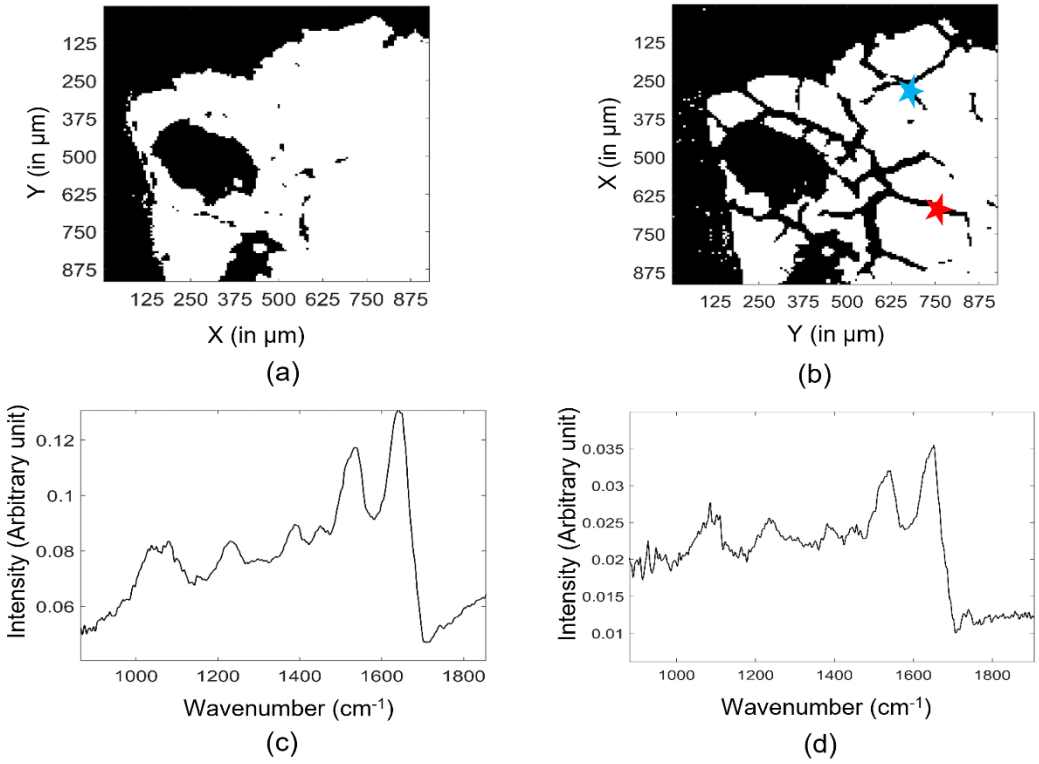


Figure 30. Impact du nombre de CPs inclus dans la matrice d'interférence I sur les partitions KMeans à deux classes estimées sur la combinaison $\{\{a, r, c_0\}$ de coefficients de régression EMSC pour une image spectrale FTIR acquise sur un échantillon congelé de carcinome de côlon xénogreffé. (a, b) Partitions KMeans pour I composée de (a) 2 CPs, et (b) 14 CPs. Les pixels noirs et blancs correspondent respectivement aux pixels d'OCT et de tissu estimés. (c, d)) Exemples de spectres bruts correspondants aux pixels identifiés comme non tissulaires par l'approche multivariée utilisant une matrice d'interférence I composée de 14 CPs et marqués sur (b) par : une croix rouge (c), et une croix bleue (d).

Ainsi, quel que soit le nombre de CPs utilisé pour construire la matrice d'interférence I du modèle d'EMSC, notre méthode MA est capable d'identifier avec une grande précision les pixels non tissulaires sur des images spectrales FTIR acquises sur des échantillons congelés enrobés dans de l'OCT. Pour en plus permettre à notre méthode MA l'élimination, pour des analyses ultérieures, des spectres tissulaires bruités, acquis sur des pixels avec peu de tissu ou contaminés par l'OCT, alors un modèle d'interférence incluant plus de CP sera privilégié lors

de l'application de l'EMSC. Notre méthode MA est donc flexible puisque, en fonction des objectifs d'une étude, elle peut éliminer d'une image FTIR uniquement les pixels non tissulaires, ou en plus éliminer les pixels tissulaires de mauvaise qualité.

IV. C. 4. Impact du choix du spectre de référence sur la méthode multivariée

Afin d'étudier le comportement de notre méthode MA de détection des pixels non tissulaires en fonction du choix du spectre de référence utilisé dans l'EMSC, nous avons comparé les résultats de l'application de MA obtenus pour différents choix de ce spectre de référence :

- en prenant un spectre moyen calculé sur une image FTIR acquise sur une zone d'OCT pur (donc un spectre moyen d'OCT pur) ;
- en prenant un spectre moyen calculé sur une image FTIR acquise sur une zone tissulaire pure (donc un spectre moyen de tissu) ;
- en prenant un spectre moyen d'une image FTIR mixte, c'est à dire contenant à la fois des contributions de tissu et d'OCT.

La première étape de cette analyse a consisté à déterminer l'impact du choix du spectre référence sur la détermination de la meilleure combinaison de coefficients de régression EMSC pour la séparation des pixels tissulaires et non tissulaires. Pour chaque spectre de référence utilisé et pour chaque combinaison possible de coefficients, une partition KMeans à deux clusters a été estimée et a alimenté les quatre indices de validité. Pour chaque indice de validité, les 3 meilleures combinaisons de coefficients de régression EMSC conduisant aux meilleures valeurs d'indice de validité ont été déterminées. Les résultats obtenus sont illustrés dans la Table 9. D'après ces résultats, les meilleures combinaisons des coefficients EMSC sont conservées quel que soit le spectre de référence considéré. En effet, les deux combinaisons consensuelles restent toujours $\{a, r, c_0\}$ et $\{a, c_0\}$. En terme d'estimation des meilleures combinaisons de coefficients de régression d'EMSC, notre méthode MA est donc très stable en fonction du choix du spectre de référence du modèle d'EMSC.

Spectre de référence	Spectre moyen de tissu pur	Spectre moyen d'OCT pur	Spectre moyen d'une image mixte
Meilleures combinaisons	a, c_0 a, r, c_0	a, c_0 a, r, c_0	a, c_0 a, r, c_0

Table 9. Combinaisons consensuelles de coefficients de régression EMSC estimées par les quatre indices de validité sur l'image spectrale FTIR acquise sur une coupe tissulaire de carcinome du côlon xénogreffé (Souris #1, ROI #1 de la Table 7), pour différentes définitions du spectre de référence utilisé dans le modèle d'EMSC**.**

Afin de visualiser le comportement de notre méthodologie MA pour la distinction entre les pixels tissulaires et non tissulaires, les partitions KMeans à deux classes estimées sur la combinaison optimale de coefficients d'EMSC $\{a, \mathrm{r}, c_0\}$ pour chaque cas d'étude sont illustrées sur la Figure 31. Les résultats de la détection des spectres non tissulaires par notre méthode MA varient en fonction du spectre de référence considéré dans l'EMSC. En effet, l'utilisation d'un spectre de référence ne contenant aucune contribution tissulaire (Figure 31(e)) conduit à une confusion dans l'identification des pixels tissulaires de ceux non tissulaires (Figure 31(b)). En effet, les pixels du support sont confondus avec des pixels tissulaires.

Au contraire, l'utilisation d'un spectre moyen d'une image FTIR acquise sur un échantillon composé d'OCT et de tissu (Figure 31(d)) ou d'un spectre de tissu pur (Figure 31(f)) conduit à des partitions KMeans très similaires qui discriminent parfaitement les pixels tissulaires de ceux non tissulaires, comme montrées respectivement sur les Figure 31(a) et Figure 31(c).

Afin d'expliquer cette dépendance des résultats de notre approche MA au choix du spectre de référence, la Figure 6 présente les images en niveaux de gris reconstruites à partir du coefficient r estimé en utilisant dans l'EMSC un spectre de référence défini par un spectre moyen d'OCT pur (Figure 32(a)) ou par un spectre tissulaire pur (Figure 32(d)). Lorsqu'un spectre moyen d'OCT est utilisé comme spectre de référence, alors les spectres tissulaires et de support sont très mal modélisés puisqu'ils ne contiennent pas d'OCT. Ainsi, leur erreur de modélisation r est grande (Figure 32(a)) et leur coefficient de régression a (représentant la contribution du spectre de référence) est faible (Figure 32(c)). L'opposé est observé pour les spectres d'OCT qui exhibent un fort coefficient a (Figure 32(c)) et un faible résidu r (Figure 32(a)). La confusion entre pixels tissulaires et pixels de support est donc inévitable. Lorsqu'un spectre tissulaire pur est utilisé comme spectre de référence, alors les spectres d'OCT et de support sont

mal modélisés, induisant un faible coefficient a (Figure 32(d)) car ils ne contiennent pas de tissu, et un fort résidu r (Figure 32(b)). L'opposé est observé pour les spectres tissulaires qui exhibent un fort coefficient a (Figure 32(d)) et un faible résidu r (Figure 32(b)).

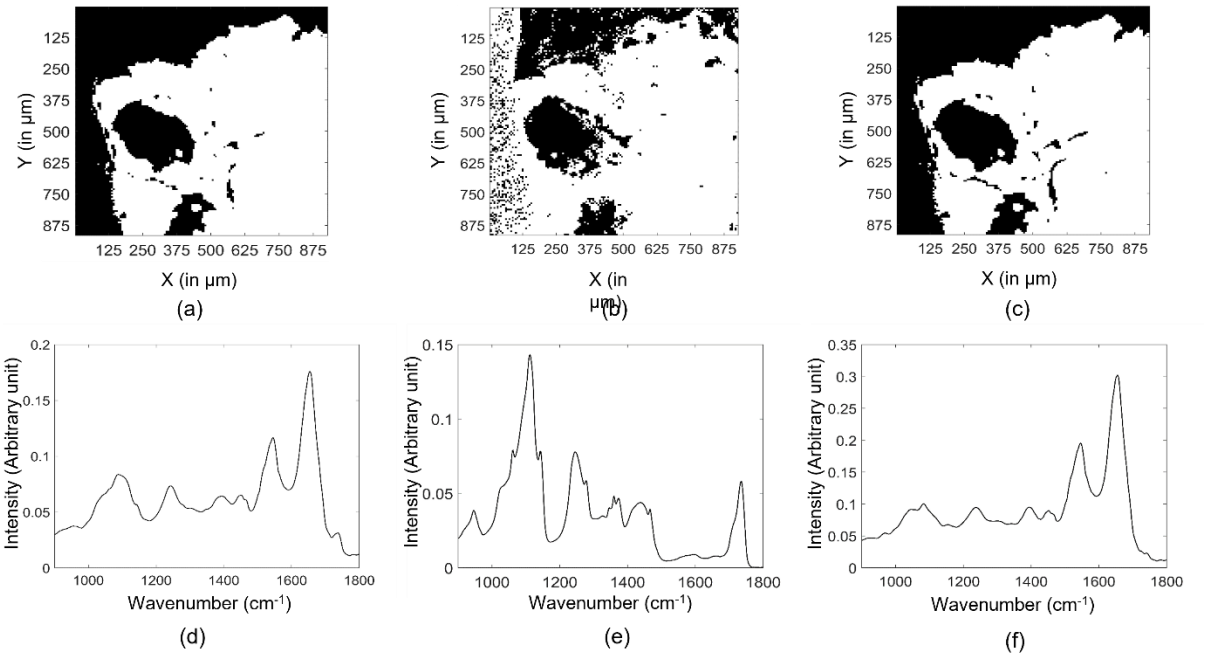


Figure 31. Impact du type de spectre de référence considéré dans l'EMSC sur l'efficacité de l'approche multivariée à détecter les spectres non tissulaires. (a-c) Partitions KMeans à deux classes sur la combinaison optimale $\{a, r, c_0\}$ en utilisant pour spectre de référence : (a) un spectre moyen de l'image FTIR analysée, (b) un spectre moyen d'OCT pur, (c) un spectre moyen de tissu pur. Les pixels noirs et blancs correspondent respectivement aux pixels estimés comme non tissulaires et tissulaires. (d-f) Spectres de référence utilisés pour générer les partitions KMeans des figures (a-c) : (d) un spectre moyen de l'image FTIR analysée, (e) un spectre moyen d'OCT pur, (f) un spectre moyen de tissu pur.

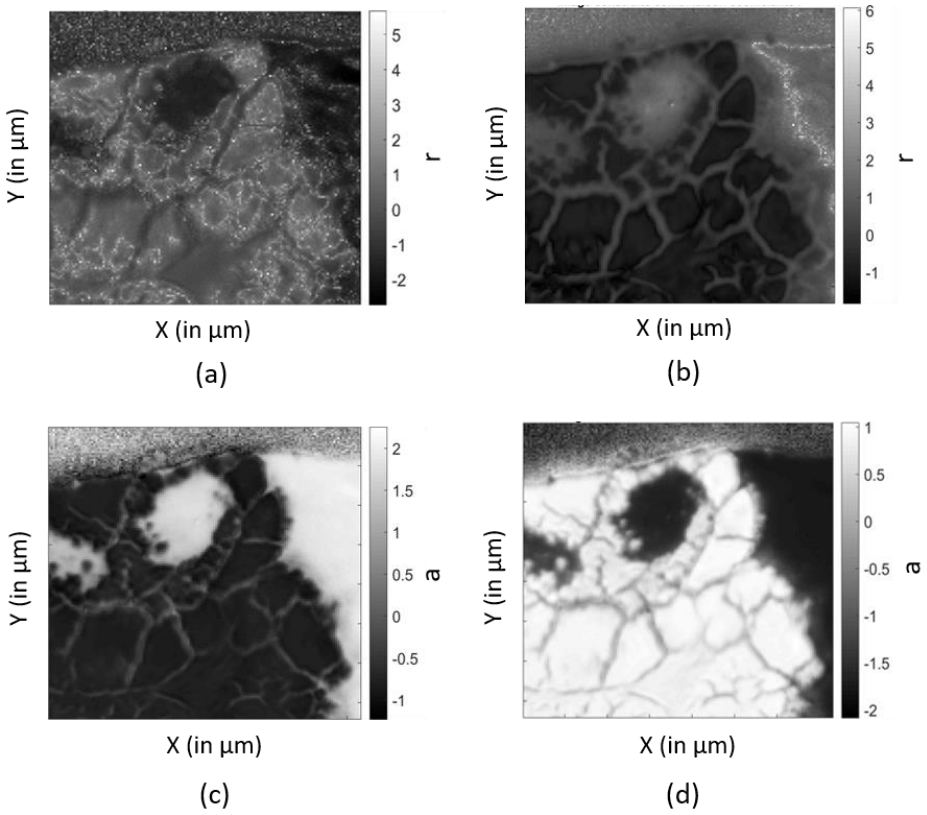


Figure 32. Impact du type de spectre de référence sur les coefficients de régression EMSC. Images en niveaux de gris reconstruites à partir du résidu r obtenu pour un spectre de référence choisi (a) comme un spectre d'OCT pur ou (b) comme un spectre tissulaire pur. Images reconstruites à partir du coefficient a obtenu pour un spectre de référence choisi (c) comme un spectre d'OCT pur ou (d) comme un spectre tissulaire pur.

Dans l'ensemble, ces résultats prouvent que notre méthodologie est robuste aux principaux paramètres de l'EMSC, à savoir le spectre de référence et le modèle d'interférence.

Il est à noter que l'optimisation des combinaisons optimales des coefficients de régression de l'EMSC par notre méthode afin d'identifier les pixels tissulaires devrait être effectuée pour chaque image individuelle afin de prendre en compte la variabilité de signal OCT et le bruit inhérents à chaque image. Cependant, on peut raisonnablement penser que ces caractéristiques restent relativement invariables pour des acquisitions réalisées sur le même type d'échantillons provenant du même hôpital, préparés selon le même protocole et analysés par imagerie IR en utilisant les mêmes paramètres d'acquisition. Dans ce cas, les combinaisons optimales estimées sur quelques échantillons peuvent être appliquées sur toute la cohorte.

IV. D. Conclusion

Dans cette étude, nous avons testé et validé le potentiel de transfert de notre méthodologie d'analyse multivariée (MA) des coefficients de régression d'EMSC pour la détection automatique des pixels non tissulaires sur des coupes tissulaires de tissus congelés. En utilisant la combinaison optimale des coefficients de régression EMSC estimée par les indices de validité, nous avons démontré l'efficacité de notre méthodologie pour séparer automatiquement les pixels non tissulaires, spécifiques du milieu d'enrobage OCT et du substrat en CaF_2, de ceux du tissu sur des images IR acquises sur des carcinomes de côlon xénogreffés et de la muqueuse saine de souris. De plus, nous avons montré la robustesse de notre méthode face aux principaux composants de l'EMSC, à savoir le modèle d'interférence et le spectre de référence. En outre, en adaptant la complexité du modèle d'interférence de l'EMSC, notre méthode peut également éliminer pour une analyse ultérieure les spectres tissulaires bruités ou contaminés par de l'OCT.

IV. E. Résultats supplémentaires

Notre méthode de détection automatique de contribution des signaux due au milieu d'enrobage a été testée aussi sur d'autres types d'échantillons, tels que le cheveu.

Appliquée sur des images spectrales IR acquises sur des coupes longitudinales de cheveu, les résultats de ce test ont validé l'efficacité de notre méthode à modéliser le signal d'OCT et son élimination avec une bonne précision. Etant donné que ce projet sur les cheveux a un caractère confidentiel, les images et les résultats de ce test ne seront pas montrés dans ce manuscrit.

IV. F. Perspectives

Notre méthodologie MA a été appliquée sur l'empreinte digitale des spectres IR, c'est-à-dire sur la gamme 900-1800 cm^{-1}. Or, il a été démontré que l'analyse par la méthode de séparation de source MCR-ALS d'images spectrales IR acquises sur des échantillons biologiques enrobés dans de l'OCT est améliorée lorsque la gamme spectrale complète (400-4000 cm^{-1}) est exploitée (153). Il est donc important dans la suite de ce travail d'étudier la plus-value apportée à notre méthode en exploitant le spectre entier afin d'améliorer la différenciation entre les contributions biologiques et non biologiques.

En outre, notre méthodologie MA a démontré son potentiel de transfert sur des échantillons de tissus congelés afin de détecter et éliminer les pixels non tissulaires (OCT et support). Cependant, l'utilisation de données simulées serait d'une valeur inestimable pour évaluer objectivement l'efficacité de notre méthodologie. En effet, l'utilisation de données simulées où toutes les sources de variabilité sont maitrisées et les labels de chaque spectre sont connus (gold standard) peut apporter des preuves solides de notre méthode à détecter les pixels non tissulaires, comme nous l'avions fait sur des tissus fixés en paraffine dans le chapitre précédent (154).

Enfin, le meilleur moyen de valider concrètement notre méthode pour des tissus enrobés dans de l'OCT serait d'ôter chimiquement l'OCT sur les coupes tissulaires déjà analysées par imagerie IR. En effet, nous disposerions ainsi d'un réel gold standard permettant une évaluation précise des performances de notre méthode. Récemment, une méthode a été proposée dans la littérature pour réaliser cette étape (159).

Chapitre V : Développement d’une nouvelle méthode de sélection non supervisée de variables basée sur les algorithmes génétiques.

V. A. Préambule

V. A. 1. Contexte et objectif

L'histopathologie spectrale IR a montré un grand potentiel pour l'identification fiable des structures tissulaires, sans préparation spécifique de l'échantillon, facilitant sa possible intégration dans le protocole de routine clinique. Jusqu'à présent, la majorité des études en histopathologie spectrale IR exploite toutes les informations contenues dans la gamme spectrale allant de 900 à1800 cm^{-1}, constituant la région d'empreinte digitale. Or, il est possible d'extraire au sein de cette gamme spectrale des biomarqueurs plus spécifiques des structures histologiques. L'importance de cette sélection de variables a été soulignée par plusieurs auteurs, en particulier pour améliorer l'efficacité des modèles de prédiction tout en réduisant leur complexité (93,160). Cependant, cette étape de sélection de variables est supervisée en histopathologie spectrale IR. La composition des échantillons tissulaires doit donc être parfaitement connue *a priori*. Or, de par la complexité des échantillons biologiques analysés et la difficulté d'accès à un gold-standard précis, cette composition est généralement inconnue et estimée par une étape de classification non supervisée réalisée sur l'ensemble de la gamme spectrale. Les étapes de sélection de variables et de construction d'un modèle de prédiction sont donc tributaires de cette partition estimée de façon non supervisée sur l'ensemble de la gamme spectrale.

Il est donc indispensable de proposer de nouveaux outils pour la sélection non supervisée de variables spécifiques à l'histopathologie spectrale. A cette fin, un nouvel outil pourrait être développé en s'inspirant des méthodes les plus efficaces de sélection supervisée de variables, à savoir les méthodes intégrées « embedded », pour rechercher des variables préservant la structure spatiale des structures histologiques recherchées. Cette méthode devrait donc combiner un critère de mesure de la qualité d'une partition estimée par classification non supervisée et un critère d'homogénéité spatiale des clusters constituant cette partition. Pour construire cette méthode, notre choix s'est tourné vers les algorithmes génétiques de par leur flexibilité, leur puissance et leurs applications à de l'optimisation multi-objectifs.

Cependant, ce travail a été structuré en plusieurs étapes. Dans un premier temps, nous avons

développé une nouvelle méthode de sélection non supervisée de nombres d'onde basée sur l'optimisation par algorithme génétique d'un indice de validité pour évaluer la qualité d'une partition estimée par clustering. Ce travail fait l'objet de la suite de ce chapitre.

Cet algorithme a ensuite été adapté à la recherche de variables conservant une cohérence spatiale des partitions estimées afin d'exploiter l'information spatiale intrinsèque des images spectrales IR. Cette partie du travail, pas complètement achevée, est présentée dans la partie « Résultats supplémentaires » de ce chapitre.

V. A. 2. Sélection non supervisée de variables par algorithme génétique

Le but de ce travail est d'identifier les L variables (ou nombres d'onde dans nos applications) qui partitionnent au mieux, au sens d'un indice de validité, un jeu de données en K classes estimées par clustering. L et K sont deux paramètres qui seront fixés par l'utilisateur.

L'entité de base manipulée par un algorithme génétique s'appelle un individu (ou chromosome) composé de gènes. Chaque chromosome va permettre d'explorer l'espace des solutions à la recherche de la solution optimale au problème posé. Dans ce travail, chaque gène sera une variable (ou nombre d'onde). Un individu sera donc composé de L variables.

La structure classique de notre AG est composée des étapes suivantes : i) une initialisation aléatoire de la population composée de N_P individus, ii) une évaluation des individus par des scores calculés par une fonction d'évaluation (fitness function en anglais) que nous présenterons plus bas, iii) une mise à l'échelle exponentielle de ces scores, iv) une sélection d'individus, nommés parents, par roue de la fortune, v) le croisement de ces parents pour générer de nouveaux individus appelés enfants, vi) une mutation aléatoire de quelques gènes pour certains enfants choisis aléatoirement, vii) une étape d'élitisme permettant de conserver les meilleurs individus (en terme de fonction d'évaluation) dans la génération suivante, viii) la constitution de la génération suivante composée des enfants et des élites. Les étapes ii à viii sont répétées T fois. A l'issue de cette procédure, l'individu ayant le score calculé par la fonction d'évaluation le plus faible sera retenu comme meilleure solution, c'est-à-dire qu'il sera composé des L variables (nombres d'onde) permettant d'obtenir la meilleure partition possible en termes d'indice de validité.

Dans ce travail, deux types de fonctions d'évaluation ont été testées : i) la fonction objectif de KMeans, ii) des indices de validités (101) à savoir (XB), (DB), (PBM) et (SWC).

De plus, afin d'éviter la sélection répétée d'un nombre d'onde, donc d'assurer la sélection de L variables (nombres d'onde) différentes, une contrainte de diversité des variables a été ajoutée. Appliquée à chaque individu, cette contrainte consiste à remplacer chaque répétition d'une variable par une nouvelle variable sélectionnée aléatoirement mais excluant les variables composant déjà cet individu. A chaque fois que les gènes d'un individu sont modifiés, c'est-à-dire lors de l'initialisation, du croisement, et de la mutation, cette contrainte est appliquée.

La taille de la population N_P et le nombre d'itérations T sont deux paramètres qui doivent également être fixés par l'utilisateur et qui conditionnent la convergence d'un AG vers une solution optimale.

V. A. 3. Résultats et discussion

Dans un premier temps, notre algorithme a été vérifié sur un jeu de données simulant une image spectrale IR d'un tissu biologique composé de deux structures histologiques. Parmi les 451 variables représentant des nombres d'onde dans la gamme 900-1800 cm^{-1}, seules 4 ont été simulées pour être discriminantes entre les deux groupes de spectres.

Dans un premier temps, la fonction objectif de l'algorithme KMeans a été choisie comme fonction d'évaluation de notre AG. En utilisant une recherche par grille appliquée sur la taille de la population N_P et le nombre d'itérations T, la convergence de notre algorithme et sa stabilité ont été prouvées. A l'issue de cette étape, N_P et T ont été fixés à 500. En termes de variables estimées, notre AG est apte à identifier les nombres d'onde discriminants. De plus, une précision de 100% a été obtenue sur les partitions KMeans estimées à partir des nombres d'onde trouvés par notre AG, surpassant ainsi les faibles performances obtenues (55,74%) lors de l'exploitation de toute la gamme spectrale, c'est-à-dire sans sélection de variables. Cependant, cette fonction d'évaluation a une limite. Les nombres d'onde ne sont pas estimés en fonction de leur pouvoir discriminant.

Afin d'améliorer notre algorithme, nous avons donc remplacé la fonction d'évaluation par des indices de validité. Quel que soit l'indice utilisé, la convergence et la stabilité de notre algorithme sont préservées. De plus, les nombres d'onde sont estimés dans l'ordre décroissant

de leur pouvoir discriminant, et les partitions KMeans estimées sur ces ensembles de variables sont identiques au modèle original.

Dans un second temps, nous avons testé notre méthodologie sur trois véritables jeux de données IR publiques et labélisées (spectres points) avec des complexités croissantes, en utilisant les indices de validité comme fonction d'évaluation.

Le premier jeu de données est composé des spectres de réflectance diffuse dans l'infrarouge moyen (DRIFT-MIR) acquis sur des cafés lyophilisés produits à partir de deux espèces : i) Arabica et ii) Robusta. L'application de notre méthode se fera donc en recherchant $K = 2$ classes. Bien que plus lente et plus variable à cause de la plus grande complexité de ce jeu de données par rapport aux données simulées, la convergence de notre algorithme est une nouvelle fois montrée. De plus, notre algorithme a identifié seulement 13 (3%) nombres d'onde parmi 451 comme hautement discriminants. Ces variables sont regroupées principalement sous forme de deux bandes spectrales, à savoir 1660-1670 cm^{-1} et 1690-1715 cm^{-1} attribuées à la caféine. Avec une sélection de variables par notre approche, la précision de KMeans atteint 99,14%, alors qu'elle est de 89,58% sans sélection de variables.

Le deuxième jeu de données est composé des spectres FTIR acquis en mode de réflectance totale atténuée (ATR) sur des viandes hachées fraîches de poulet, de porc et de dindes. $K = 3$ classes seront donc recherchées par notre algorithme pour ce jeu de données. Notre algorithme permet encore une fois une grande réduction de dimension. En effet, seuls 14 (3,4%) parmi les 413 nombres d'onde ont été identifiés comme hautement discriminants. Ces nombres d'ondes sont regroupés dans la bande spectrale 1530-1560 cm^{-1} spécifique de la teneur en protéines. Notre approche permet d'obtenir une précision de KMeans de 75%, qui diminue à 58% sans sélection de variables.

Le troisième jeu de données est composé de spectres ATR-FTIR acquis sur six espèces différentes de spores stockés dans deux conditions distinctes, à savoir frais ou archivé. De plus, les six espèces étudiées n'appartiennent qu'à trois genres différents, induisant une possible similitude spectrale élevée entre espèces appartenant au même genre. Ces sources multiples de variabilité spectrale et le nombre $K = 6$ de classes recherchées rendent donc ce jeu de données complexe. Face à ce challenge, notre algorithme a pu sélectionner 29 (5%) parmi 571 nombres d'onde comme discriminants. Ces nombres d'onde appartiennent aux bandes vibrationnelles 800-830 cm^{-1} et 1620-1650 cm^{-1} caractéristiques des glucides, de la chitine et des protéines qui

sont les principaux composants de ces spores. Une troisième bande spectrale discriminante moins visible apparaissant à 980-1080 cm^{-1} peut être attribuée en partie aux acides gras libres, glycolipides et phospholipides, indiquant des changements chimiques induits par les conditions de stockage. L'application de notre AG sur ce jeu de données permet à KMeans d'atteindre une précision de 77%, à comparer aux 59% obtenus en exploitant toute la gamme spectrale.

L'ensemble de ces résultats confirme l'efficacité de notre algorithme à identifier les variables discriminantes pertinentes pour améliorer les performances du clustering par KMeans dans un ensemble de données complexes. En outre, un comportement intéressant de notre algorithme est sa capacité à identifier des bandes spectrales plutôt que des nombres d'onde isolés. Ce comportement est en adéquation avec la colinéarité connue des variables dans des spectres IR d'échantillons biologiques, qui se traduit par la définition de bandes spectrales.

Un inconvénient majeur de notre méthode est sa gourmandise en termes de temps de calcul. En effet, elle est basée sur une application répétée d'un AG pour différents nombres de gènes. Dans ce travail, nous avons donc proposé deux procédures accélérées de notre AG, qui consistent à appliquer R fois notre AG pour un grand nombre de variables recherchées (par exemple 100), pour une taille de la population N_P et un nombre d'itérations T fixés. La première version ne répète que quelques fois ($R = 10$ par exemple) notre AG mais en imposant des valeurs fortes à N_P et T (par exemple 500) afin d'assurer la convergence de l'algorithme vers une solution quasi optimale. Au contraire, la deuxième version répète l'algorithme de très nombreuses fois ($R = 100$ par exemple) mais en imposant de valeurs faibles à N_P et T (par exemple 50) afin d'assurer la rapidité de l'AG. Appliquées sur les trois jeux de données réelles, ces deux versions permettent d'économiser jusqu'à 98% du temps de calcul tout en préservant la précision de la sélection de variables et du clustering.

Un article, en phase de finalisation, va être soumis au journal Chemical Science pour valoriser cette partie.

V. B. Article #2 : “Unsupervised feature selection by a genetic algorithm for mid-infrared spectral data”

Boutegrabet, W., Piot, O., Guenot, D., Gobinet, C.

A soumettre dans Chemical Science.

Unsupervised feature selection by a genetic algorithm for mid-infrared spectral data

Warda Boutegrabet[1, 2], Olivier Piot[2, 3], Dominique Guenot[1], Cyril Gobinet[2]*

[1] Université de Strasbourg (Unistra), Institut National de la Santé et de la Recherche Médicale, IRFAC Inserm U1113, 3 avenue Molière, 67200 Strasbourg, France

[2] Université de Reims Champagne-Ardenne, BioSpecT EA 7506, 51 rue Cognacq-Jay, 51097 Reims, France

[3] Platform of Cellular and Tissular Imaging (PICT), 51 rue Cognacq-Jay, 51097 Reims, France

*corresponding author

Abstract:

Dimensional reduction of highly multidimensional datasets such as those acquired by Fourier-transform infrared spectroscopy (FTIR) is a critical step in the data analysis workflow. To achieve this goal, numerous feature selection methods have been developed and applied in a supervised context, i.e. using a priori knowledge about data usually in the form of labels for classification or quantitative values for regression. For this, genetic algorithms have been largely exploited due to their flexibility and global optimization principle. However, few applications in an unsupervised context have been reported in infrared spectroscopy. The aim of this article is to propose a new unsupervised feature selection method based on a genetic algorithm using a fitness function. The most efficient fitness function is based on the measurement of a validity index value computed from KMeans partitions. Evaluated on a simulated dataset and validated and tested on three real-world infrared spectroscopic datasets, our developed algorithm is able to find the spectral descriptors improving clustering accuracy and simplifying the spectral interpretation of results.

1. Introduction

Infrared (IR) spectroscopy is a recognized label-free analytical technique to probe the biochemical composition of samples with minimal preparation. Based on IR light absorption by chemical bonds of molecules composing the analyzed sample, the versatility of this technology has been proved by its numerous applications on different sample types, such as cells, tissues, solids and powders, in very different scientific fields, such as food quality [1], biological [2], biomedical [3] , pharmaceutical [4] and forensic [5] sciences. The highly multidimensional nature of IR spectra makes this technology highly informative about the sample composition. This apparent intrinsic complexity can however be managed thanks to recent advances in chemometrics permitting to disentangle the useful information. Clustering, classification and regression [6,7] are multivariate data analysis tools that have been widely applied to IR spectroscopy and contributed to its popularity.

However, working on the full-spectrum range can degrade the performance of the above-cited chemometrics methods for two reasons: i) among the hundreds of spectroscopic variables or features, many can be useless or uninformative, ii) the high dimensionality of spectroscopic data coupled to the usually limited number of analyzed samples makes evident the curse of dimensionality [8]. A usually advised and currently applied solution is thus to reduce the spectroscopic data dimensionality. This critical step can be done (i) either by feature extraction [9] performed by unsupervised methods such as Principal Component Analysis, Independent Component Analysis, Vector Quantization or Non-negative Matrix Factorization, or by supervised methods such as Partial Least Squares, (ii) or either by feature selection [10] performed for example by a Genetic Algorithm (GA). A GA is a type of evolutionary algorithms inspired by the process of natural selection and genetics in biological evolution to resolve complex optimization problems [28]. A GA starts with the initialization of a population composed of individuals, each representing a potential solution to the optimization problem. Then, this population evolves by applying successive operators named chromosome evaluation, parent selection, parent cross-over and mutation to generate new individuals, called children. Next, an elitism process saves the best few parents. Children and elite parents thus compose the next population of individuals. Numerous repetitions of these steps ensure the convergence of the algorithm to a sub-optimal solution of the defined problem.

While feature extraction transforms the data into a new coordinate space, feature selection just selects some features in the original coordinate space without any transformation. Interpretation of data is thus simpler with feature selection than with feature extraction.

In IR spectroscopy, feature selection is popular when associated with supervised classification since it eases spectral, chemical and biological interpretations, reduces model complexity and improves classification accuracy [11–14]. Due to its flexibility and ability to efficiently explore large search spaces to estimate near-optimal solutions in complex problems, GA [15] has been largely applied for such supervised feature selection in infrared spectroscopy [11,16–18].

However, labeled data are not always available, preventing the application of such supervised feature selection. In this article, we thus propose a new feature selection method based on the application of a GA characterized by two main originalities. Firstly, the proposed approach is unsupervised by definition and thus not requires data ground truth. Secondly, a new efficient GA fitness function is constructed using a validity index computed from KMeans partitions estimated on the data.

The efficiency of our method is proved on simulated and real-world infrared spectroscopy datasets by comparing performance of KMeans when applied on the full-spectral range versus on the GA selected feature subset.

2. Methodological development of a new feature selection method using a genetic algorithm

The working hypothesis of the proposed method is that, in an unsupervised context, the discriminant features optimize a quality measure of a data partition estimated by unsupervised classification. In this work, KMeans was chosen as the clustering method.

2.1. KMeans clustering

K-means is one of the simplest and most popular unsupervised classification algorithms aiming at partitioning a dataset $\boldsymbol{X} = \{\boldsymbol{x}_i / \boldsymbol{x}_i \in \mathbb{R}^D\}_{i=1}^N$ of N objects composed of D features into K disjoint clusters defining a partition $\boldsymbol{P} = \{\boldsymbol{P_1}, \boldsymbol{P_2}, \dots, \boldsymbol{P_K}\}$ by minimizing the total within-cluster variation defined as:

$$KM(\boldsymbol{P}) = \sum_{k=1}^{K} \sum_{\boldsymbol{x}_i \in \boldsymbol{P_k}} \|\boldsymbol{x}_i - \boldsymbol{c}_k\|^2 \quad (1)$$

where $\|.\|^2$ is the squared Euclidean norm and $\boldsymbol{c}_k$ is the centroid of $\boldsymbol{P_k}$. Each cluster k is thus composed of N_k objects.

After a random initialization, the convergence of the algorithm is ensured by the iterative and alternated updates of the object assignation to clusters and of their centroids [19].

A natural measure of data partition quality can thus be the KMeans objective function $KM(\boldsymbol{P})$. However, mathematical tools referred as validity indices have been specifically designed to objectively measure the quality of a partition [20], and have thus been investigated in this work.

2.2. Validity index for partition quality evaluation

A validity index is a mathematical function designed to estimate the quality of a clustering solution. Traditionally, it is used to compare several partitions estimated by the same clustering algorithm using different parameter values, such as the number of clusters [20].

Most of the existing validity indices are functions of cluster cohesion (within-group dispersion) and cluster separation (between-group dispersion). The best partition maximizes both the inter-cluster separation and the intra-cluster cohesion which have numerous different possible mathematical definitions, leading to the development of various validity indices in literature. From the results of previous studies of our group on vibrational spectroscopic datasets of biological samples [21–23], the validity indices identified as the most efficient in combination with KMeans, that were consequently used in the present work, were Xie-Beni (XB) [24], Alternative Simplified Silhouette Width Criterion (ASSWC) [25], Davies-Bouldin (DB) [26] and Pakhira-Bandyopadhyay-Maulik (PBM) [27].

Xie-Beni (XB). This index is defined as the ratio between the average cluster compactness $\pi = \frac{1}{N}\sum_{k=1}^{K} \sum_{\boldsymbol{x}_i \in \boldsymbol{P_k}} \|\boldsymbol{x}_i - \boldsymbol{c}_k\|^2$ and the minimum between-cluster separation measure $s = \min_{l \in \{1,\dots,K\}, m \in \{1,\dots,K\}, l \neq m} \{\|\boldsymbol{c}_l - \boldsymbol{c}_m\|^2\}$, i.e $XB(\boldsymbol{P}) = \frac{\pi}{s}$. The best partition minimizes the XB index value [24].

Alternative Simplified Silhouette Width Criterion (ASSWC). This index is a variant of the well-

known but computationally expensive Silhouette Width Criterion (SWC) index [20]. It is defined as $ASSWC(\boldsymbol{P}) = \frac{1}{N}\sum_{i=1}^{N} s(\boldsymbol{x}_i)$ where $s(\boldsymbol{x}_i) = \frac{b_i}{a_i+\varepsilon}$, with $a_i = \|\boldsymbol{x}_i - \boldsymbol{c}_k\|$ being the distance between the data point $\boldsymbol{x}_i$ belonging to cluster k and its centroid $\boldsymbol{c}_k$, $b_i = \min_{m\in\{1,\dots,K\},m\neq k}\{\|\boldsymbol{x}_i - \boldsymbol{c}_m\|\}$ is the distance between the data point $\boldsymbol{x}_i$ and the centroid of the nearest neighbor cluster. The term ε is a small constant (fixed to 10^{-6} in this work) used to prevent division by zero when $a_i = 0$. The best partition maximizes the ASSWC index value [25].

Davies-Bouldin (DB). This index is defined as the average ratio between the compactness and separation of each cluster with its most similar neighbor cluster:

$$DB(\boldsymbol{P}) = \frac{1}{K}\sum_{k=1}^{K} \max_{m\in\{1,\dots,K\},m\neq k}\left\{\frac{\left(\frac{1}{N_k}\sum_{\boldsymbol{x}_i\in\boldsymbol{P_k}}\|\boldsymbol{x}_i - \boldsymbol{c}_k\|^2 + \frac{1}{N_m}\sum_{\boldsymbol{x}_i\in\boldsymbol{P_m}}\|\boldsymbol{x}_i - \boldsymbol{c}_m\|^2\right)}{\|\boldsymbol{c}_k - \boldsymbol{c}_m\|}\right\}$$

The best partition minimizes the DB index value [26].

Pakhira-Bandyopadhyay-Maulik (PBM). This index is defined as the square ratio between a normalized cluster separation defined as $D_N = \frac{\max_{l,m=1,\dots,K}\|\boldsymbol{c}_l - \boldsymbol{c}_m\|}{K}$ and a normalized cluster cohesion defined as $E_N = \frac{\sum_{k=1}^{K}\sum_{\boldsymbol{x}_i\in\boldsymbol{P_k}}\|\boldsymbol{x}_i - \boldsymbol{c}_k\|}{\sum_{i=1}^{N}\|\boldsymbol{x}_i - \bar{\boldsymbol{x}}\|}$ where $\bar{\boldsymbol{x}}$ is the mean data object, i.e. $PBM(\boldsymbol{P}) = \left(\frac{D_N}{E_N}\right)^2$. The best partition maximizes the PBM index value [27].

2.3. Unsupervised variable selection by GA

Using KMeans clustering and validity indices, we developed a new feature selection approach based on a GA which will be described below.

In this work, the general procedure of GA descripted in the introduction has been adapted to propose a new unsupervised feature selection method improving the clustering of vibrational spectroscopic data. The general architecture of this proposed GA is depicted in Figure 1, and its detailed description is presented below.

Problem definition and encoding. Consider the vector $\boldsymbol{w} \in \mathbb{R}^D$ composed of the D different variables for which each object $\boldsymbol{x}_i$ has been observed. The aim of this study is to identify the L most discriminant

variables of $\boldsymbol{w}$ to partition the dataset $\boldsymbol{X}$ into K clusters using unsupervised classification. A solution to this problem is encoded by an individual, named chromosome, $\boldsymbol{z}_n \in \mathbb{R}^L$ composed of L genes. Each gene $z_{n,l}$, $1 \leq l \leq L$, is the value of a variable composing $\boldsymbol{w}$. Thus, $\boldsymbol{z}_n$ is composed of L variable values.

Population initialization. At each iteration t, the population is composed of N_P chromosomes, also called a generation, mathematically represented by the set $Z^{(t)} = \left\{\boldsymbol{z}_n^{(t)} \,/\, \boldsymbol{z}_n^{(t)} \in \mathbb{R}^L\right\}_{n=1}^{N_P}$. The initial population $\boldsymbol{Z}^{(0)}$ was generated by uniformly and randomly selecting N_P solutions in the search space, i.e. each initial chromosome $\boldsymbol{z}_n^{(0)}$, $1 \leq n \leq N_P$, is generated by randomly selecting L variables from $\boldsymbol{w}$.

Constraint to ensure feature diversity in chromosomes. In order to ensure the selection of L different features for each chromosome, i.e. to avoid the selection of a same feature several times for the same chromosome, a feature diversity constraint was applied to each chromosome each time its composition is modified. This constraint consists in replacing each repeated feature in a chromosome $\boldsymbol{z}_n^{(t)}$ by a new variable value uniformly and randomly selected in $\boldsymbol{w}$ excluding the variable values composing the chromosome $\boldsymbol{z}_n^{(t)}$, i.e. $\boldsymbol{w} \setminus \boldsymbol{z}_n^{(t)}$ in mathematical notations.

Chromosome evaluation. At each iteration *t*, the N_P chromosomes composing the population compete. The quality of each chromosome $\boldsymbol{z}_n^{(t)}$ is quantified by an evaluation metric, named a fitness function, $F: \mathbb{R}^L \to \mathbb{R}$, $\boldsymbol{z}_n^{(t)} \longmapsto F\left(\boldsymbol{z}_n^{(t)}\right)$. The lower the fitness function score, the nearer the chromosome is from the optimal solution of the defined problem. This fitness function must be carefully defined since it is closely related to the problem objective. As previously mentioned, the aim of this work being to realize an unsupervised selection of the variables which provide the best partition $\boldsymbol{P}$ of the dataset $\boldsymbol{X}$ into K clusters, the fitness function can be (i) defined as the KMeans objective function $KM(\boldsymbol{P})$, or (ii) based on one of the possible VI, such as the XB validity index value $XB(\boldsymbol{P})$, the inverse of the ASSWC validity index value $ASSWC(\boldsymbol{P})$, the DB validity index value $DB(\boldsymbol{P})$, the inverse of the PBM validity index value $PBM(\boldsymbol{P})$. The evaluation of the fitness function for each chromosome $\boldsymbol{z}_n^{(t)}$ is realized as follows: (i) each object $\boldsymbol{x}_i$ of the dataset $\boldsymbol{X}$ is reduced to the L data values acquired at the L variable values composing $\boldsymbol{z}_n^{(t)}$, giving the reduced dataset $\widetilde{\boldsymbol{X}}^{\boldsymbol{z}_n^{(t)}}$, (ii) the KMeans algorithm is applied on $\widetilde{\boldsymbol{X}}^{\boldsymbol{z}_n^{(t)}}$,

estimating the partition $\boldsymbol{P}(\widetilde{\boldsymbol{X}}^{\mathbf{z}_n^{(t)}})$, (iii) the fitness function score $F\left(\mathbf{z}_n^{(t)}\right)$ is computed on the estimated partition $\boldsymbol{P}(\widetilde{\boldsymbol{X}}^{\mathbf{z}_n^{(t)}})$.

Fitness scaling. To ensure convergence of the algorithm and to completely explore the search space, it is necessary to scale the scores of the fitness function over the whole population. In this work, an exponential scaling was applied to the fitness function scores as follows: $F_s\left(\boldsymbol{z}_n^{(t)}\right) = \left(F\left(\boldsymbol{z}_n^{(t)}\right)\right)^{E(t)}$ with $E(t) = 0.1\tan\left(\frac{t}{T+1} \times \frac{\pi}{2}\right)$ where $t \in \{0, \dots, T\}$ is the current iteration and T is the total number of iterations [29].

Parent selection. At each iteration t, only a portion of the chromosomes is randomly selected as parents for breeding, according to their scaled fitness function scores. In this study, the roulette wheel strategy [30] has been used to select $2(N_P - N_e)$ parents with a probability proportional to the inverse of their scaled fitness values $F_s\left(\boldsymbol{z}_n^{(t)}\right)$, where N_e is the number of elite parents as explained below. In this way, this rule promotes the selection of chromosomes with the lowest fitness values. Combined with the exponential scaling, the roulette wheel strategy will ensure a high diversity into the selected parents during the first iterations since $E(t) \approx 0$ thus $F_s\left(\boldsymbol{z}_n^{(t)}\right) \approx 1$ for each chromosome when $t \ll T$. On the contrary, during the last iterations, this strategy will focus on a specific region of the search space, i.e. it will precisely refine the solution, by privileging the selection of chromosomes with the smallest fitness function values. At the end of this step, the selected parents compose the set $Y^{(t)} = \left\{\boldsymbol{y}_n^{(t)} \,/\, \boldsymbol{y}_n^{(t)} \in \mathbb{R}^L\right\}_{n=1}^{2(N_P - N_e)}$.

Parent crossover. The goal of crossover is to combine parts of the genetic information of pairs of parents belonging to $Y^{(t)}$ in order to generate children chromosomes $V^{(t)} = \left\{\boldsymbol{v}_n^{(t)} \,/\, \boldsymbol{v}_n^{(t)} \in \mathbb{R}^L\right\}_{n=1}^{N_P - N_e}$. In this work, the uniform crossover has been applied, i.e. each gene

$v_{n,l}^{(t)}$ of each child chromosome $\boldsymbol{v}_n^{(t)}$ is randomly selected from the corresponding parent genes $\left\{y_{2n-1,l}^{(t)} \; ; \; y_{2n,l}^{(t)}\right\}$ using an uniform distribution. This crossover operator can generate chromosomes containing repeated features. The constraint to ensure feature diversity in chromosomes defined above is thus applied on each generated child.

Mutation. Mutation corresponds to the random change of gene values to maintain genetic diversity from one generation to the next and to avoid the premature algorithm convergence to a local optimum [31]. In this work, a uniform mutation has been applied to the children chromosomes $V^{(t)}$ [15]. Consequently, each gene has a probability P_M to be replaced by another feature value w randomly selected in $\boldsymbol{w} \in \mathbb{R}^D$ using an uniform distribution. This mutation operator can generate chromosomes containing repeated features. The constraint to ensure feature diversity in chromosomes defined above is thus applied on each mutated chromosome. In this work, P_M was empirically fixed to 0.1.

Elitism. To ensure that the quality of the estimated solutions will not decrease through generations, the elitism operator has been applied [15]. It consists in preserving the N_e unaltered best chromosomes, i.e. those having the smallest fitness function scores, from the current generation $Z^{(t)}$ to the next $Z^{(t+1)}$. In this work, N_e was fixed to 2.

Next generation. The next population $\mathbf{Z}^{(t+1)} = \left\{\mathbf{z}_n^{(t+1)} \,/\, \mathbf{z}_n^{(t+1)} \in \mathbb{R}^{\mathrm{L}}\right\}_{n=1}^{N_P}$ is composed of the N_e elite chromosomes and of the $N_P - N_e$ children generated by the crossover operator. Through iterations, the population size is thus constant and equal to N_P.

Stopping criterion. In order to ensure the convergence of the proposed GA, the steps from the chromosome evaluation to the next generation are repeated T times, i.e. till the maximal number of iterations T fixed by the user is reached. At the end, the chromosomes of the last population $\mathbf{Z}^{(T+1)}$ are evaluated and the one with the smallest fitness value is considered as the best solution

to the encoded problem and is denoted by $\boldsymbol{z}_{best}$. Two different information are thus available, namely the L most discriminant features composing $\boldsymbol{z}_{best}$ and the corresponding KMeans partition $\boldsymbol{P}(\boldsymbol{z}_{best})$.

The convergence of a GA is known to be highly dependent on the fitness function F, the number of iterations T and the population size N_P. The impact of these parameters on our proposed GA will be discussed in the "Results and discussion" section.

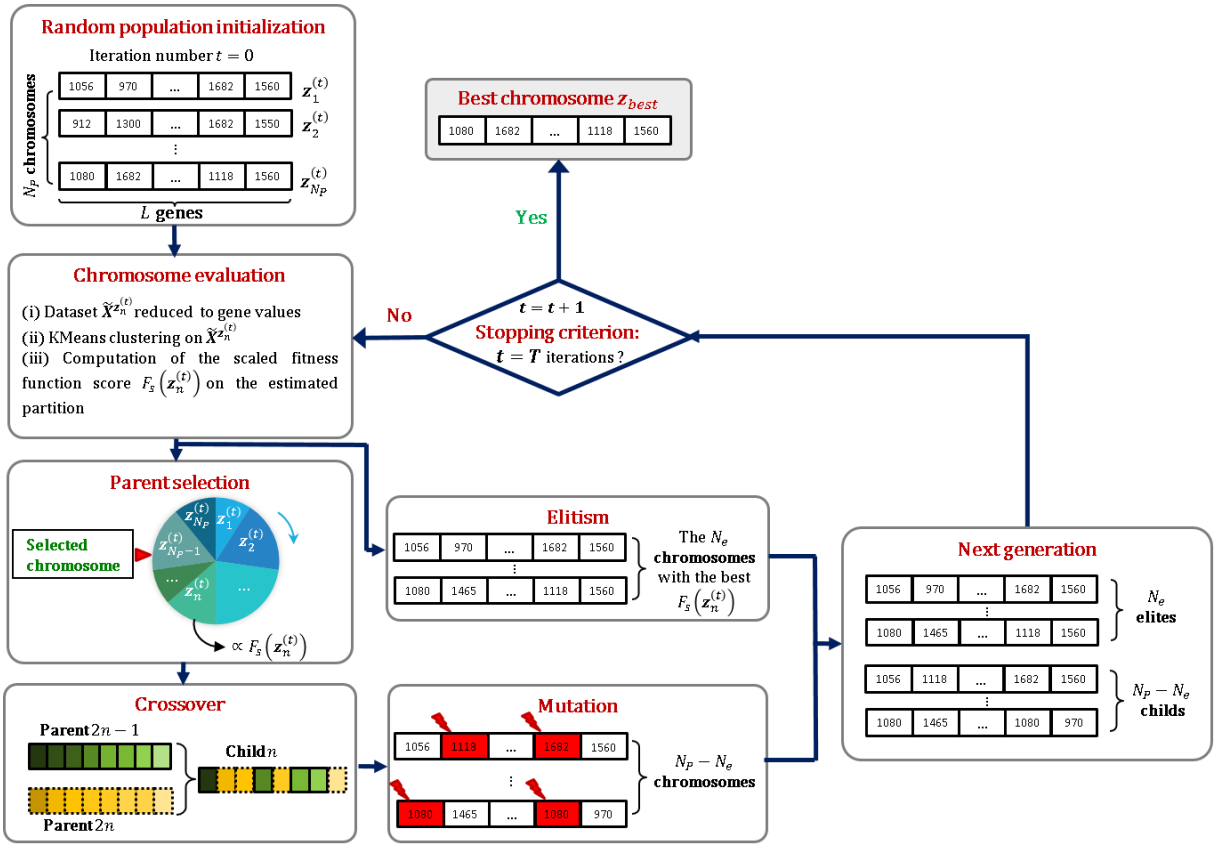


Figure 1. Flowchart of the developed GA for unsupervised feature selection. The fitness function can be the KMeans objective function or based on a VI measure.

3. Dataset description

To assess the performance of the proposed unsupervised feature selection algorithm based on GA, two types of datasets were used, i.e. an artificial spectral image and real Fourier transform infrared (FTIR) spectral datasets.

3.1. Simulated spectral dataset

Before being tested on real-life datasets, the performance of a new algorithm is usually evaluated on a completely under-control simulated dataset [22]. In consequence, the efficiency of our proposed algorithm was first tested on a simulated dataset mimicking infrared data of a biological tissue acquired in transmission mode. To simplify, we chose to construct an artificial spectral image split into two equivalent parts as if this construction corresponds to two histological structures whose respective spectral signatures are very similar. This artificial spectral image is composed of $10 \times 20 = 200$ spectra, each containing $D = 451$ variables representing the acquired wavenumbers $\boldsymbol{w} = \{w_j \in \mathbb{R}\}_{j=1}^{451} \in \mathbb{R}^{451}$ from 900 to 1800 cm^{-1} with a step of 2 cm^{-1}, i.e. $\boldsymbol{w} = \{900,902, \dots ,1798,1800\} \in \mathbb{R}^{451}$. The unfolded simulated spectral image is thus represented by the set $\boldsymbol{X} = \{\boldsymbol{x}_i / \ \boldsymbol{x}_i \in \mathbb{R}^{451}\}_{i=1}^{200}$. Each half of this image is associated with spectra of a specific artificial histological structure, i.e. the two structures are respectively composed of the subsets of spectra $\boldsymbol{X^{(1)}} = \left\{\boldsymbol{x}_i^{(1)} / \ \boldsymbol{x}_i^{(1)} \in \mathbb{R}^{451}\right\}_{i=1}^{100} = \{\boldsymbol{x}_i \in \boldsymbol{X}\}_{i=1}^{100}$ and $\boldsymbol{X^{(2)}} = \left\{\boldsymbol{x}_i^{(2)} / \ \boldsymbol{x}_i^{(2)} \in \mathbb{R}^{451}\right\}_{i=1}^{100} = \{\boldsymbol{x}_i \in \boldsymbol{X}\}_{i=101}^{200}$. The spatial distribution of these two simulated histological structures is visible in Figure 2(a).

From the mean spectrum $\bar{\mathbf{x}} = \{\bar{x}_j \in \mathbb{R}\}_{j=1}^{451}$ of a real FTIR image acquired on a formalin-fixed paraffin-embedded normal human côlon tissue section, each intensity $x_{ij}^{(k)}$ of the ith simulated spectrum at the jth wavenumber w_j for the kth histological structure was generated using the following model:

$$x_{ij}^{(k)} = \bar{x}_j + \xi_{ij}^{(k)}$$

where $\xi_{ij}^{(k)} \sim \mathcal{N}\left(m_j^{(k)}, \sigma_j^{(k)}\right)$ is a Gaussian random variable of mean $m_j^{(k)}$ and standard deviation $\sigma_j^{(k)}$, for each $i \in \{1, \dots ,100\}$, each $j \in \{1, \dots ,451\}$ and each $k \in \{1,2\}$.

Among the 451 wavenumbers, the four wavenumbers $\boldsymbol{z} = \{w_{79} = 1056, w_{110} = 1118, w_{331} = 1560, w_{392} = 1682\}$ cm^{-1} were randomly selected to be the discriminant wavenumbers between the two simulated histological structures. Thus, for these wavenumbers, $\xi_{ij}^{(k)}$ was generated using the parameters described in **Table 10**. These discriminant wavenumbers were furthermore characterized by their discriminating degree, i.e. their ability to discriminate the two histological structures, shown on the last line of **Table 10**. The discriminating degree δ_j of the jth wavenumber w_j is here defined as the absolute value of the difference between the means of the two histological structures evaluated at the jth wavenumber w_j, i.e. $\delta_j = \left|\bar{x}_j^{(1)} - \bar{x}_j^{(2)}\right| = \left|\left(\bar{x}_j + m_j^{(1)}\right) - \left(\bar{x}_j + m_j^{(2)}\right)\right| = \left|m_j^{(1)} - m_j^{(2)}\right|$. As can be seen in this table, these four wavenumbers have different discriminating degrees, decreasing from $z_1 = w_{79} = 1056$ cm^{-1} to $z_4 = w_{392} = 1682$ cm^{-1}.

For the remaining 447 wavenumbers, we have chosen $m_j^{(1)} = m_j^{(2)} = 0$ (resulting in a $\delta_j = 0$) and $\sigma_j^{(1)} = \sigma_j^{(2)} = 0.01$, in order to simulate non-discriminant wavenumbers.

Table 10. Means and standard deviations of the Gaussian distributions used to generate the synthetic dataset, and the resulting discriminating degrees, at the four discriminant wavenumbers.

Variable number *j*	79	110	331	392
Wavenumber w_j (cm^{-1})	1056	1118	1560	1682
Mean $m_j^{(1)}$	0.014	0.010	0.006	0.002
Standard deviation $\sigma_j^{(1)}$	0.001	0.001	0.001	0.001
Mean $m_j^{(2)}$	-0.014	-0.010	-0.006	-0.002
Standard deviation $\sigma_j^{(2)}$	0.001	0.001	0.001	0.001
Discriminating degree δ_j	0.028	0.020	0.012	0.004

The mean spectrum of this simulated dataset, its one standard deviation envelope and the four discriminant wavenumbers are represented in Figure 2(b). In order to highlight the four discriminant wavenumbers, Figure 2(c) shows the absolute value of the difference between the mean spectra associated with the two artificial histological structures, i.e. the experimental discriminating degree δ at each wavenumber.

Then, our algorithm was tested on three real publicly-accessible datasets, corresponding to IR analysis of lyophilized coffees, fresh meats and fresh and archived fungal spores.

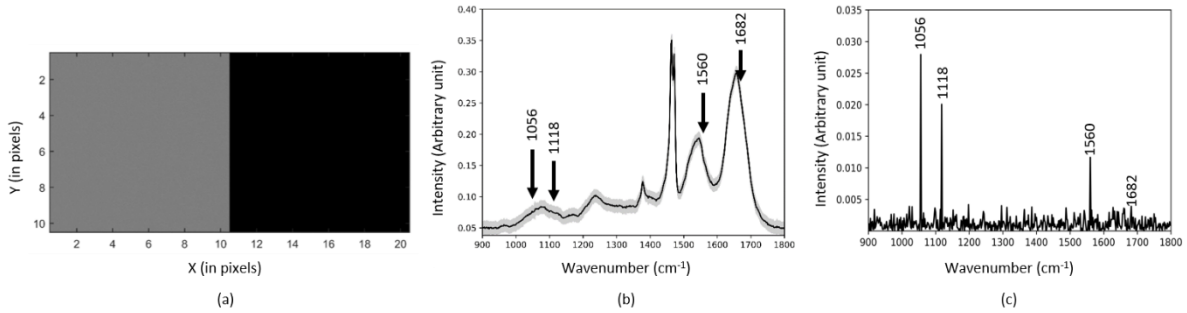


Figure 2. Simulated dataset. (a) Spatial distribution of the two histological structures represented by gray and black pixels respectively. (b) Mean spectrum (solid black line) and the one standard deviation envelope (shaded area) of the simulated dataset. The four discriminant wavenumbers are identified by black arrows. (c) Experimental discriminating degree δ at each wavenumber. The four discriminant features are identified by their wavenumbers.

3.2. DRIFT-MIR public dataset of lyophilized coffees

This dataset is composed of 56 diffuse reflectance Fourier transformed mid-infrared (DRIFT-MIR) spectra acquired in the 810-1910 cm^{-1} spectral range on lyophilized coffees produced from two species: (i) 29 samples of Arabica and (ii) 27 samples of Robusta. A complete description of acquisition parameters and of dataset characteristics is available in [32,33].

In order to reduce the data variability mainly due to physical effects, Extended Multiplicative Signal Correction (EMSC) has been applied to neutralize the baseline variability modeled by a polynomial function and to normalize the spectra on the mean dataset spectrum [34]. In this

study, a 0-order polynomial function has been determined as the most efficient for our feature selection algorithm (data not shown).The means of the spectra preprocessed by EMSC and their one-standard deviation envelopes are presented for each coffee variety on Figure 3(a). The main spectral differences between the two coffee species are visible in the 1650-1750 cm^{-1} spectral range.

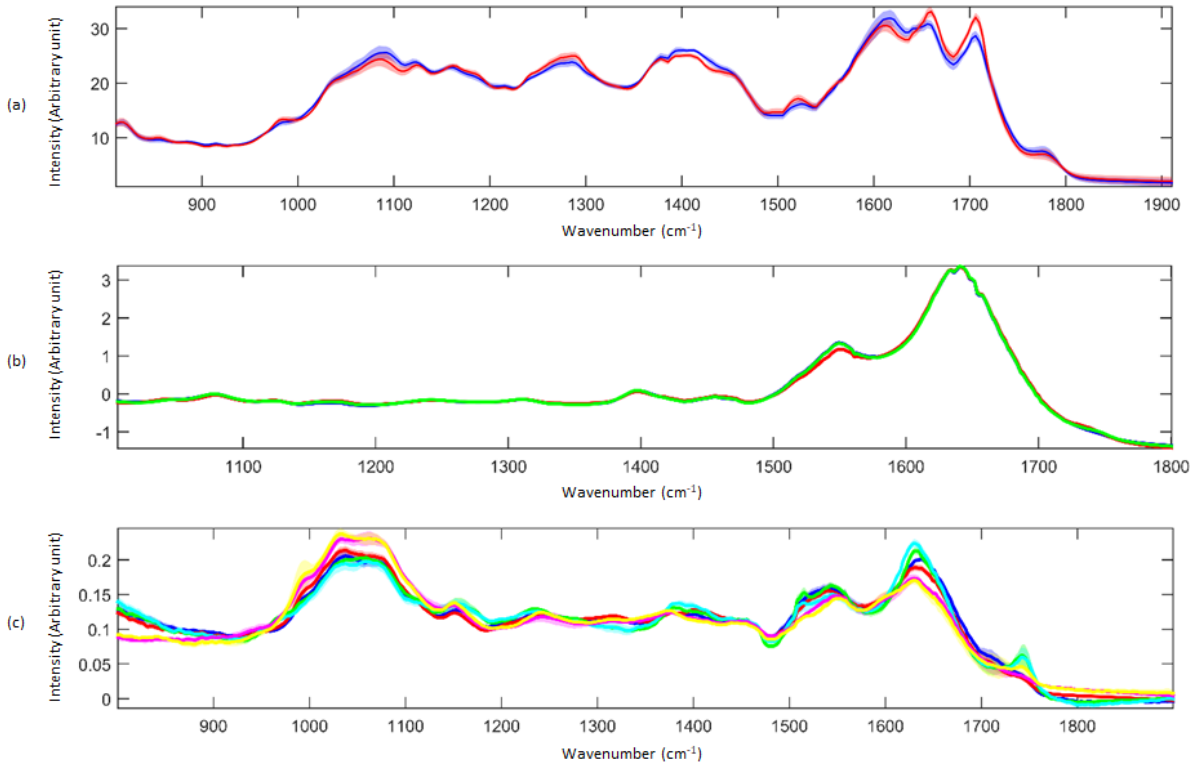


Figure 3. EMSC preprocessed mean spectra (solid lines) and their one standard deviation envelopes (shaded areas) for: (a) the lyophilized coffee dataset (Arabica in blue and Robusta in red), (b) the fresh meat dataset (chicken in blue, pork in red and turkey in green), (c) the fresh and archived fungal spore dataset (Geastrum triplex in blue, Scleroderma citrinum in red, Scleroderma areolatum in green, Lycoperdon perlatum in cyan, Lycoperdon pyriforme in magenta, Geastrum fimbriatum in yellow).

3.3. ATR-FTIR public dataset of fresh meats

The second real dataset is composed of 60 attenuated total reflectance (ATR)-FTIR spectra recorded in the 800 to 4000 cm^{-1} spectral range on fresh minced meats of (i) 20 chicken, (ii) 20 pork and (iii) 20 turkey independent samples. The spectra were limited to the 1000-1800 cm^{-1} spectral range according to the full experimental details described in [35].

These spectra were preprocessed by EMSC to correct the baseline effects and to normalize their intensities [34]. To model the baseline, a second order polynomial function was chosen because optimizing KMeans clustering applied on the full 1000-1800 cm^{-1} spectral range. The mean spectra and their one-standard deviation envelopes of each species are shown in Figure 3(b). As can be seen, the spectra of each species are very similar. The major spectral differences between the three species can be noted in the 1520-1565cm^{-1} spectral range specific to protein contribution according to [35].

3.4. ATR-FTIR public dataset of fresh and archived fungal spores

The third dataset consists of six fresh and archived (i.e. stored at room temperature) fungal spore samples belonging to three different genera and two different species per genus (Lycoperdon. perlatum, Lycoperdon pyriforme, Scleroderma areolatum, Scleroderma citrinum, Geastrum triplex, Geastrum fimbriatum), analysed by ATR-FTIR spectroscopy.

In order to decrease the computational cost, 15 spectra per species in the 800-1900 cm^{-1} spectral range have been considered in our study. For a full experimental and spectral data description, the reader can refer to [36].

The spectra were preprocessed by EMSC [34] using a second-order polynomial function to model the baseline in accordance with the original paper [36]. The means of the preprocessed spectra with their one-standard deviation envelopes are depicted in Figure 3(c). Spectral differences between species are visible in the 800-900 cm^{-1}, 1000-1100cm^{-1} and 1600-1660 cm^{-1} spectral ranges specific to carbohydrate and protein signatures.

3.5. Programming environment

Data analysis realized in this work, including pre-processing, KMeans clustering and feature selection by GA, was carried out using in-house scripts written in Matlab (The Mathworks, Natick, MA) run on a computer equipped with a 3.20 GHz Intel® Xeon® W-2104 processor, 32 Go RAM and 4 cores.

4. Results and discussion

The evaluation of the proposed algorithm was first performed on the simulated dataset and on the DRIFT-MIR public dataset of lyophilized coffees, described in the previous section. For each of these two datasets, the stability of the proposed algorithm in terms of the chosen fitness functions and in terms of the selected variables has been first analyzed. Then, the impact of the variable selection on the KMeans clustering accuracy has been investigated. Finally, our proposed algorithm was validated on the two other public datasets described previously, i.e., the ATR-FTIR public datasets of fresh meats and of fresh and archived fungal spores.

4.1. Benefits of feature selection by GA on the simulated dataset

For this dataset, our goal is to find the best variable subset permitting to distinguish the two spectally-simulated histological structures as presented on Figure 2(a).

4.1.1. Failure of full-spectrum KMeans clustering

To justify the necessity of variable selection for this dataset, a two-class KMeans clustering was first performed on the full-range spectra composed of 451 wavenumbers from 900 cm^{-1} to 1800 cm^{-1}, i.e. without variable selection. As shown on Figure 4(a), the estimated KMeans partition is randomly distributed and is thus completely uncorrelated to the spatial distribution of the two simulated histological structures shown on Figure 2(a). Indeed, the pixels belonging to the first structure are distributed at 63% in cluster 1 and 37% in cluster 2, and for the second structure

at 73% in cluster 1 and 27% in cluster 2, resulting in a clustering accuracy of 55%, explaining the poor quality of the estimated partition.

Furthermore, the analysis of the estimated centroids of Figure 4(b) does not reveal any significant difference between the two clusters, which was expected due to the way that the dataset was generated. In addition, the difference spectrum does not permit to enhance any significant differences (Figure 4(c)). Indeed, the four discriminants wavenumber of the model, represented by red crosses on Figure 4(c), do not present higher intensity than the other uninformative wavenumbers.

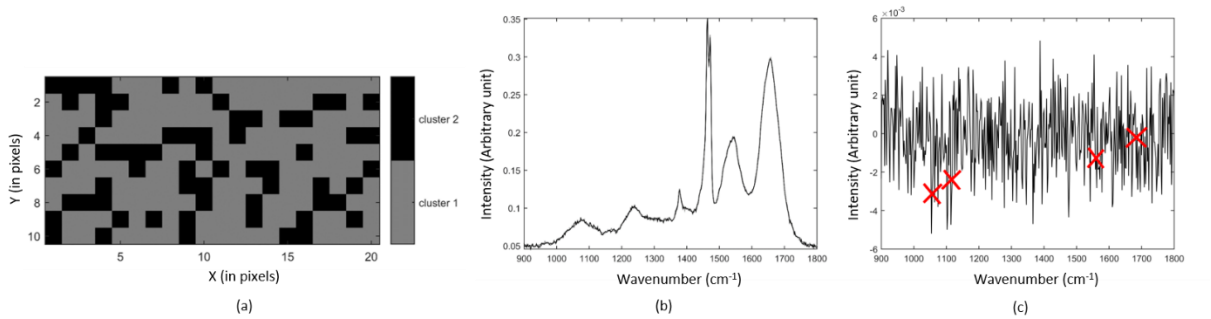


Figure 4. KMeans results estimated on the simulated FTIR spectral image using the full spectral range 900-1800 cm^{-1}. (a) Partition where grey and black pixels correspond to spectra attributed to the first and second clusters respectively. (b) The centroids of the two estimated clusters. (c) Difference spectrum between the centroids of the two estimated clusters. The intensities at the four discriminant wavenumbers are displayed by red crosses.

However, KMeans clustering is known to be highly dependent on the initialization step. In order to show that the previous results were not marginal and possibly due to a specific bad initialization, KMeans was run with 100 different initializations. The sensitivities and accuracy for the two simulated tissue structures were computed for each run in order to evaluate the performance of the KMeans, and summarized in term of mean ± standard deviation. The sensitivities were evaluated to (53.79 ± 26.25) % for the first structure and to (57.69 ± 25.55) % for the second structure, proving the highly variability of the KMeans results. Overall, the

accuracy was estimated to (55.74 ± 4.48) % showing the limitation inefficiency of KMeans to retrieve the real classes when applied to the full-range spectra. In fact, this can be explained by two factors: (i) the high number (447) of irrelevant features compared to the small number (4) of discriminant features and (ii) the high variability of the irrelevant features compared to the weak intensity of the discriminant features. These two effects misguided the KMeans algorithm which converged to wrong solutions.

These results clearly demonstrate the need to perform feature selection in order to improve the clustering accuracy. The following of this article will thus be dedicated to this task using a GA combined to an objective fitness function, for making this processing totally unsupervised.

4.1.2. Success of the proposed feature selection GA with the fitness function defined as the KMeans objective function

In a first experiment, the GA fitness function was defined as the KMeans objective function.

Before evaluating the performance of the proposed approach, it is important to study its convergence and its stability in term of fitness value, and its repeatability in term of the estimated discriminant features. Indeed, the convergence of a GA is highly sensitive to the choice of its parameters. Choosing the appropriate number of iterations T and the population size N_P is thus a key aspect of GA convergence. It is thus important to study its stability and repeatability in function of these parameters.

Furthermore, conceptually, a higher number of genes L leads to a more complex problem to be solved by the GA, inducing the need to fix higher values of T and N_P to ensure convergence. For this reason, we realized the convergence and stability study of the GA for a gene number $L = 10$ greater than the real number of discriminant features fixed to 4 in this simulated dataset. In order to find the $L = 10$ most discriminant wavenumbers partitioning the simulated IR spectral image into two clusters, the convergence analysis of the proposed algorithm was realized by a grid search implemented for $N_P \times T \in \{50,100, \dots ,450,500\}^2$. Furthermore, for

each couple of N_P and T, the algorithm was repeated 10 times in order to evaluate its stability or variability.

The results of this experiment are visible in Figure 5(a) depicting the mean and the standard deviation of the fitness function score in function of N_P and T over the 10 repetitions. Whatever the population size, the higher the number of iterations, the lower the mean and standard deviation of the fitness function score. The number of iterations is thus a crucial parameter for our proposed method, whereas the fitness function score is relatively independent of the population size.

More precisely, when the number of iterations is higher than 250, the proposed algorithm reaches a stability area characterized by small means and weak standard deviations of the fitness function score, proving the repeated GA convergence to close solutions over the 10 repetitions. To ensure the algorithm convergence, it is vital to select the number of iterations greater than 250, whatever the number of searched discriminant features L.

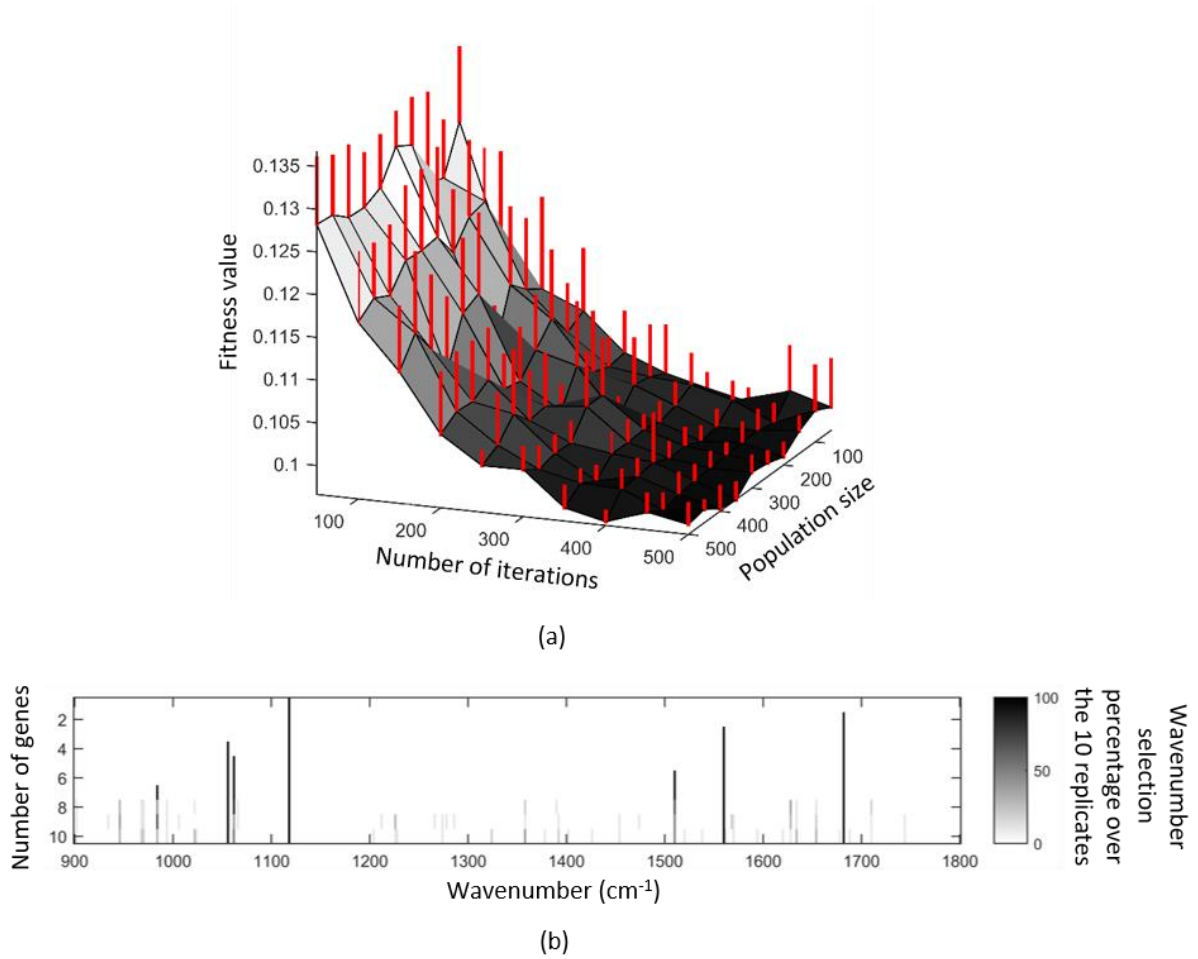


Figure 5. Study of the stability of the proposed GA on the simulated FTIR spectral image using KMeans objective function as the fitness function. (a) Mean fitness function value computed over 10 replicates in function of the population size and the number of iterations in order to find the 10 most discriminant features. Red error bars represent the standard deviation of the fitness function value. (b) Wavenumber selection percentage over 10 replicates in function of the number of searched discriminant features.

After having defined the conditions of stability and convergence, it is now important to evaluate the ability of the algorithm to select correctly the relevant discriminant wavenumbers and how it behaves in function of the number of searched discriminant features L. For this purpose, the algorithm was applied on the simulated dataset using a number of iterations T and a population size N_P of 500 in order to ensure convergence to sub-optimal solutions for searching a number

of discriminant features *L* ranging from 1 to 10. Furthermore, to study the ability of our algorithm to repeatedly detect the discriminant wavenumbers, 10 independent runs were performed for each value of *L*.

The results of this experiment are summarized on Figure 5(b) depicting the percentage of selection of the wavenumbers by our GA over the 10 runs in function of the number *L* of searched discriminant features. From a visual analysis of this figure, the algorithm behavior is dependent on the number of searched discriminant features leading to three main results. For a number of genes *L* varying from 1 to 4, the proposed GA is able to sequentially estimate the four simulated discriminant features with 100% of selection over the 10 runs, represented by black vertical line segments on Figure 5(b). However, the order in which the discriminant wavenumbers are estimated (firstly 1118 cm^{-1}, secondly 1682 cm^{-1}, thirdly 1560 cm^{-1}, and fourthly 1056 cm^{-1}) clearly differs from the simulated model description (firstly 1056 cm^{-1}, secondly 1118 cm^{-1}, thirdly 1560 cm^{-1}, and fourthly 1682 cm^{-1}) as described in section 3.1 and visible on Figure 2(c). The GA using the KMeans objective function is thus unable to sequentially select the wavenumbers in function of their discriminatory degree.

For a number of genes varying from 5 to 7, the GA completes the list of the four real discriminant features by sequentially selecting the uninformative wavenumbers 1062 cm^{-1}, 1510 cm-^{1}, 984 cm^{-1}. Their selection percentage being 100%, these wavenumbers are wrongly considered as highly discriminant by the algorithm. They will thus be named “false informative” in the remaining of the article.

For a higher number of genes, the four real discriminant wavenumbers are still correctly identified with 100% of selection. The remaining wavenumbers are estimated with low selection frequencies represented by gray rectangles on Figure 5(b), indicating their irrelevancy as discriminant features. This can be explained by the fact that, at each run, the algorithm is obliged to estimate as many wavenumbers as fixed by the number of genes *L*. The four real

discriminant wavenumbers inducing a high decrease of the fitness function score are always identified by the GA. On the contrary, for the remaining wavenumbers being not discriminant, their selection induces very small variations of the fitness function. The number of iterations being limited to 500, the GA converges to different solutions i.e. to different estimated remaining wavenumbers at each run, inducing the gray rectangles observed on Figure 5(b). Furthermore, for a fixed number of iterations T, the higher the number of searched discriminant features L, the lower the selection percentage of these remaining non-discriminant wavenumbers since the problem complexity increases in function of the number of genes, requiring a higher number of iterations to reach the optimal solution.

To summarize these results, the true discriminant wavenumbers can easily be visualized on Figure 5(b) by a long vertical black line segment corresponding to early and repeatedly selected wavenumbers.

To prove the efficiency of our feature selection method, its impact on the clustering performance must be analyzed. Whatever the number of searched features $1 \leq L \leq 10$, a KMeans clustering applied on the data reduced to the selected wavenumbers always resulted in a 100% sensitivity, 100% specificity and thus 100% accuracy, compared to the low performance obtained when using the entire spectral range as described above. Taken together, these results provide evidence that our proposed feature selection GA is efficient even when the number of genes is reasonably higher than the real number of discriminant features.

4.1.3. Improvement of the proposed feature selection GA by using a validity index as the fitness function

In the previous section using the KMeans objective function as the fitness of the GA, we have shown that our algorithm was not able to select the features in function of their discriminatory degree. This result is coherent since the aim of the KMeans objective function is to find the clusters that minimize the total intracluster distances, without taking into account the separation

between clusters. Validity indices thus appear as relevant mathematical tools to measure the discriminatory power of features in an unsupervised configuration. We thus proposed a new version of our GA by replacing the KMeans objective function by a validity index value (XB and DB) or its inverse value (PBM and ASSWC) as the fitness function as explained in section 2.3.

We first begin the performance analysis of our GA using the inverse of the PBM validity index value as the fitness function. The fitness function being modified, the stability analysis of the proposed GA to select the appropriate parameters ensuring the convergence of the GA has been again performed using the same experimental setup as in the previous section. As visible on Figure 6(a), the general shape of the 3D map is similar to the one obtained with the KMeans objective function used as the fitness function. Once again, the convergence of the algorithm is relatively independent of the population size N_P but strongly dependent of the number of iterations T. However, the convergence of the algorithm, characterized by the flat region on the 3D map, is obtained sooner, i.e. for T higher than 100, whatever N_P.

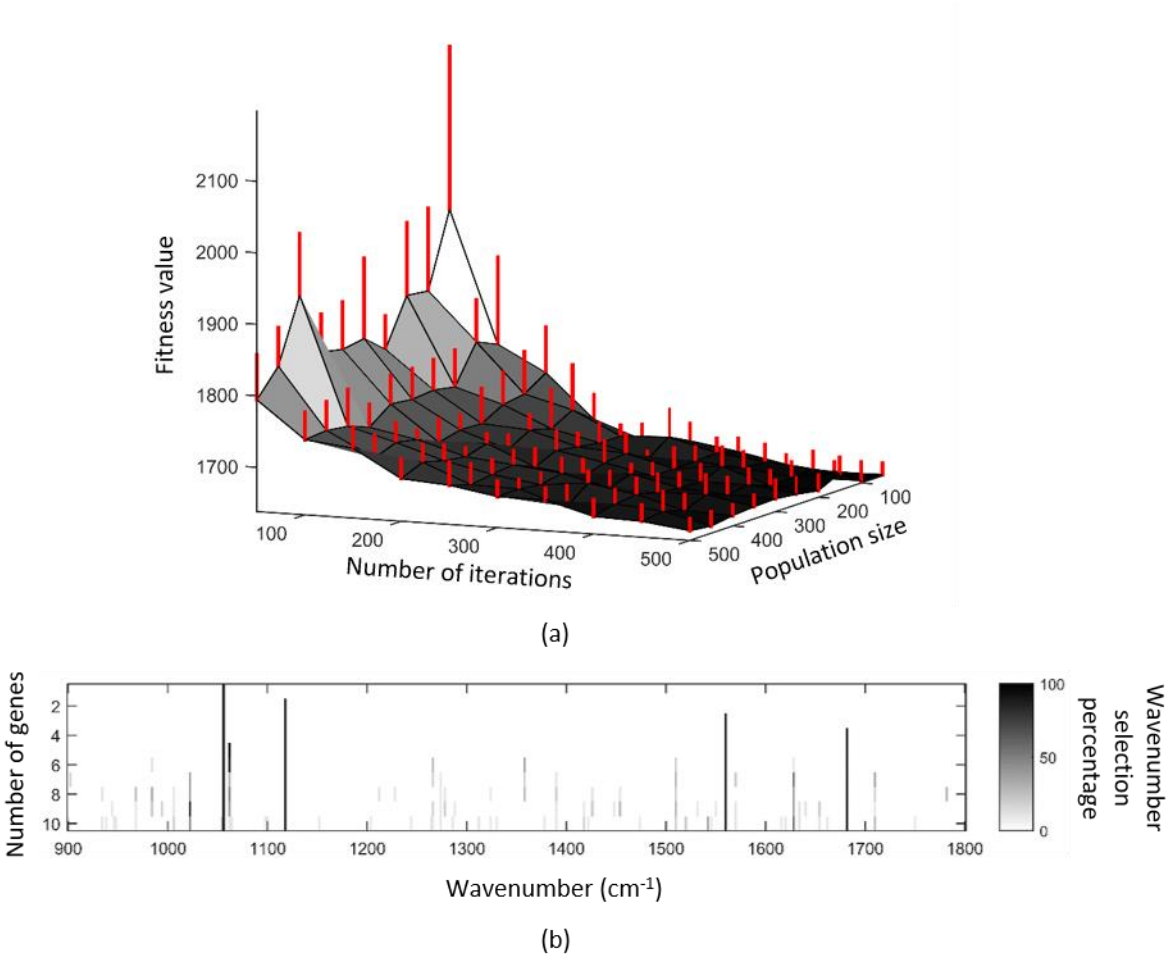


Figure 6. Study of the stability of the proposed GA on the simulated FTIR spectral image using the inverse of the PBM validity index value as the fitness function. (a) Mean fitness function value computed over 10 replicates in function of the population size and the number of iterations in order to find the 10 most discriminant features. Red error bars represent the standard deviation of the fitness function value. (b) Wavenumber selection percentage over 10 replicates in function of the number of searched discriminant features.

The same results were obtained by using XB, DB or the inverse of the ASSWC as the fitness function of our GA (Figure S1(a, c, e)). This faster and reproducible convergence of our GA, whatever the validity index, proves that validity indices are especially well designed to feature selection. Such as in the previous section, our algorithm has been applied in the following using N_P and T equal to 500 in order to ensure convergence to sub-optimal solutions.

The stability of the GA being proved whatever the chosen validity index, the next step consists in studying the ability of the algorithm to select correctly the relevant discriminant wavenumbers in function of the number of searched discriminant features L using the same parameters as in the previous section, i.e. 10 independent runs for each L in {1,…,10}.

From Figure 6(b) depicting the wavenumber selection frequency over the 10 runs in function of L, a first interesting result is that, for a number of genes varying from 1 to 4, our modified GA is able to perfectly and repeatedly estimate the four discriminant features in the same discriminant order defined in the generative model represented on Figure 2(c), i.e. firstly 1056 cm^{-1}, secondly 1118 cm^{-1}, thirdly 1560 cm^{-1}, and fourthly 1682 cm^{-1}. The modified GA is thus certainly able to sequentially select the wavenumbers in function of their discriminatory degree. For a higher number of genes, the four real discriminant variables are still correctly identified. Contrary to the previous section using the KMeans objective function as the fitness function, the remaining selected wavenumbers are no longer false informative, but interference wavenumbers with low selection frequencies.

Concerning the clustering performance on data reduced to the selected wavenumbers, 100% of sensitivity, specificity and accuracy, whatever the number of searched features L in {1,…,10}. The same results were obtained with the other validity indices, as shown in Figure S1(b, d, f).

Taken together, these results lead to several conclusions. First, validity indices as fitness function give better results compared to those obtained with KMeans objective function since, undoubtedly, they can distinguish the relevant discriminant wavenumbers from the non-informative ones, and at the same time correctly sorting them in function of their discriminatory degree. Second, the proposed modified GA is robust to the choice of the number of iterations, the population size and the number of searched discriminant features, facilitating its application and reducing the computation time by avoiding a tedious grid search of these parameters. Third, the developed GA achieves an efficient unsupervised feature selection that greatly reduces the

data dimensionality simplifying further processing and that greatly improves the clustering accuracy.

On the basis of these results, the improved version of the GA, i.e. validity indices as fitness function, will be applied in the remaining of the article.

4.2. Evaluation of the proposed feature selection GA on the DRIFT-MIR public dataset of lyophilized coffees

After having evaluated the performance of our unsupervised feature selection by GA on a simulated dataset, we will now test it on more complex datasets, such as real-world ones. This section presents the results of our GA applied on the labeled DRIFT-MIR coffee spectra described in the "Dataset description" section. The goal is to find the feature subset able to distinguish between the two coffee species, i.e. Arabica and Robusta, in an unsupervised manner, i.e. without using the data labels.

In a first step, it is important to study the convergence and stability of our GA on this real dataset. Thus, the GA has been run using the same experimental setup as in the previous section, except for the number of discriminant features fixed to $L = 100$ in order to take into account the complexity of this real world application.

The results of this stability study are illustrated in Figure 7(a) for the EMSC preprocessed spectra using the inverse of the PBM validity index value as the fitness function. Compared to the results obtained on the simulated dataset, the identification of a flat region on the 3D map is less evident due to a higher variability. This can be explained by several causes: (i) in a real dataset, the frontier between discriminant and non-discriminant features is fuzzier than for the simulated dataset for which the discriminant status of a feature is binary; (ii) as it will be described latter, only 10 features are discriminant for this problem while 100 features are searched in this convergence analysis. The complexity of the problem is thus higher, requiring more iterations to converge to a near-optimal solution. However, as will be seen below and

explained latter in section 3.5, the incomplete convergence of the GA does not prevent the identification of the discriminant features.

In the following, our GA has thus been applied using a number of iterations and a population size of 500 since these values lead to the best convergence (Figure 7(a)). Furthermore, the real number of discriminant features being unknown for this dataset, the number of searched discriminant features L varied in {10, 20, …,100}. For each L, the GA was repeated 10 times to identify the relevant discriminant features by their selection frequency as reported in Figure 7(b).

On Figure 7(b), the most discriminant features selected by our algorithm are easily visible as long vertical black line segments. As a result, it can be remarked that the difference between the two coffee species is mainly due to two spectral bands, i.e. 1660-1670 cm^{-1} and 1690-1715 cm^{-1}. Then, for a number of genes higher than 40, a third discriminant spectral band appears at 1620-1630 cm^{-1}. A less discriminant fourth band appears around 1100 cm^{-1}. These results are in accordance with previous studies [32,33] stating that the main spectral differences between Arabica and Robusta are visible in the spectral range 1550-1750 cm^{-1} attributed to caffeine and around 1100 cm^{-1} associated to chlorogenic acid. Indeed, these two kinds of beans are known to mainly differ in their content of these two compounds. An interesting behavior of our GA using the inverse of PBM as fitness function is that, once selected, these bands remain discriminant whatever the number of genes, as can be seen on Figure 7(b).

This intrinsic selection by our algorithm of adjacent discriminant wavenumbers indicates an existing collinearity or correlation between these features and is in complete adequacy with the nature of these spectral data. Indeed, IR spectral data are well known to be composed of dependent neighbor wavenumbers which are at the origin of the spectral bands associated with molecular vibrations composing an IR spectrum.

Globally, when all the gene numbers tested by our algorithm (i.e. L in {10, 20, …,100}) are considered and by fixing a selection frequency threshold at 90%, our improved GA selected 14 among 451 available features, i.e. 3%, as highly discriminant, mainly constituting the three previously found as highly discriminant bands, i.e. 1620-1630 cm^{-1},1660-1670 cm^{-1} and 1690-1715 cm^{-1}. Furthermore, the remaining variables are selected with relatively low selection percentages characterizing their irrelevancy. The proposed method is thus efficient to select a small subset of discriminant features among a high number of uninformative ones in a real-world environment.

In a second time, it is important to test the effectiveness of the selected features to improve the clustering performance. Thus, for each number of searched discriminant wavenumbers l in L in {10, 20, …,100}, a two-cluster KMeans was applied on the dataset reduced to the selected features. Knowing the ground truth of this dataset, averaged sensitivity and specificity over the 10 GA repetitions were computed as depicted in Figure 7(c). The best accuracy value of 99.14% was obtained using $L = 10$ discriminant wavenumbers. It can be noticed that these good performances are relatively maintained until $L = 40$, then the accuracy decreases as the number of searched features increases. Taken together, these results prove that KMeans clustering with feature selection outperforms full-spectrum KMeans reaching an accuracy of 89.58%. A particular attention must be paid to the number of searched discriminant features since a too high number will surely include a lot of useless features leading to the deterioration of the clustering results. When the GA is repeated, this limit can be circumvented by using a feature selection frequency threshold while conserving high KMeans accuracy, as will be discussed in section 4.5.

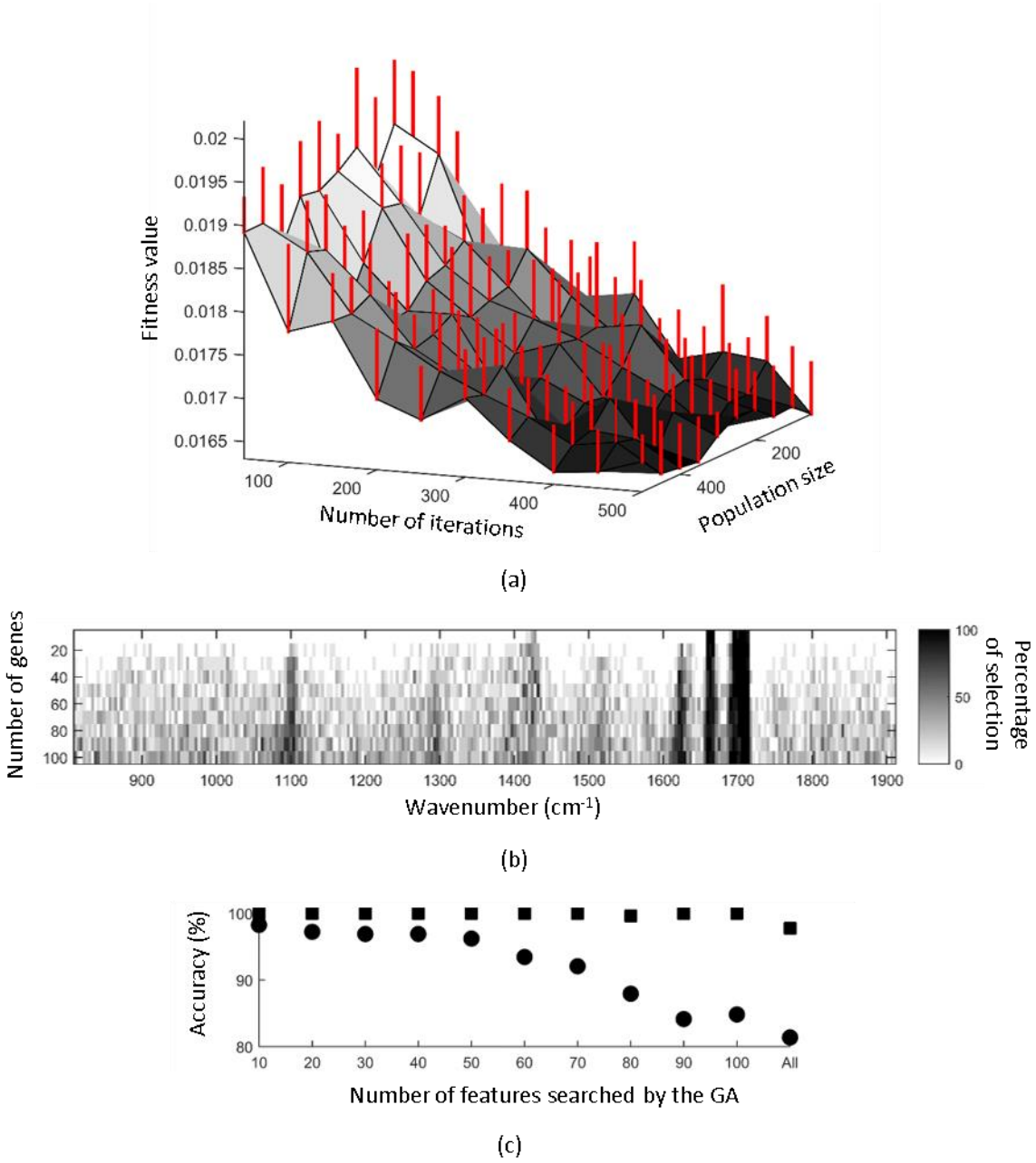


Figure 7. Study of the stability of the proposed GA on the coffee public dataset pre-processed by EMSC using the inverse of the PBM validity index value as the fitness function. (a) Mean fitness function value computed over 10 replicates in function of the population size and the number of iterations in order to find the 100 most discriminant features. Red error bars represent the standard deviation of the fitness function value. (b) Wavenumber selection percentage over 10 replicates in function of the number of searched discriminant features. (c) KMeans

accuracies for Arabica (circles) and Robusta (squares) in function of the number of searched features by the GA. 'All' refers to the use of the full spectral range.

Another important parameter is the validity index used as the fitness function of the GA. To study its influence, the GA has been run for each validity index on the EMSC preprocessed coffee dataset. From the results visible in Figures S2(a, c, e), ASSWC, DB and XB and do not identify the 1620-1630 cm^{-1} and 1660-1670 cm^{-1} bands as discriminant. Furthermore, the 1690-1715 cm^{-1} is correctly identified whatever the number of searched discriminants features only by ASSWC. DB and XB are able to correctly identify this band only for a number of discriminant features not exceeding 50. Moreover, from Figures S2(d, f), the sensitivities and specificities using DB and XB are globally worse than those obtained by PBM in the same conditions. Furthermore, if the number of searched discriminant features is higher than 80, then the feature selection is inefficient since these performance measures are worse than those obtained with the entire spectral range. On the contrary, ASSWC gives very good results (sensitivity of 96.55% and specificity of 100%) whatever the number of searched variables (Figure S2(b)). In addition, KMeans objective function appears totally inappropriate to select discriminant features. The choice of the validity index is thus very important for real world applications. These results demonstrating the superiority of PBM to capture the structure of infrared data are in accordance with previous studies showing PBM as the most efficient validity index for the estimation of the number of clusters on infrared images acquired on normal human and mouse côlon tissues [23].

Taken together, these results show that the choice of the fitness function are very important to ensure good performance of our algorithm. Additionally, the subset of discriminant variables estimated by our algorithm leads to better clustering performance and to a simplification of the biological interpretation of the results by reducing the data complexity to few discriminant spectral bands.

4.3. Validation of the proposed feature selection GA on the ATR-FTIR public dataset of fresh meats

In order to validate the proposed methodology, our GA has been applied on the ATR-FTIR fresh meat spectra, using the following experimental setup: i) a number of iterations N_P and a population size T fixed to 500, ii) the number of searched discriminant features L varying in {10, 20, …, 100}, iii) a number of GA repetitions fixed to 10 to assess the selection frequency of features, iv) a 3-cluster KMeans, v) the inverse of the PBM validity index value as fitness function, vi) an EMSC polynomial order of 2 for ensuring the best KMeans results on the full spectral range.

Figure 8(a) shows that the discriminant features between the three meat categories selected by our algorithm appeared condensed in the spectral band 1530-1560 cm^{-1} specific of protein content. This result was in accordance with previous studies describing different contributions of proteins between these meat types [37]. Furthermore, our algorithm presented a coherent behavior with the spectroscopic data type since the collinearity of the selected features is preserved, without the need of an extra parameter to control it.

Such as in the previous section, when all the gene numbers tested by our algorithm (i.e. L in {10, 20, …,100}) are considered and by fixing a selection frequency threshold at 90%, our algorithm permitted a great dimension reduction since only 14 (3.4%) among the 413 wavenumbers have been identified as highly discriminant as visible on Figure 8(a) as long vertical black line segments in the spectral range 1530-1560 cm^{-1}.

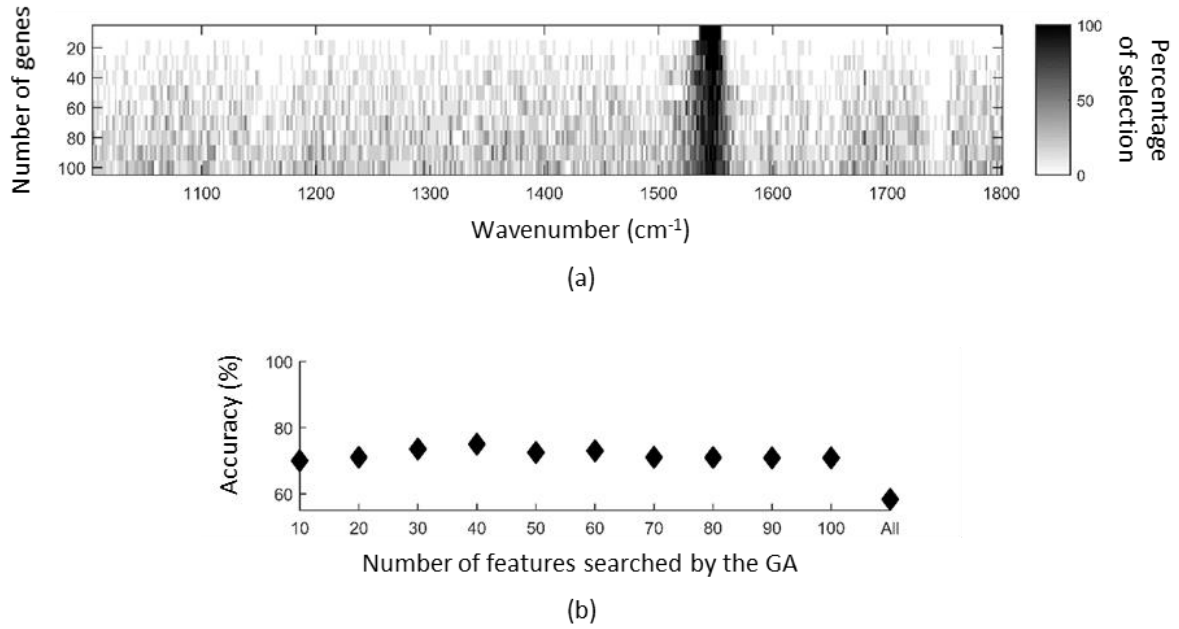


Figure 8. Validation of the proposed GA on the meat public dataset pre-processed by EMSC using a second order polynomial function and using the inverse of the PBM validity index value as the fitness function. (a) Wavenumber selection percentage over 10 replicates in function of the number of searched discriminant features. (b) KMeans accuracy in function of the number of searched features by the GA. 'All' refers to the use of the full spectral range.

To show the efficiency of our GA to improve the clustering performance, a three-cluster KMeans was performed on the dataset reduced to the selected features, by varying the number of searched discriminant wavenumbers L in $\{10, 20, \ldots, 100\}$. The average accuracy over 10 repetitions was computed for each value of L as summarized in Figure 8(b). A first significant result is that KMeans clustering with feature selection outperforms full-spectrum KMeans results, whatever the number of determined discriminant wavenumbers L. Indeed, KMeans clustering with feature selection reaches a minimum of 70% of accuracy for $L = 10$ features and a maximum of 75% obtained for $L = 40$ features, while 58% for full-spectrum KMeans. A second result is the relative stability of the clustering accuracy whatever the number of

searched discriminant wavenumbers *L*, indicating that the number of searched discriminant features does not need to be exactly specified.

All these results validate the efficiency of our unsupervised GA to select relevant discriminant features for improving KMeans clustering performance.

4.4. Validation of the proposed feature selection GA on ATR-FTIR public dataset of fresh and archived fungal spores

For testing the performance of our feature selection GA on a more complicated classification issue, our methodology has been applied on the preprocessed ATR-FTIR spectra of spores in order to distinguish the six different species, using the procedure described in the previous section, except for the KMeans number of clusters fixed to 6. The obtained results are summarized in Figure 9.

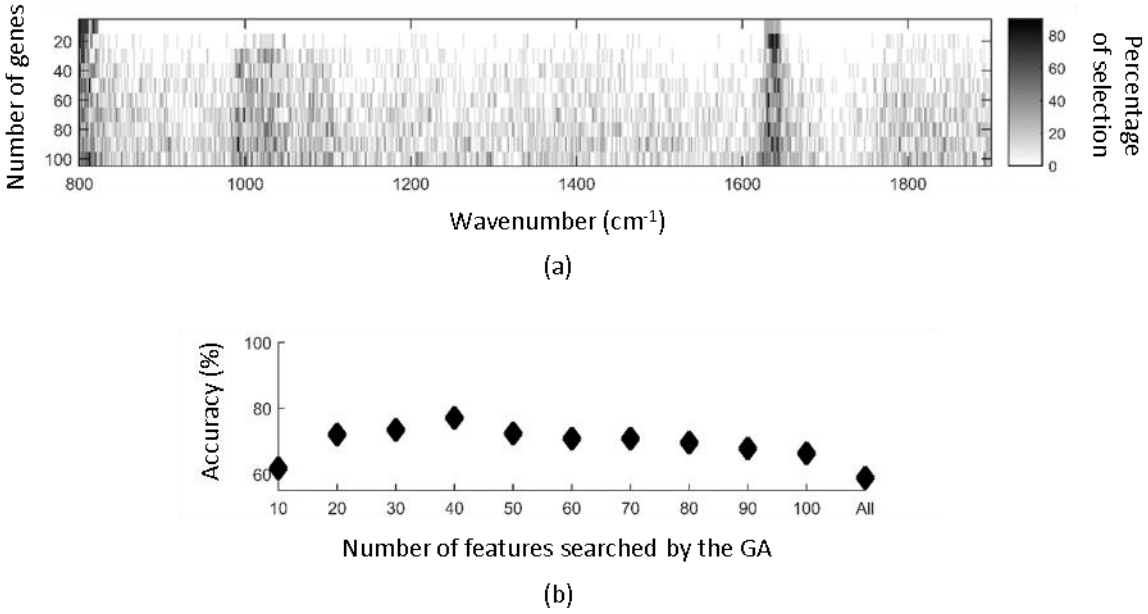


Figure 9. Study of the stability of the proposed GA on the spore public dataset pre-processed by EMSC using a 2-order polynomial function and using the inverse of the PBM validity index value as the fitness function. (a) Wavenumber selection percentage over 10 replicates in

function of the number of searched discriminant features. (b) KMeans accuracy in function of the number of searched features by the GA. 'All' refers to the use of the full spectral range.

As illustrated in Figure 9(a) depicting the selection frequency, our algorithm found adjacent discriminant features belonging to the 800-830 cm^{-1} and 1620-1650 cm^{-1} vibrational bands characteristic of carbohydrate, chitin and proteins which are the main components of spores. A third less visible discriminant spectral band appearing at 980-1080 cm^{-1} can be attributed in part to free fatty acids, glycolipid and phospholipids, indicating spore chemical changes induced by storage conditions [36].

From Figure 9(a), no long vertical black line segments are visible. Indeed, when all the gene numbers tested by our algorithm (i.e. L in {10, 20, …,100}), very few wavenumbers can be defined as highly discriminant, i.e. having a selection frequency over 90%. The selection frequency threshold must be decreased to 80% in order to select 29 among 571 (i.e. 5%) wavenumbers as discriminant. Contrary to the previous datasets, the difficulty for our GA to find highly discriminant features can be explained by the increased data complexity. Indeed, this dataset has been acquired on six different spore species. Furthermore, for each species, the samples were stocked in two distinctive conditions (fresh and archived). And finally, the studied species belong to only three different genera; thus species belonging to the same genus are likely to have a certain level of biochemical similarity. Altogether, these conditions represent various variability sources that can limit the identification of discriminant features.

As shown in Figure 9(b), the KMeans average accuracy computed over 10 repetitions of our algorithm is always better using our feature selection method, with a minimal value of 62% using only $L = 10$ genes, compared to the full-spectrum KMeans reaching 59%. In function of the number of searched features, the average accuracy increases till $L = 40$ to reach its maximum accuracy of 77%, and then monotonically decreases. These results confirm the

efficiency of our algorithm to identify relevant discriminant features and simultaneously to improve KMeans clustering performance in a complex classification issue.

4.5. Simplified and faster procedures for feature selection by GA

From the previous results presented in Figures 6(a), 7(a), 8(a) and 9(a), it appeared that the top discriminant features determined by our GA (represented by black line segments on those figures) are independent of the number of searched features L. It is therefore not necessary to repeat the GA for different number of genes L. In order to exploit this property, we investigated two simplified and faster configurations of our method.

The first configuration, named SFGA1, consisted in running only few times the feature selection GA in order to find a high number of genes L with a high population size N_P and a high number of GA iterations T in order to ensure convergence to sub-optimal solutions. Then, only the features with a percentage of selection above a predefined threshold are selected.

The second configuration, named SFGA2, aimed at further reducing the computation time by applying the GA with small population size N_P and number of iterations T, but still searching for a high number of genes L. However, such small values of N_P and T, and high value of L prevent the algorithm to converge to sub-optimal solutions. By repeating a high number of times the algorithm in this configuration, only the most discriminant features are hypothesized to be repeatedly identified. Then, only the features with a percentage of selection above a predefined threshold are selected.

These two faster procedures were tested on the three real datasets studied previously. The computational times will be also reported later.

For the first configuration SFGA1, the GA was repeated 10 times searching for $L = 100$ genes with a population size $N_P = 500$ and a number of iterations $T = 500$. So, for these datasets, the results of this procedure correspond to the last line of Figures 7(a), 8(a) and 9(a), and to the points with abscissa of 100 on Figures 7(b), 8(b) and 9(b).

Concerning the second configuration SFGA2, the GA was repeated 100 times searching also for $L = 100$ genes but with much smaller population size $N_P = 50$ and number of iterations $T = 50$. The percentage of selection of the wavenumbers are presented for each of the three real datasets on Figure 10. The most frequently selected wavenumbers are very similar to those obtained by the complete procedure or by the SFGA1 configuration, even if their selection percentages are smaller due to the small values of N_P and T.

In order to study the influence of the selection percentage threshold on the results, the KMeans clustering accuracy and the number of selected features were measured for the different possible values of this threshold as depicted on Figures 11 and 12 for SFGA1 and SFGA2, respectively. Whatever the complexity of the dataset, the accuracy is relatively stable in function of the threshold. Too extreme thresholds (close to 0 or 1) must however be avoided. The shape of the curve of the number of selected features can be roughly divided into two parts while the threshold decreases: (i) a constant and slow increase, (ii) then a rapid increase. This sigmoid shape can be interpreted as the ability of our algorithm to clearly distinguish discriminant features from uninformative ones in all the studied datasets.

In order to compare the complete procedure with the two proposed simplified and faster configurations, the best results, i.e. those having the highest KMeans accuracy, are summarized for the three datasets in Tables 2 to 4. Whatever the dataset, the three algorithm configurations give almost identical accuracies. SFGA1 and SFGA2 give their best results for quite similar ranges of feature selection percentage threshold and number of selected features. This last parameter is always higher than the one estimated by the complete procedure, due to the fixed and possibly overestimated value of $L = 100$ searched discriminant features for SFGA1 and SFGA2, and also due to the small population size and number of iterations used in SFGA2. Concerning the computation time, the complete procedure is the greediest, while SFGA1 and SFGA2 save around 86% and 98% of computation time respectively.

The proposed methods being unsupervised, the aim is to use them in real life applications where the data labels are unknown, preventing the computation of KMeans accuracy. In this case, a good choice of the selection percentage threshold can be the value just before the number of selected features begins to increase rapidly. From Figures 11(d-f) resulting from the application of SFGA1, the following threshold could be chosen: (i) 0.7 leading to an accuracy of 94.64% and a number of selected features of 22 with a computational time of 4.51 hours for the lyophilized coffee dataset, (ii) 0.5 leading to an accuracy of 70% and a number of selected features of 37 with a computational time of 1.02 hours for the fresh meat dataset, (iii) 0.4 leading to an accuracy of 75.56% and a number of selected features of 68 with a computational time of 1.09 hours for the fresh and archived fungal spore dataset. From Figures 12(d-f) resulting from the application of SFGA2, the following threshold should be chosen: (i) 0.49 leading to an accuracy of 94.64% and a number of selected features of 21 with a computational time of 0.69 hours for the lyophilized coffee dataset, (ii) 0.35 leading to an accuracy of 70% and a number of selected features of 27 with a computational time of 0.16 hours for the fresh meat dataset, (iii) 0.21 leading to an accuracy of 78.89% and a number of selected features of 129 with a computational time of 0.20 hours for the fresh and archived fungal spore dataset. Even not optimal, these thresholds lead to good solutions which are almost identical for SFGA1 and SFGA2.

Altogether, these results show that the two proposed simplified configurations of our algorithm give results similar to the complete procedure but with a fast computation time. Furthermore, the feature selection percentage can be easily chosen without significantly deteriorate the accuracy. These two configurations should thus be preferred in practical applications where computation time is a limiting factor.

We can reasonably assume that the three presented procedures do not converge to the optimal solutions on the presented real world datasets due to the small values used for the population

size N_P and the number of iterations T compared to the problem complexity. However, finding some of the most discriminant features induces a large decrease of the fitness function. By repeating the procedure several times, the accumulated statistics permit to overcome this convergence problem and to save computational time while still revealing the most discriminant features.

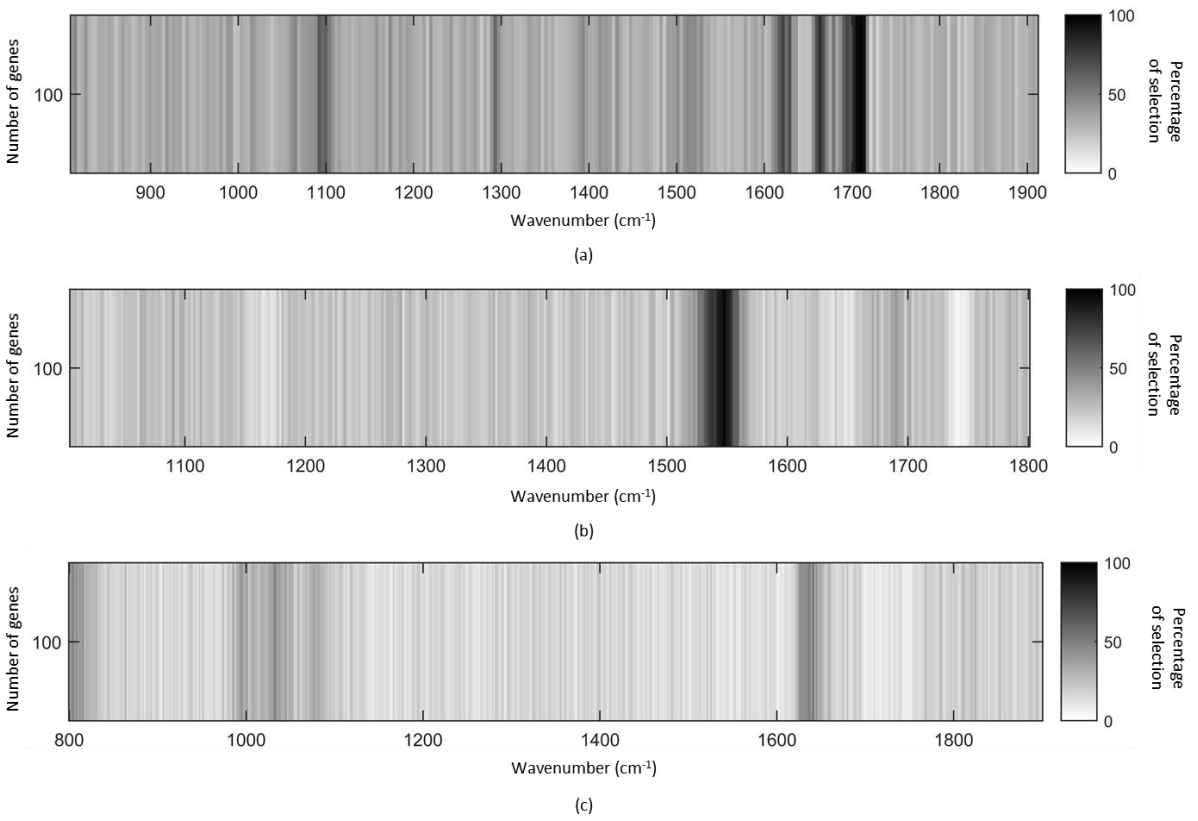


Figure 10. Wavenumber selection percentage estimated by the SFGA2 configuration for the EMSC pre-processed public datasets: (a) DRIFT-MIR spectra of lyophilized coffees, (b) ATR-FTIR spectra of fresh meats, (c) ATR-FTIR spectra of fresh and archived fungal spores.

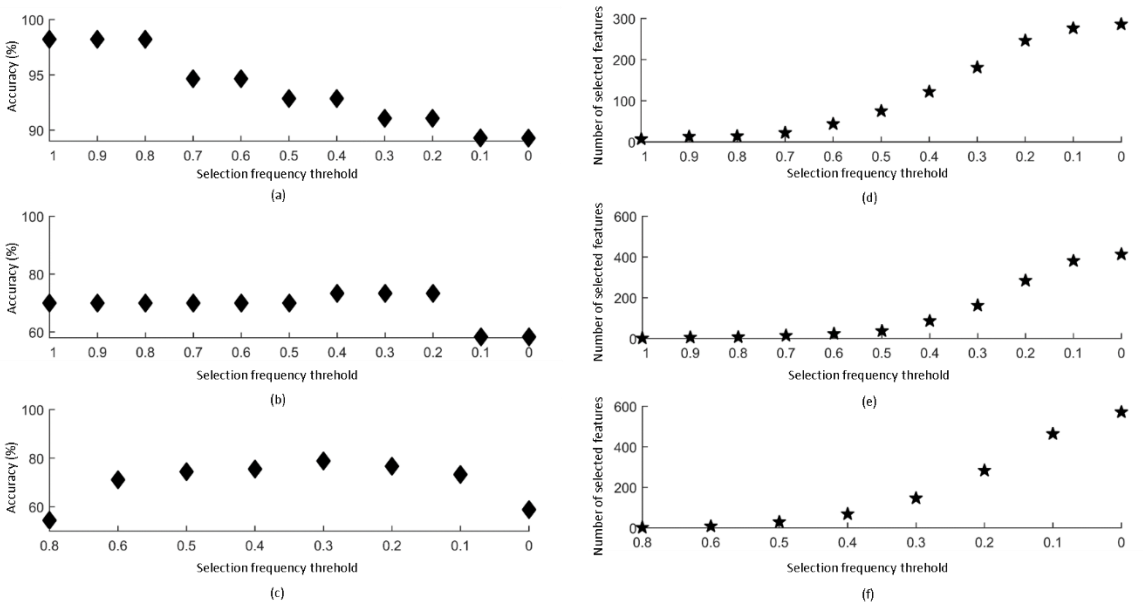


Figure 11. Results of the SFGA1 configuration applied on the real datasets. Evolution of the accuracy of KMeans clustering (a, b, c) and of the number of selected features (d, e, f) in function of the selection frequency threshold for the EMSC pre-processed public datasets: (a, d) DRIFT-MIR spectra of lyophilized coffees, (b, e) ATR-FTIR spectra of fresh meats, (c, f) ATR-FTIR spectra of fresh and archived fungal spores.

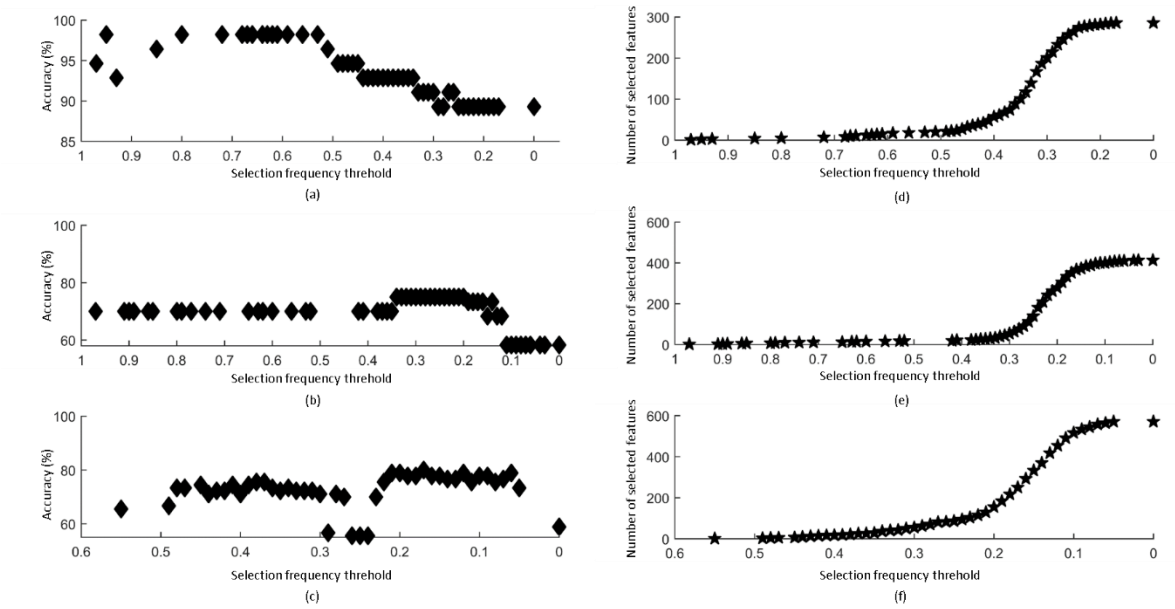


Figure 12. Results of the SFGA2 configuration applied on the real datasets. Evolution of the accuracy of KMeans clustering (a, b, c) and of the number of selected features (d, e, f) in

function of the selection frequency threshold for the EMSC pre-processed public datasets: (a, d) DRIFT-MIR spectra of lyophilized coffees, (b, e) ATR-FTIR spectra of fresh meats, (c, f) ATR-FTIR spectra of fresh and archived fungal spores.

Table 2. Best performance comparison of the different GA configurations applied on the DRIFT-MIR public dataset of lyophilized coffees. *[a ; b] corresponds to values within the range from a to b.

Performance / **Procedure**	**Number of selected features / total number of features**	**Percentage selection threshold**	**KMeans accuracy**	**Computational time (in hours)**
Complete	10 / 451	0%	99.14%	30.58
SFGA1	[7 ; 14]* / 451	≥80%	98.21%	4.51
SFGA2	[5 ; 19]* / 451	[53% ; 80%]*	98.21%	0.69

Table 3. Best performance comparison of the different GA configurations applied on the ATR-FTIR public dataset of fresh meats. *[a ; b] corresponds to values within the range from a to b.

Performance / **Procedure**	**Number of selected features / total number of features**	**Percentage selection threshold**	**KMeans accuracy**	**Computational time (in hours)**
Complete	40 / 413	0%	75.00%	7.26
SFGA1	[86 ; 284]* / 413	[20% ; 40%]*	73.33%	1.02
SFGA2	[30 ; 279]* / 413	[20% ; 34%]*	75.00%	0.16

Table 4. Best performance comparison of the different GA configurations applied on the ATR-FTIR public dataset of fresh and archived fungal spores.

Performance / Procedure	Number of selected features / total number of features	Percentage selection threshold	KMeans accuracy	Computational time (in hours)
Complete	40 / 571	0%	77.00%	8.69
SFGA1	146 / 571	30%	78.89%	1.09
SFGA2	250 / 571	17%	80.00%	0.20

5. Conclusion:

Vibrational spectroscopies probe the biochemical composition of samples by recording light intensities at several hundreds of different wavenumbers. However, only a subset of these numerous wavenumbers is useful for the discrimination between different sample categories. Several feature selection methods have been developed in order to identify these discriminant features in a supervised configuration. However, labels of spectra are not always available. In this case, data analysis is usually non-optimally realized on the full spectral range or on manually selected subsets. In order to overcome this problem, we thus proposed in this article a new unsupervised feature selection method based on a GA measuring the discriminant power of a subset of features by validity indices evaluated on KMeans partitions.

Evaluated on a simulated dataset and on three real spectroscopic datasets, our unsupervised algorithm has been proved efficient to estimate discriminant features. Applied on the features selected by our algorithm, KMeans achieved better performance than when applied on the full spectral range. An interesting behavior of our algorithm is its ability to find discriminant

spectral bands instead of isolated wavenumbers. Presented results also showed the importance of data preprocessing and optimization of their parameters for feature selection and forthcoming clustering. GA being known as time consuming, we also proposed accelerated procedures saving up to 98% of computational time while preserving clustering accuracy.

From an applicative point of view, our method should be applied on vibrational images acquired on few samples in order to identify for example the wavenumbers best revealing a tissue composition. Then, the acquisition of vibrational images on other samples of the same type of tissue should be accelerated by focusing the acquisition on these precise wavenumbers, for example using the Quantum Cascade Laser (QCL) technology.

The objective of this article was to device the general framework of our unsupervised feature selection method. However, several points still need to be in-depth studied. A first perspective is to test other clustering methods instead of KMeans. The framework of our method being simple and flexible, this point could be easily studied. A second perspective is to test other methods for the different GA steps, i.e. chromosome evaluation, fitness scaling, parent selection, parent crossover and mutation. Indeed, the convergence speed of GA definitely depending on these parameters, a proper choice is imperative to insure optimal performance of the algorithm. A third perspective is to favor the selection of a single variable per band in order to avoid to fix the number of searched features to a high value and to favor diversity in the selected biochemical compounds. This improvement could be realized by adding a penalty term to the fitness function.

6. References

[1] Sun D-W. Infrared spectroscopy for food quality analysis and control. Academic Press; 2009. 445 p.

[2] Barth A, Haris PI. Biological and Biomedical Infrared Spectroscopy. IOS Press; 2009. 449 p.

[3] Severcan F, Haris PI. Vibrational Spectroscopy in Diagnosis and Screening. IOS Press; 2012. 433 p.

[4] Brittain HG. Chapter Four - Mid-infrared spectroscopy of pharmaceutical solids. In: Brittain HG, éditeur. Profiles of Drug Substances, Excipients and Related Methodology. Academic Press; 2018. p. 321-58.

[5] Ewing AV, Kazarian SG. Infrared spectroscopy and spectroscopic imaging in forensic science. Analyst. 2017;142(2):257-72.

[6] Mark H, Jr JW. Chemometrics in spectroscopy. Elsevier; 2010. 560 p.

[7] Morais CLM, Lima KMG, Singh M, Martin FL. Tutorial: multivariate classification for vibrational spectroscopy in biological samples. Nat Protoc. 2020;15(7):2143-62.

[8] Donoho DL. High-dimensional data analysis: The curses and blessings of dimensionality. In: Ams Conference on Math Challenges of the 21st Century. 2000.

[9] Sorzano COS, Vargas J, Montano AP. A survey of dimensionality reduction techniques. arXiv. 2014;1403.2877.

[10] Kumar V. Feature selection: A literature review. SmartCR. 2014;4(3):211-29.

[11] Elliott GN, Worgan H, Broadhurst D, Draper J, Scullion J. Soil differentiation using fingerprint Fourier transform infrared spectroscopy, chemometrics and genetic algorithm-based feature selection. Soil Biology and Biochemistry. 2007;39(11):2888-96.

[12] Balabin RM, Smirnov SV. Variable selection in near-infrared spectroscopy: Benchmarking of feature selection methods on biodiesel data. Analytica Chimica Acta. 2011;692(1):63-72.

[13] Xiaobo Z, Jiewen Z, Povey MJW, Holmes M, Hanpin M. Variables selection methods in near-infrared spectroscopy. Analytica Chimica Acta. 2010;667(1):14-32.

[14] Andersen CM, Bro R. Variable selection in regression—a tutorial. Journal of Chemometrics. 2010;24(11-12):728-37.

[15] Goldberg D. Genetic algorithms in search optimization and machine learning. 1988.

[16] Niazi A, Leardi R. Genetic algorithms in chemometrics. Journal of Chemometrics. 2012;26(6):345-51.

[17] Arakawa M, Yamashita Y, Funatsu K. Genetic algorithm-based wavelength selection method for spectral calibration. Journal of Chemometrics. 2011;25(1):10-9.

[18] Leardi R. Application of genetic algorithm–PLS for feature selection in spectral data sets. Journal of Chemometrics. 2000;14(5-6):643-55.

[19] MacQueen J. Some methods for classification and analysis of multivariate observations. In: Proceedings of the fifth Berkeley symposium on mathematical statistics and probability. Oakland, CA, USA; 1967. p. 281-97.

[20] Arbelaitz O, Gurrutxaga I, Muguerza J, Pérez JM, Perona I. An extensive comparative study of cluster validity indices. Pattern Recognition. 2013;46(1):243-56.

[21] Alsamad F, Gobinet C, Vuiblet V, Jaisson S, Piot O. Towards normalization selection of Raman data in the context of protein glycation: application of validity indices to PCA processed spectra. Analyst. 2020;145(8):2945-57.

[22] Boutegrabet W, Guenot D, Bouché O, Boulagnon-Rombi C, Marchal Bressenot A, Piot O, et al. Automatic identification of paraffin pixels on FTIR images acquired on FFPE human samples. Anal Chem. 2021;93(8):3750-61.

[23] Nguyen TNQ, Jeannesson P, Groh A, Guenot D, Gobinet C. Development of a hierarchical double application of crisp cluster validity indices: a proof-of-concept study for automated FTIR spectral histology. Analyst. 2015;140(7):2439-48.

[24] Xie XL, Beni G. A validity measure for fuzzy clustering. IEEE Transactions on Pattern Analysis and Machine Intelligence. 1991;13(8):841-7.

[25] Vendramin L, Campello RJGB, Hruschka ER. On the comparison of relative clustering validity criteria. In: Proceedings of the 2009 SIAM International Conference on Data Mining (SDM). Society for Industrial and Applied Mathematics; 2009. p. 733-44.

[26] Davies DL, Bouldin DW. A cluster separation measure. IEEE Transactions on Pattern Analysis and Machine Intelligence. 1979;PAMI-1(2):224-7.

[27] Pakhira MK, Bandyopadhyay S, Maulik U. Validity index for crisp and fuzzy clusters. Pattern Recognition. 2004;37(3):487-501.

[28] Holland JH. Adaptation in natural and artificial systems: an introductory analysis with applications to biology, control, and artificial intelligence. MIT Press; 1992. 228 p.

[29] Farah I, Nguyen TNQ, Groh A, Guenot D, Jeannesson P, Gobinet C. Development of a memetic clustering algorithm for optimal spectral histology: application to FTIR images of normal human côlon. Analyst. 2016;141(11):3296-304.

[30] Sastry K, Goldberg DE, Kendall G. Genetic algorithms. In: Burke EK, Kendall G, éditeurs. Search methodologies: introductory tutorials in optimization and decision support techniques. Boston, MA: Springer US; 2014. p. 93-117.

[31] Samsudin SH, Shafri HZM, Hamedianfar A, Mansor S. Spectral feature selection and classification of roofing materials using field spectroscopy data. JARS. 2015;9(1):095079.

[32] Briandet R, Kemsley EK, Wilson RH. Discrimination of Arabica and Robusta in instant coffee by Fourier transform infrared spectroscopy and chemometrics. J Agric Food Chem. 1996;44(1):170-4.

[33] Downey G, Briandet R, Wilson RH, Kemsley EK. Near- and mid-Infrared spectroscopies in food authentication: coffee varietal identification. J Agric Food Chem. 1997;45(11):4357-61.

[34] Afseth NK, Kohler A. Extended multiplicative signal correction in vibrational spectroscopy, a tutorial. Chemometrics and Intelligent Laboratory Systems. 2012;117:92-9.

[35] Al-Jowder O, Kemsley EK, Wilson RH. Mid-infrared spectroscopy and authenticity problems in selected meats: a feasibility study. Food Chemistry. 1997;59(2):195-201.

[36] Zimmermann B, Tkalčec Z, Mešić A, Kohler A. Characterizing aeroallergens by infrared spectroscopy of fungal spores and pollen. PLOS ONE. 2015;10(4):e0124240.

[37] Paul AA, Southgate DA, MacCance RA. The composition of foods. HM Stationery Office; 1978.

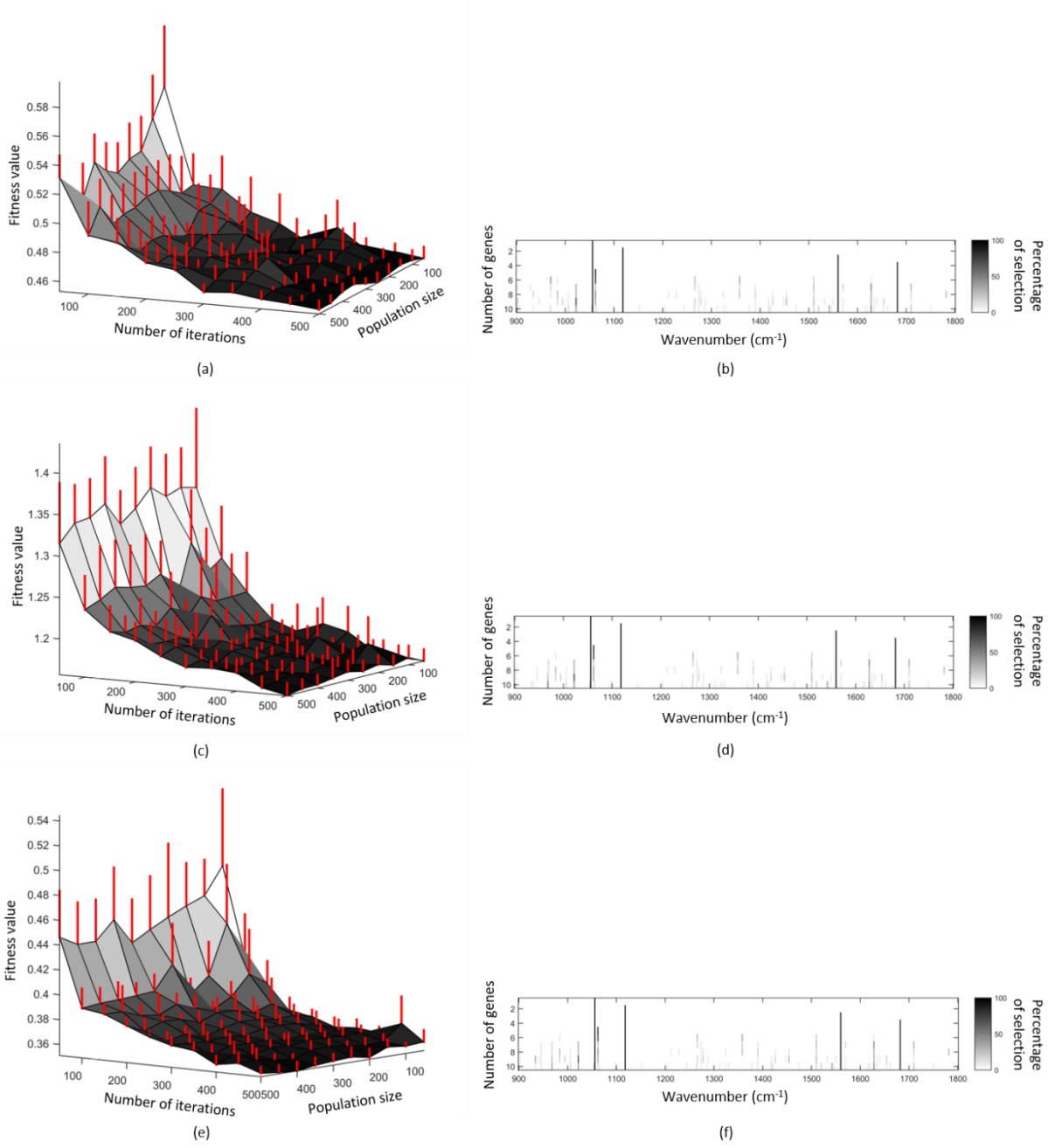


Figure S1. Study of the stability of the proposed GA on the simulated FTIR spectral image using: (a, b) the inverse of the ASSWC validity index value, (c, d) the DB validity index value and (e, f) the XB validity index value as the fitness function. (a, c, e) Mean fitness function value computed over 10 replicates in function of the population size and the number of iterations in order to find the 10 most discriminant features. Red error bars represent the standard deviation of the fitness function value. (b, d, f) Wavenumber selection percentage over 10 replicates in function of the number of searched discriminant features.

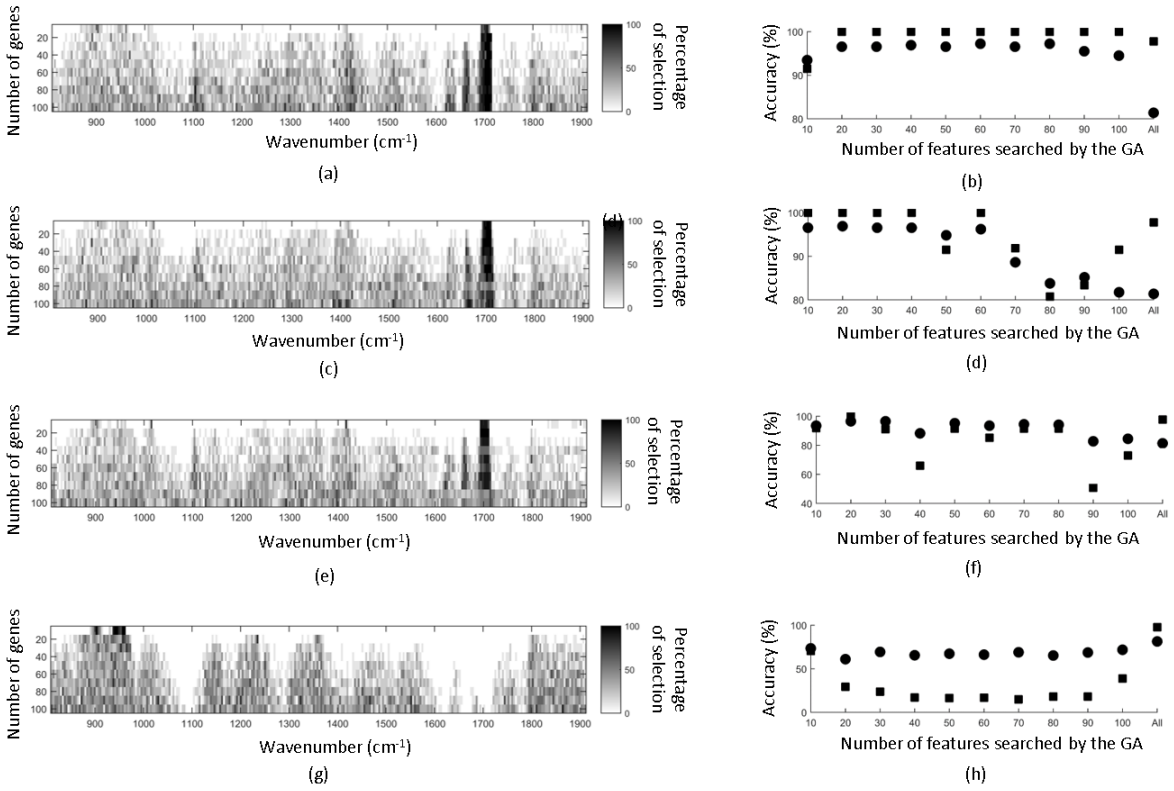


Figure S2. Study of the stability of the proposed GA on the preprocessed coffee public dataset using: (a, b) the inverse of the ASSWC validity index value, (c, d) the DB validity index value, (e, f) the XB validity index value and (g, h) the KMeans objective function as the fitness function. (a, c, e, g) Wavenumber selection percentage over 10 replicates in function of the number of searched discriminant features. (b, d, f, h) KMeans accuracies for Arabica (circles) and Robusta (squares) in function of the number of searched features by the GA. 'All' refers to the use of the full spectral range.

V. C. Résultats supplémentaires

V. C. 1. Application à des images spectrales IR de cancer colique

Nous avons testé l'efficacité de notre approche de sélection non supervisée de variables sur des images FTIR acquises sur des coupes tissulaires paraffinées (FFPE) provenant d'échantillons de patients avec une tumeur colique humaine obtenus auprès du CHU de Reims (service d'oncologie digestive et laboratoire de bio-pathologie). L'objectif était de distinguer les structures histologiques principales de ces échantillons humains. Un exemple est proposé sur la Figure 34 (a) qui représente une coupe colorée à l'H&E de l'un de ces échantillons tumoraux humains. On y discerne les structures histologiques suivantes : les cellules tumorales et le stroma.

Cette image spectrale a été prétraitée en appliquant les prétraitements routiniers décrits dans le chapitre II, à savoir une correction atmosphérique, la transformation des intensités de transmittance en absorbance, et la réduction de la gamme spectrale à l'empreinte digitale 900-1800 cm-1. Puis, elle a été prétraitée par EMSC comme décrit dans le chapitre 2 afin de corriger les spectres de la ligne de base modélisée par une fonction polynomiale d'ordre 4 et de la variabilité du signal de la paraffine, et de normaliser les spectres autour du spectre moyen de cette image. Ensuite, les spectres de la paraffine ont été repérés et éliminés en appliquant la procédure multivariée développée dans le chapitre III.

Notre méthode de sélection non supervisée de variable par AG a ensuite été appliquée sur les spectres tissulaires de cette image FTIR en utilisant l'indice de validité PBM dans la procédure complète, c'est-à-dire pour un nombre de variables recherchées L appartenant à $\{10, 20, \dots, 100\}$, une taille de population N_P et un nombre d'itérations T fixés à 500. De plus, cet échantillon étant composée principalement de deux structures histologiques, $K = 2$ classes ont été recherchées.

L'étude de stabilité de notre AG en termes de variables estimées est présentée sur la Figure 33. Notre AG a sélectionné 12 (3%) nombres d'onde parmi les 451 comme hautement discriminants, c'est-à-dire que leur fréquence de sélection est supérieure à 90 %. Ces nombres d'onde sont adjacents et sont facilement repérables (Figure 33) par de longs segments noirs

verticaux. Ces nombres d'onde, principalement compris dans la gamme spectrale 1600-1650 cm-1, appartiennent à la bande Amide I et sont caractéristiques des protéines qui sont les principaux composants de ces deux structures histologiques (cellules tumorales et mucus).

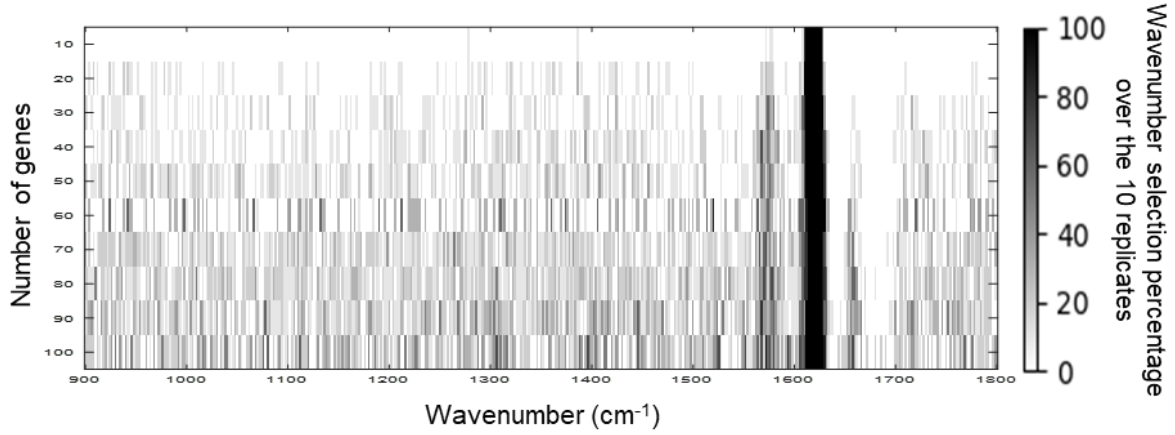


Figure 33. Pourcentage de sélection des nombres d'onde par notre AG répété 10 fois en fonction du nombre de variables discriminantes recherchées en utilisant l'indice de validité PBM.

Appliqué à ces nombres d'onde discriminant, KMeans retrouve les structures histologiques du tissu étudié, comme montré sur la Figure 34(b). Cette partition est identique à celle obtenue en exploitant toute la gamme spectrale 900-1800 cm^{-1}. En effet, contrairement aux données utilisées dans la section précédente pour lesquelles les classes étudiées ne différaient spectralement que dans quelques bandes, les structures histologiques observées ici ont des compositions biomoléculaires assez différentes pour engendrer des différences spectrales (même ténues) sur l'ensemble de la gamme 900-1800 cm^{-1}. Cependant, l'application de notre méthode n'est pas inutile puisqu'elle permet d'identifier les quelques nombres d'onde réellement utiles pour distinguer les classes histologiques, permettant ainsi une meilleure interprétation biologique des résultats. En effet, ces variables identifiées par notre AG sont corrélées à des groupes chimiques fonctionnels spécifiques des structures histologiques de notre échantillon.

Appliquée à un grand nombre d'échantillons, notre méthodologie pourrait donc mener à la définition de marqueurs spectraux spécifiques de la diversité des différentes structures histologiques présentes dans le tissu analysé.

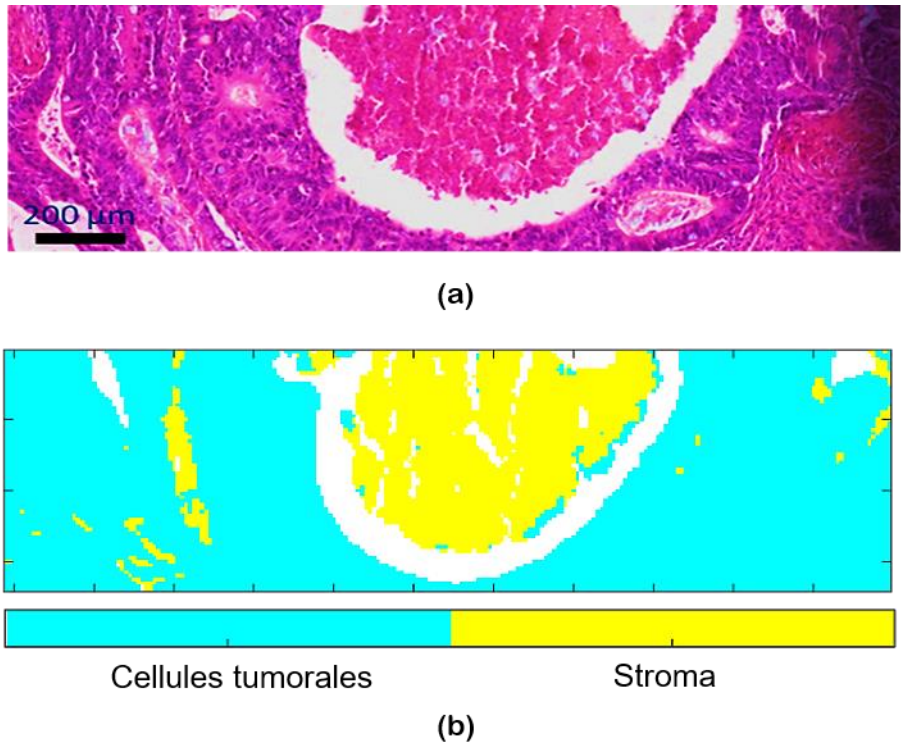


Figure 34. (a) Image de la coupe tissulaire adjacente colorée à l'H&E. (b) Partition KMeans à deux classes après la sélection de variables par notre AG.

V. C. 2. Incorporation d'une contrainte d'homogénéité spatiale

Une propriété importante des échantillons biologiques, très rarement exploitée en histopathologie spectrale IR, est leur structure spatiale. En effet, les pixels d'une image spectrale IR ne sont pas distribués aléatoirement, mais une cohérence spatiale existe entre pixels voisins.

Afin d'exploiter cette propriété fondamentale des images IR, nous avons intégré une mesure d'hétérogénéité spatiale dans la fonction d'évaluation de notre AG. Pour une partition **P** estimée par KMeans sur les données réduites aux L variables composant l'individu **z**, une hétérogénéité locale $H_i(\mathbf{P})$ du clustering est mesurée dans un voisinage 3×3 de chaque pixel i. Notons K_i le nombre de clusters différents présents dans ce voisinage 3×3. Alors l'hétérogénéité locale est définie par :

$$H_i(\mathbf{P}) = K_i - 1 \qquad \text{Éq (18)}$$

Ainsi, si tous les pixels de ce voisinage appartiennent à la même classe, alors il n'y a pas d'hétérogénéité et $H_i(\mathbf{P}) = 0$, donc l'homogénéité est complète. Plus le nombre de classes présentes dans ce voisinage est grand, et plus l'hétérogénéité locale $H_i(\mathbf{P})$ est grande.

L'hétérogénéité moyenne sur l'image complète se calcule donc par :

$$H(\mathbf{P}) = \frac{\sum_{i=1}^{N} H_i(\mathbf{P})}{N} \qquad \text{Éq (19)}$$

où N représente le nombre total de pixels de l'image.

La nouvelle fonction d'évaluation de notre AG est donc définie par :

$$F(\mathbf{z}) = IV(\mathbf{P}) \times H(\mathbf{P}) \qquad \text{Éq (20)}$$

où $IV(\mathbf{P})$ correspond à la valeur de l'indice de validité calculé sur la partition **P** estimée par KMeans sur les données réduites aux L variables composant l'individu **z**.

Nous rappelons que $IV(\mathbf{P})$ s'exprime différemment en fonction de l'indice de validité choisi :

- $IV(\mathbf{P}) = XB(\mathbf{P})$ pour l'indice de validité XB ;
- $IV(\mathbf{P}) = DB(\mathbf{P})$ pour l'indice de validité DB ;
- $IV(\mathbf{P}) = \frac{1}{PMB(\mathbf{P})}$ pour l'indice de validité PBM ;
- $IV(\mathbf{P}) = \frac{1}{ASSWC(\mathbf{P})}$ pour l'indice de validité ASSWC.

Afin de tester notre approche, nous avons simulé une nouvelle image multidimensionnelle à partir de celle simulée et décrite dans l'article à soumettre précédent. Cette nouvelle image mime donc encore une image FTIR acquise sur un tissu de côlon normal composé de deux structures histologiques. Elle est composée de 200 spectres d'une résolution spectrale de 2 cm^{-1}. Chaque spectre contient donc 451 variables représentant les nombres d'onde acquis dans la gamme spectrale 900 à 1800 cm^{-1}. Chaque structure histologique simulée est donc représentée par 100 spectres simulés différents dont les distributions spatiales sont visibles sur la Figure 35(a). Parmi les 451 nombres d'onde, les même quatre nombres d'onde $\mathbf{z_s} = \{1056, 1118, 1560, 1682\}$ cm^{-1} sont discriminants des deux structures histologiques simulées. Par contre, ce jeu de données incorpore quatre nouveaux nombres d'onde $\mathbf{z_p} = \{950, 1200, 1300, 1500\}$ cm^{-1} choisis pour être des nombres d'onde parasites qui induisent une partition KMeans optimale en deux classes spatialement très hétérogènes visibles sur la Figure 35(b).

La moyenne de ces spectres ainsi simulés et leur enveloppe à un écart-type sont visibles sur la Figure 36. Afin de quantifier la puissance des variables à discriminer les groupes parasites ou bien les structures histologiques simulées, nous avons défini un degré de discrimination δ défini en chaque nombre d'onde comme la valeur absolue de la différence entre les spectres moyens soit des deux groupes parasites, soit des deux structures histologiques simulées. Comme montré

sur la Figure 37(a-b), un degré de discrimination δ identique aux nombres d'onde spatialement discriminants a été accordé aux nombres d'onde parasites.

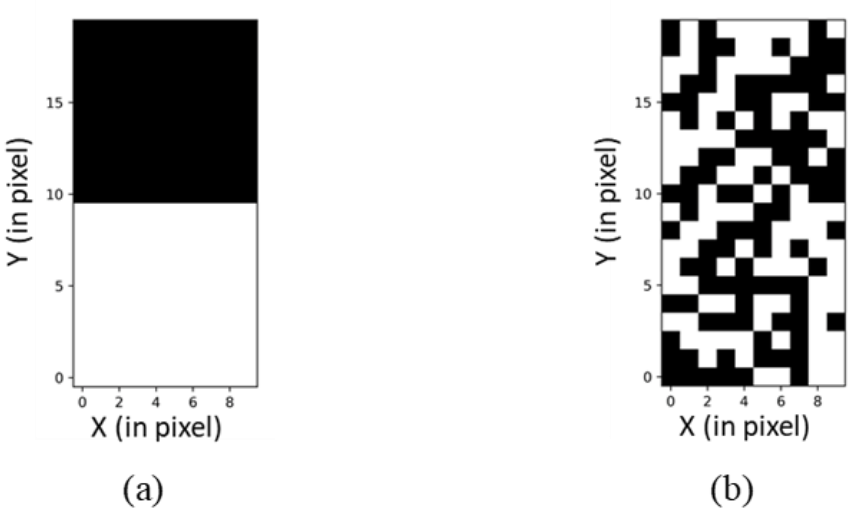


Figure 35. (a) Distribution spatiale des deux structures histologiques simulées représentées respectivement par des pixels blancs et noirs, (b) Distribution spatiale des deux groupes parasites simulés représentés respectivement par des pixels blancs et noirs.

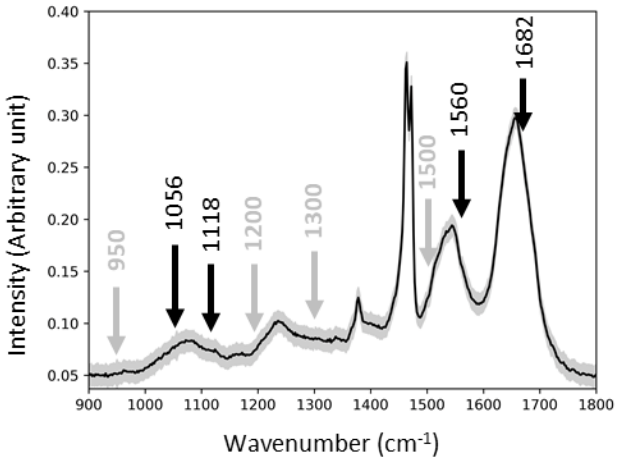


Figure 36. Spectre moyen (ligne noire continue) et son enveloppe à un écart-type (zone ombrée) du jeu de données simulées. Les flèches noires et grises identifient respectivement les quatre nombres d'onde spatialement discriminants et les quatre nombres d'onde parasites.

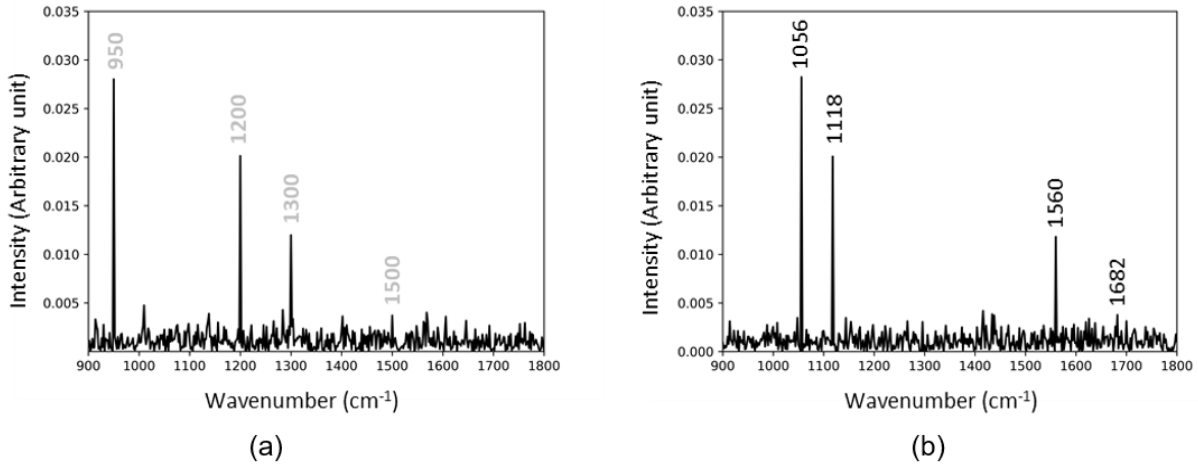


Figure 37. (a) Degré de discrimination δ des nombres d'onde calculé à partir des deux classes parasites. Les quatre nombres d'onde parasites sont identifiés par leurs nombres d'onde. (b) Degré de discrimination δ des nombres d'onde calculé à partir deux structure tissulaires simulées. Les quatre nombres d'onde discriminants sont identifiés par leurs nombres d'onde.

Un KMeans à deux classes a été appliqué sur la gamme spectrale complète (900-1800cm-1) de l'image simulée. La partition estimée est visible sur la Figure 38(a). Cette partition est spatialement distribuée aléatoirement et est complètement décorrélée de la distribution spatiale des deux structures histologiques simulées visible sur la Figure 35(a) et de la distribution spatiale parasite visible sur la Figure 35(b). De plus, la différence entre les deux centroïdes estimés ne révèle aucune différence significative entre les deux clusters, en particulier aux nombres d'onde discriminants et parasites, comme illustré sur la Figure 29(a-b). Cela confirme la nécessité de procéder à une sélection de variables afin d'améliorer les performances du clustering par KMeans.

Afin d'évaluer les performances de notre nouvelle architecture d'AG, les deux versions (avec indice de validité et avec ajout d'une contrainte d'homogénéité spatiale) ont été comparées sur ce jeu de données simulées.

Tout d'abord, la convergence et la stabilité des deux versions ont été étudiées en termes de valeur de la fonction d'évaluation en appliquant chaque version 10 fois pour chaque couple de valeurs $(n_P, t) \in \{50, 100, \dots, 500\}^2$ où n_P et t représentent respectivement la taille de la population et le nombre d'itérations de l'AG. Comme visible sur la Figure 39, les deux versions de notre algorithme présentent des cartographies 3D de convergences similaires et stables. La convergence de l'algorithme est relativement indépendante de la taille de la population mais fortement dépendante du nombre d'itérations. Cependant, la convergence de l'algorithme vers

la solution sous-optimale (région plate) est obtenue pour une taille de population supérieure à 300.

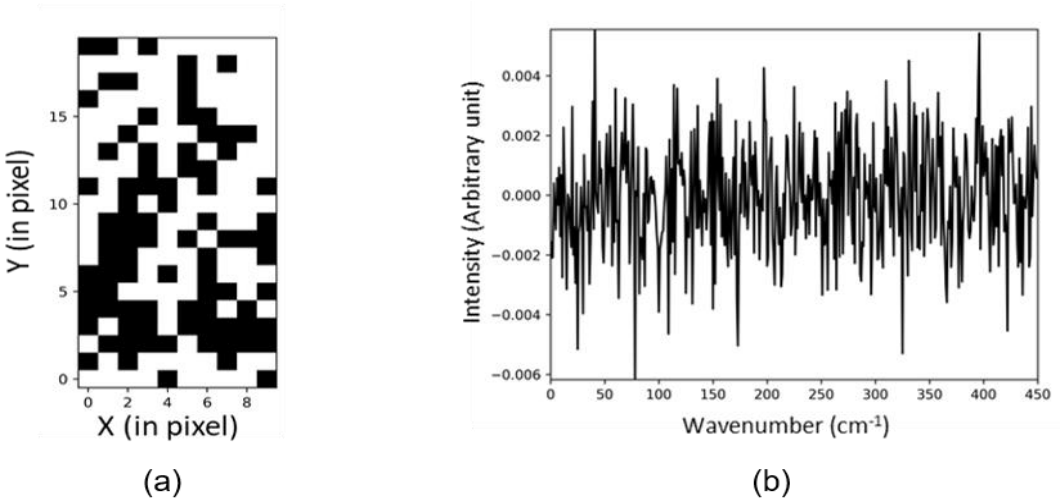


Figure 38. Résultats d'un KMeans à deux clusters, appliqué sur les données simulées complètes, c'est-à-dire en utilisant la gamme spectrale entière (900-1800 cm^{-1}). (a) Partition estimée. (b) Spectre de différence entre les deux centroïdes.

Nous avons ensuite vérifié la capacité de la nouvelle version de notre AG à identifier les nombres d'ondes discriminants des deux structures histologiques simulées. Dans ce but, les deux versions de notre AG ont été appliquées sur l'image FTIR simulée avec les paramètres suivants : i) un nombre d'itérations et une taille de population fixés à 500, ii) le nombre de variables discriminantes recherchées variant dans {1, 2, …, 10}, iii) un nombre de répétitions de l'AG fixé à 10 pour estimer une fréquence de sélection des nombres d'onde, iv) un KMeans à 2 clusters, v) XB comme indice de validité utilisé pour calculer la fonction d'évaluation dans les deux versions de l'AG. Les résultats obtenus sont illustrés sur la Figure 40.

Les résultats obtenus sans critère d'homogénéité spatiale montrent que l'AG est incapable de sélectionner uniquement les nombres d'ondes discriminants des structures histologiques simulées comme présenté sur la Figure 40(a). Les premiers nombres d'onde estimés sont les nombres d'onde parasites. Puis au-delà de 4 variables recherchées, les nombres d'ondes spécifiques des structures histologiques simulées s'ajoutent à la sélection. Pour 4 variables, KMeans atteint une précision autour de 50%.

La nouvelle fonction fitness qui combine à la fois l'indice de validité et impose une contrainte de cohérence spatiale s'est avérée très efficace pour la sélection des variables les plus

discriminantes des structures histologiques simulées (Figure 40(b)). Avec cette sélection de variables, KMeans atteint une précision de 100%.

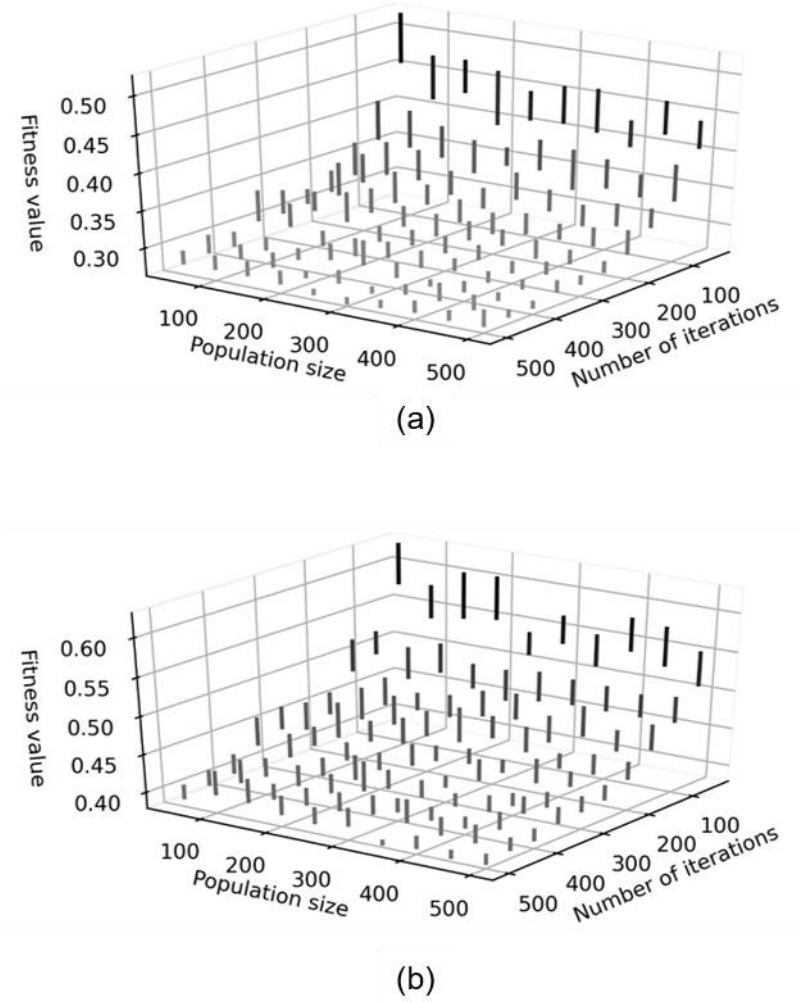


Figure 39. Etude de la stabilité en termes de valeur de la fonction d'évaluation des deux versions proposées de sélection non-supervisée de variables appliquées sur l'image spectrale FTIR simulée en fonction de la taille de la population et du nombre d'itérations. Sur chaque figure sont représentées des barres symbolisant les valeurs moyennes et les écarts-types de la fonction d'évaluation calculée pour 10 répétitions des AG. (a) Fonction d'évaluation égale à l'indice de validité XB. (b) Fonction d'évaluation définie par l'équation (Éq (21)) en utilisant l'indice de validité XB.

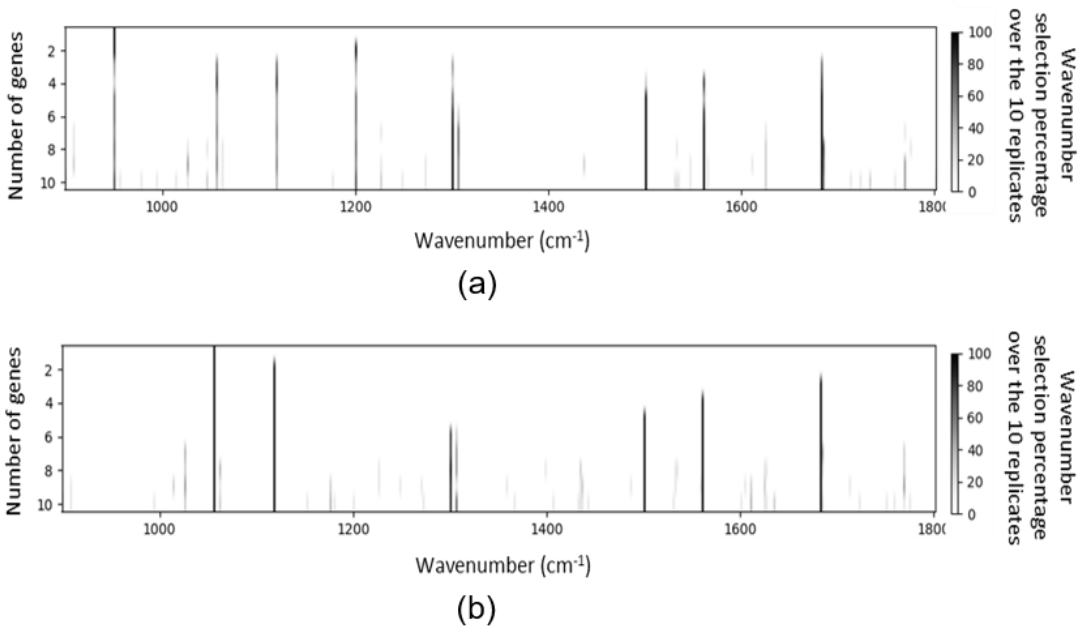


Figure 40. Etude de la stabilité en terme de sélection des nombres d'onde des deux versions proposées de l'AG appliquées sur l'image spectrale FTIR simulée en fonction du nombre de variables discriminantes recherchées. Pour chaque version et chaque nombre de variables recherchées, l'AG est appliqué 10 fois. (a) Fonction d'évaluation égale à l'indice de validité XB. (b) Fonction d'évaluation définie par l'équation (Éq (22)) en utilisant l'indice de validité XB.

L'ensemble de ces résultats préliminaires montre l'intérêt d'incorporer une contrainte d'homogénéité spatiale dans la fonction d'évaluation de notre AG. En effet, cette contrainte permet une sélection efficace des variables spatialement discriminantes même en présence de variables parasites discriminantes par chance.

V. D. Perspectives

D'un point de vue applicatif, la première version de notre AG doit être validée sur d'autres images spectrales IR acquises sur des échantillons biologiques réels afin d'identifier les nombres d'onde révélant le mieux leurs compositions tissulaires. Pour tirer complètement parti de notre approche, il faudrait la confronter à des problématiques pour lesquelles seules quelques bandes spectrales sont corrélées au problème posé. Par exemple, notre approche pourrait être pertinente dans l'étude de la réponse de patients cancéreux à des traitements. En effet, on peut supposer que si des différences spectrales existent entre des échantillons de patients bon répondeurs et mauvais répondeurs, alors ces différences doivent être très ténues et spécifiques à quelques différences biomoléculaires.

La deuxième version de notre AG n'a pas encore été validée sur des données réelles. Pour le moment, son application sur des images spectrales IR d'échantillons biologiques n'améliore pas les résultats d'un KMeans appliqué sur toute la gamme spectrale ou de la première version de notre AG. En effet, les échantillons analysés durant cette thèse étaient constitués de structures tissulaires ayant des signatures spectrales IR différentes sur l'ensemble de la gamme spectrale analysée 900-1800 cm-1. Ces différences se suffisent à elle-même pour retrouver les structures spatiales des échantillons. Nous pensons que notre algorithme pourrait avoir un intérêt sur des images vibrationnelles bruitées. C'est pourquoi nous envisageons de l'appliquer sur des images Raman acquises avec des faibles temps d'accumulation sur des échantillons biologiques.

Ces méthodes de sélection non supervisée de variable pourraient indirectement permettre d'accélérer l'acquisition d'images spectrales sur des grosses cohortes d'échantillon. En effet, les nombres d'onde discriminants estimés à partir de quelques échantillons pourraient orienter l'acquisition rapide d'images spectrales IR à l'aide de la technologie QCL (Quantum Cascade Laser). On peut imaginer le même genre de scénario en imagerie Raman pour identifier des nombres d'ondes qui seront ensuite utilisés pour la prise d'images Raman rapide par la technologie CARS (Coherent Anti-stokes Raman Scattering) ou la technologie SRS (Stimulated Raman Scattering).

Une autre perspective serait de tester d'autres méthodes pour les différentes étapes de l'AG. En effet, pour chaque étape de notre AG, nous avons choisi une méthode parmi les nombreuses variantes existantes. Or, la vitesse de convergence d'un AG étant dépendante de ces paramètres, un bon choix est impératif pour assurer des performances optimales de l'algorithme.

Une dernière perspective est de privilégier la sélection d'une seule variable par bande afin de favoriser la diversité des composés biochimiques sélectionnés et d'éviter de fixer le nombre de variables recherchées à une valeur élevée. Cette amélioration pourrait être réalisée en ajoutant un terme de pénalité à la fonction d'évaluation.

Chapitre VI : Conclusion et perspectives

VI. A. Conclusion

Au cours de ce travail de thèse, pour répondre au besoin majeur d'optimisation et d'automatisation de l'histopathologie spectrale IR, différents développements méthodologiques ont été réalisés sur des échantillons de tumeurs coliques humaines.

Le premier enjeu était de s'affranchir des limites liées aux conditions de préparation des échantillons biologiques, plus particulièrement lorsque les échantillons sont inclus en paraffine. Une correction optimale de la contribution spectrale de la paraffine est nécessaire pour pouvoir interpréter correctement les variations chimiques détectées au sein des tissus par imagerie IR. La correction EMSC (Extended Multiplicative Signal Correction) a été initialement développée comme méthode de prétraitement des images spectrales de tissus paraffinés. Elle permet de normaliser les spectres, corriger des effets de la ligne de base, neutraliser la variabilité du signal de la paraffine et éliminer les spectres aberrants « non tissulaires » ou « outliers » ; cette approche permet d'éviter un déparaffinage chimique des échantillons avant leur analyse. Dans ce travail, nous avons mis au point une méthode automatique de détection des spectres non tissulaires. Cette méthode innovante repose sur une analyse multivariée des coefficients estimés par l'EMSC. Cette méthodologie consiste à appliquer une classification non-supervisée de type KMeans à deux classes, sur toutes les combinaisons possibles de coefficients de régression d'EMSC. Quatre indices de validité (Xie-Beni (XB), Davies-Bouldin (DB), Pakhira–Bandyopadhyay–Maulik (PBM) et Silhouette-Width-Criterion (SWC)) sont ensuite appliqués sur cette partition pour en estimer la qualité. La combinaison de coefficients estimés par l'EMSC optimisant les valeurs de ces indices est retenue de manière objective. Notre méthodologie a été testée sur des données simulées où les différentes sources de variabilité parasite existant dans des spectres infrarouges (paraffine, ligne de base, bruit) sont contrôlées, puis validée sur des images spectrales acquises sur des coupes fines de tumeurs coliques humaines, issues de blocs fixés au paraformaldéhyde puis inclus en paraffine (FFPE). Nous avons utilisé des outils statistiques de visualisation de données multidimensionnelles, tels que t-SNE, qui sont encore très peu utilisés pour l'analyse des spectres vibrationnels. La validation de notre approche, quant à elle, a été réalisée au moyen de l'indice de Jaccard calculé entre une image traitée et une image de référence ; plus précisément entre la partition KMeans à deux classes estimée à partir de la combinaison optimale des coefficients de régression de l'EMSC

et la partition KMeans à deux classes estimée sur l'image spectrale IR de la même coupe tissulaire mais après déparaffinage chimique. Pour les données simulées, l'image de référence est accessible à partir de la contribution de la composante tissulaire utilisée pour la construction de ces données. En effet, un indice de Jaccard > 0.90 sur les données simulées, et > 0,84 sur les données réelles a permis de confirmer la validité de notre approche. Quel que soit le type de données (simulées ou réelles), la meilleure combinaison de coefficients de régression de l'EMSC pour éliminer les pixels non tissulaires s'avère prendre en compte les coefficients de régression du spectre de référence et du spectre moyen du milieu d'inclusion, ainsi que la constante du polynôme utilisé pour modéliser la ligne de base.

L'application de cette nouvelle méthode totalement automatisée permet donc de sélectionner parfaitement les pixels spécifiques des structures histologiques au sein des images spectrales, évitant ainsi d'éventuels biais dus aux pixels du milieu d'inclusion lors de traitements ultérieurs par classification.

Afin de tester la robustesse de notre approche à un autre type de milieu d'inclusion, nous avons appliqué la même méthodologie à des coupes de côlon normal et tumoral, issues de blocs non fixés mais congelés dans de la colle histologique de type OCT. En effet, la signature spectrale de la colle histologique OCT est complexe avec des bandes de vibrations se superposant aux caractéristiques du tissu. La même tendance a été observée en ce qui concerne les meilleures combinaisons des coefficients d'EMSC permettant d'identifier automatiquement les pixels non tissulaires (OCT et support en CaF_2). Cette mise au point a permis d'identifier avec efficacité les pixels d'OCT-CaF_2 sur ces images spectrales IR de tissu colique.

Après ces développements visant à optimiser le prétraitement en identifiant automatiquement les pixels aberrants non tissulaires, nous avons dans un deuxième temps développé un algorithme génétique pour une sélection non-supervisée de variables (nombres d'ondes) dans le but d'identifier des marqueurs spectraux associés à une classification. L'importance de cette sélection de variables a été soulignée dans la littérature, en particulier pour améliorer l'efficacité de modèles de prédiction. Cependant, dans la grande majorité des études, cette étape de sélection de variables est réalisée de manière supervisée, ce qui suppose de connaître parfaitement le label des échantillons tissulaires et leur composition. Or, cette composition est généralement inconnue ou repose sur une connaissance qui ne se situe pas au même niveau d'informations que l'information biochimique liée à un spectre vibrationnel. Les étapes de sélection de variables et de construction d'un modèle de prédiction sont donc tributaires de cette

partition estimée de façon supervisée, pas forcément fiable. Pour pallier ce problème, nous avons développé une nouvelle méthode de sélection non supervisée de nombres d'ondes, basée sur la combinaison d'un algorithme génétique (AG) et d'un clustering par KMeans dont la qualité de partition est évaluée au moyen d'indices de validité.

Dans ce travail, notre méthodologie de sélection de variables a été évaluée sur des bases de données simulées mimant une image FTIR acquise sur un tissu de côlon normal puis validée sur des bases de données réelles publiques. La méthode a également été testée sur des données spectrales IR de coupes de tissus fixés de cancer colique. Nous avons pu montrer la stabilité de l'algorithme et sa capacité à estimer, de manière non-supervisée, un sous-ensemble de descripteurs spectroscopiques aboutissant ainsi à des partitions optimales des données.

VI. B. Perspectives

Les perspectives relatives à ce travail de recherche concernent deux volets : un volet méthodologique d'ordre chimiométrique et un volet applicatif d'ordre biologique.

VI. B. 1. Sur le plan chimiométrique

Après application des méthodes de prétraitement et sélection de variables que nous avons développées, l'implémentation d'algorithmes de classification viendra compléter la séquence chimiométrique d'analyse des tissus. Ces méthodes de classification supervisées consistent à construire des modèles de prédiction à partir de données labellisées selon des références, puis de les tester sur des jeux de données indépendantes. Il nous semble plus particulièrement intéressant d'adapter des méthodes de régressions à réponses multiples (de type Partial Least Squares multi blocs ou encore les algorithmes de deep learning) car elles présentent l'avantage de pouvoir prendre en compte plusieurs modalités de données et plusieurs références simultanément (161). Ces algorithmes permettent aussi d'accéder aux signaux spectraux impliqués dans les corrélations entre les signatures infrarouges et les informations biologiques/cliniques de référence. De plus, la combinaison de plusieurs classifieurs, plutôt que d'utiliser un seul classifieur, est une voie intéressante à explorer pour optimiser les performances diagnostiques de l'approche vibrationnelle (162,163).

L'histopathologie spectrale pourrait aussi gagner en robustesse en combinant les images des deux modalités vibrationnelles, absorption IR et diffusion Raman. L'approche la plus simple à explorer sera d'appliquer l'histopathologie spectrale automatisée séparément pour chaque modalité. Les clusters intermodalités les plus corrélés spatialement seront identifiés et les pixels communs seront alors étiquetés à une structure histologique précise avec l'aide d'un partenaire anatomopathologiste. Les pixels restants, n'appartenant pas aux deux clusters (IR et Raman) simultanément, seront écartés de l'étude. A l'issue de cette procédure, des classes homogènes de spectres multimodaux associées à des structures tissulaires précises seront construites. Cette histopathologie spectrale multimodale permettra de définir des marqueurs spectraux plus précis en termes d'informations biochimiques, du fait de la complémentarité des techniques Raman et infrarouge au niveau des modes de vibration sondées.

De plus, l'identification de nombres d'ondes d'intérêt et discriminants entre diverses structures histologiques offre la possibilité d'utiliser les nouvelles techniques d'imagerie Raman cohérente ou infrarouge par source laser QCL (Quantum Cascade Laser) qui ciblent des fréquences de vibration précises (164,165). Ces techniques présentent une sensibilité de détection significativement plus élevée que l'imagerie infrarouge par Transformée de Fourier ou que l'imagerie Raman spontanée.

VI. B. 2. Sur le plan biologique

Nos développements visent à améliorer l'histopathologie spectrale en la rendant plus robuste, notamment en étant indépendante de la subjectivité de l'opérateur. L'intérêt de ces travaux devra être démontré dans des applications translationnelles, et plus particulièrement dans la prédiction de la réponse d'une tumeur à un traitement. Rappelons que le contexte biomédical de notre étude se situe dans le cadre de la prédiction de la réponse de cancers du côlon métastasés à une chimiothérapie combinée à une thérapie ciblée. Nous avons débuté une étude de caractérisation des cellules tumorales en considérant leur hétérogénéité, et du microenvironnement (le stroma) de façon à pouvoir cibler leur impact respectif sur la réponse au traitement (166). En effet, des modèles de prédiction considérant spécifiquement l'hétérogénéité tumorale et le microenvironnement pourraient être construits et évalués. Pour cela, une caractérisation du microenvironnement d'échantillons de tumeurs coliques humaines a été réalisée au moyen de colorations et marquages immuno-histochimiques : le rapport stroma/cellules tumorales et stroma fibreux est déterminé *via* une coloration au trichrome de Masson, un marquage des cellules immunitaires avec l'expression du marqueur CD163 pour les macrophages, des marqueurs CD3 pour les lymphocytes T et CD8 pour les lymphocytes T activés, et le marqueur CD31 pour la vascularisation. Outre ces caractérisations immuno-histologiques, il est important d'y associer une caractérisation génétique et moléculaire avec par exemple le phénotype (chromosome stable, instable, ADN méthylé) et le profil mutationnel des gènes BRAF et KRAS dans la mesure où ces paramètres caractérisent une tumeur, sa progression et sa réponse aux traitements.

Bibliographie

1. Sung H, Ferlay J, Siegel RL, Laversanne M, Soerjomataram I, Jemal A, et al. Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. CA Cancer J Clin. 2021;71(3):209-49.

2. Le cancer colorectal - Les cancers les plus fréquents. Site web de l'Institut National du Cancer.

3. Binder-Foucard F, Bossard N, Delafosse P, Belot A, Woronoff A-S, Remontet L. Cancer incidence and mortality in France over the 1980-2012 period: Solid tumors. Revue d'Epidemiologie et de Sante Publique. 2014;62(2):95-108.

4. Araghi M, Soerjomataram I, Bardot A, Ferlay J, Cabasag CJ, Morrison DS, et al. Changes in colorectal cancer incidence in seven high-income countries: a population-based study. Lancet Gastroenterol Hepatol. 2019;4(7):511-8.

5. L'incidence du cancer colorectal selon le sexe et le site anatomique. INSPQ.

6. Siegel RL, Torre LA, Soerjomataram I, Hayes RB, Bray F, Weber TK, et al. Global patterns and trends in colorectal cancer incidence in young adults. Gut. 2019;68(12):2179-85.

7. Chien C, Morimoto LM, Tom J, Li CI. Differences in colorectal carcinoma stage and survival by race and ethnicity. Cancer. 2005;104(3):629-39.

8. Cheng L, Eng C, Nieman LZ, Kapadia AS, Du XL. Trends in colorectal cancer incidence by anatomic site and disease stage in the United States from 1976 to 2005. Am J Clin Oncol Cancer Clin Trials. 2011;34(6):573-80.

9. Marley AR, Nan H. Epidemiology of colorectal cancer. Int J Mol Epidemiol Genet. 2016;7(3):105-14.

10. Arnold M, Sierra MS, Laversanne M, Soerjomataram I, Jemal A, Bray F. Global patterns and trends in colorectal cancer incidence and mortality. Gut. 2017;66(4):683-91.

11. Keum NN, Giovannucci E. Global burden of colorectal cancer: emerging trends, risk factors and prevention strategies. Nat Rev Gastroenterol Hepatol. 2019;16(12):713-32.

12. Fidler MM, Soerjomataram I, Bray F. A global view on cancer incidence and national levels of the human development index. Int J Cancer. 2016;139(11):2436-46.

13. Amitay EL, Carr PR, Jansen L, Roth W, Alwers E, Herpel E, et al. Smoking, alcohol consumption and colorectal cancer risk by molecular pathological subtypes and pathways. British Journal of Cancer. 2020;122(11):1604-10.

14. Magalhães B, Peleteiro B, Lunet N. Dietary patterns and colorectal cancer: Systematic review and meta-analysis. EurJ Cancer Prev. 2012;21(1):15-23.

15. Richi EB, Baumer B, Conrad B, Darioli R, Schmid A, Keller U. Health risks associated with meat consumption: A review of epidemiological studies. International Journal for Vitamin and Nutrition Research. 2015;85(1-2):70-8.

16. Thanikachalam K, Khan G. Colorectal cancer and nutrition. Nutrients. 2019;11(1).

17. Coppedè F. Epigenetic biomarkers of colorectal cancer: Focus on DNA methylation. Cancer Lett. 2014;342(2):238-47.

18. Weisenberger DJ, Liang G, Lenz H-J. DNA methylation aberrancies delineate clinically distinct subsets of colorectal cancer and provide novel targets for epigenetic therapies. Oncogene. 2018;37(5):566-77.

19. Fearon ER, Vogelstein B. A genetic model for colorectal tumorigenesis. Cell. 1990;61(5):759-67.

20. Lynch HT, De la Chapelle A. Hereditary colorectal cancer. New Engl J Med. 2003;348(10):919-32.

21. Sinicrope FA. Lynch syndrome-associated colorectal cancer. New England Journal of Medicine. 2018;379(8):764-73.

22. Half E, Bercovich D, Rozen P. Familial adenomatous polyposis. Orphanet Journal of Rare Diseases. 2009;4(1).

23. Bettington M, Walker N, Clouston A, Brown I, Leggett B, Whitehall V. The serrated pathway to colorectal carcinoma: Current concepts and challenges. Histopathology. 2013;62(3):367-86.

24. Snover DC. Sessile serrated adenoma/polyp of the large intestine: A potentially aggressive lesion in need of a new screening strategy. Diseases of the Côlon and Rectum. 2011;54(10):1205-6.

25. Kuipers EJ, Grady WM, Lieberman D, Seufferlein T, Sung JJ, Boelens PG, et al. Colorectal cancer. Nature Reviews Disease Primers. 2015;1.

26. Birkenkamp-Demtroder K, Olesen SH, Sørensen FB, Laurberg S, Laiho P, Aaltonen LA, et al. Differential gene expression in côlon cancer of the caecum versus the sigmoid and rectosigmoid. Gut. 2005;54(3):374-84.

27. Missiaglia E, Jacobs B, D'Ario G, Di Narzo AF, Soneson C, Budinska E, et al. Distal and proximal côlon cancers differ in terms of molecular, pathological, and clinical features. Annals of Oncology. 2014;25(10):1995-2001.

28. Colorectal Cancer - Statistics. Cancer.Net. 2012.

29. Fearon ER. Molecular genetics of colorectal cancer. Annu. Rev. Pathol. Mech. Dis. 2011. 479 p. (Annual Review of Pathology: Mechanisms of Disease; vol. 6).

30. Toyota M, Ahuja N, Ohe-Toyota M, Herman JG, Baylin SB, Issa J-PJ. CpG island methylator phenotype in colorectal cancer. Proceedings of the National Academy of Sciences of the United States of America. 1999;96(15):8681-6.

31. Spirio LN, Samowitz W, Robertson J, Robertson M, Burt RW, Leppert M, et al. Alleles of APC modulate the frequency and classes of mutations that lead to côlon polyps. Nature Genetics. 1998;20(4):385-8.

32. Markowitz SD, Bertagnolli MM. Molecular Basis of Colorectal Cancer. N Engl J Med. 17 déc 2009;361(25):2449-60.

33. Collura A, Lefevre JH, Svrcek M, Tougeron D, Zaanan A, Duval A. Microsatellite instability and cancer: From genomic instability to personalized medicine. Medecine/Sciences. 2019;35(6-7):535-43.

34. Samowitz WS, Albertsen H, Herrick J, Levin TR, Sweeney C, Murtaugh MA, et al. Evaluation of a large, population-based sample supports a CpG island methylator phenotype in côlon cancer. Gastroenterology. 2005;129(3):837-45.

35. Zaanan A, Meunier K, Sangar F, Fléjou J-F, Praz F. Microsatellite instability in colorectal cancer: From molecular oncogenic mechanisms to clinical implications. Cellular Oncology. 2011;34(3):155-76.

36. Samowitz WS, Curtin K, Ma K-N, Schaffer D, Coleman LW, Leppert M, et al. Microsatellite Instability in Sporadic Côlon Cancer Is Associated with an Improved Prognosis at the Population Level. Cancer Epidemiol Biomarkers Prev. 1 sept 2001;10(9):917.

37. Rhee Y-Y, Kim K-J, Kang GH. CpG Island methylator phenotype-high colorectal cancers and their prognostic implications and relationships with the serrated Neoplasia pathway. Gut and Liver. 2017;11(1):38-46.

38. Ogino S, Kawasaki T, Kirkner GJ, Kraft P, Loda M, Fuchs CS. Evaluation of Markers for CpG Island Methylator Phenotype (CIMP) in Colorectal Cancer by a Large Population-Based Sample. The Journal of Molecular Diagnostics. 1 juill 2007;9(3):305-14.

39. Gallois C, Laurent-Puig P, Taieb J. Methylator phenotype in colorectal cancer: A prognostic factor or not? Critical Reviews in Oncology/Hematology. 2016;99:74-80.

40. Ogino S, Goel A. Molecular classification and correlates in colorectal cancer. Journal of Molecular Diagnostics. 2008;10(1):13-27.

41. Guinney J, Dienstmann R, Wang X, De Reyniès A, Schlicker A, Soneson C, et al. The consensus molecular subtypes of colorectal cancer. Nat Med. 2015;21(11):1350-6.

42. Stintzing S, Wirapati P, Lenz H-J, Neureiter D, Fischer von Weikersthal L, Decker T, et al. Consensus molecular subgroups (CMS) of colorectal cancer (CRC) and first-line efficacy of FOLFIRI plus cétuximab or bevacizumab in the FIRE3 (AIO KRK-0306) trial. Annals of Oncology. 2019;30(11):1796-803.

43. Shimizu Y, Ikeda S, Fujimori M, Kodama S, Nakahara M, Okajima M, et al. Frequent alterations in the Wnt signaling pathway in colorectal cancer with microsatellite instability. Genes, Chromosomes and Cancer. 2002;33(1):73-81.

44. Downward J. Targeting RAS signalling pathways in cancer therapy. Nature Reviews Cancer. 2003;3(1):11-22.

45. Willett CG, Chang DT, Czito BG, Meyer J, Wo J. Cancer Genome Atlas Network. Comprehensive molecular characterization of human côlon and rectal cancer. Nature 2012. (5). INTJRADIATONCOLBIOLPHYS. 1 mai 2013;86(1).

46. Thiel A, Ristimäki A. Toward a molecular classification of colorectal cancer: The role of BRAF. Frontiers in Oncology. 2013;3 NOV.

47. Weisenberger DJ, Siegmund KD, Campan M, Young J, Long TI, Faasse MA, et al. CpG island methylator phenotype underlies sporadic microsatellite instability and is tightly associated with BRAF mutation in colorectal cancer. Nature Genetics. 2006;38(7):787-93.

48. Ardekani GS, Jafarnejad SM, Tan L, Saeedi A, Li G. The Prognostic Value of BRAF Mutation in Colorectal Cancer and Melanoma: A Systematic Review and Meta-Analysis. PLOS ONE. 9 oct 2012;7(10):e47054.

49. Heldin C-H, Miyazono K, ten Dijke P. TGF-β signalling from cell membrane to nucleus through SMAD proteins. Nature. déc 1997;390(6659):465-71.

50. Grady WM, Myeroff LL, Swinler SE, Rajput A, Thiagalingam S, Lutterbaugh JD, et al. Mutational Inactivation of Transforming Growth Factor β Receptor Type II in Microsatellite Stable Côlon Cancers. Cancer Res. 15 janv 1999;59(2):320.

51. Zavadil J, Böttinger EP. TGF-β and epithelial-to-mesenchymal transitions. Oncogene. 2005;24(37):5764-74.

52. Greenblatt MS, Bennett WP, Hollstein M, Harris CC. Mutations in the p53 Tumor Suppressor Gene: Clues to Cancer Etiology and Molecular Pathogenesist. :25.

53. Malki A, Elruz RA, Gupta I, Allouch A, Vranic S, Al Moustafa A-E. Molecular mechanisms of côlon cancer progression and metastasis: Recent insights and advancements. International Journal of Molecular Sciences. 2021;22(1):1-24.

54. Van Cutsem E, Lenz H-J, Köhne C-H, Heinemann V, Tejpar S, Melezínek I, et al. Fluorouracil, leucovorin, and irinotecan plus cétuximab treatment and RAS mutations in colorectal cancer. Journal of Clinical Oncology. 2015;33(7):692-700.

55. Douillard J-Y, Oliner KS, Siena S, Tabernero J, Burkes R, Barugel M, et al. Panitumumab-FOLFOX4 treatment and RAS mutations in colorectal cancer. New England Journal of Medicine. 2013;369(11):1023-34.

56. Gianchecchi E, Delfino DV, Fierabracci A. Recent insights into the role of the PD-1/PD-L1 pathway in immunological tolerance and autoimmunity. Autoimmunity Reviews. 1 sept 2013;12(11):1091-100.

57. Xie Y-H, Chen Y-X, Fang J-Y. Comprehensive review of targeted therapy for colorectal cancer. Signal Transduction and Targeted Therapy. 2020;5(1).

58. Lièvre A, Blons H, Laurent-Puig P. Oncogenic mutations as predictive factors in colorectal cancer. Oncogene. 2010;29(21):3033-43.

59. Cutsem EV, Rougier P, Köhne C, Stroh C, Schlichting M, Bokemeyer C. 6077 A meta-analysis of the CRYSTAL and OPUS studies combining cétuximab with chemotherapy

(CT) as 1st-line treatment for patients (pts) with metastatic colorectal cancer (mCRC): Results according to KRAS and BRAF mutation status. EJC Supplements. 2009;2(7):345.

60. Di Nicolantonio F, Martini M, Molinari F, Sartore-Bianchi A, Arena S, Saletti P, et al. Wild-type BRAF is required for response to panitumumab or cétuximab in metastatic colorectal cancer. Journal of Clinical Oncology. 2008;26(35):5705-12.

61. Ardekani GS, Jafarnejad SM, Tan L, Saeedi A, Li G. The Prognostic Value of BRAF Mutation in Colorectal Cancer and Melanoma: A Systematic Review and Meta-Analysis. PLOS ONE. 9 oct 2012;7(10):e47054.

62. Webber EM, Kauffman TL, O'Connor E, Goddard KAB. Systematic review of the predictive effect of MSI status in colorectal cancer patients undergoing 5FU-based chemotherapy. BMC Cancer. 2015;15(1).

63. Goodman AM, Kato S, Bazhenova L, Patel SP, Frampton GM, Miller V, et al. Tumor Mutational Burden as an Independent Predictor of Response to Immunotherapy in Diverse Cancers. Mol Cancer Ther. 1 nov 2017;16(11):2598.

64. McGranahan N, Furness AJS, Rosenthal R, Ramskov S, Lyngaa R, Saini SK, et al. Clonal neoantigens elicit T cell immunoreactivity and sensitivity to immune checkpoint blockade. Science. 2016;351(6280):1463-9.

65. Mouw KW, Goldberg MS, Konstantinopoulos PA, D'Andrea AD. DNA Damage and Repair Biomarkers of Immunotherapy Response. Cancer Discov. 1 juill 2017;7(7):675.

66. Hebbar M, Pruvot F-R, Romano O, Triboulet J-P, de Gramont A. Integration of neoadjuvant and adjuvant chemotherapy in patients with resectable liver metastases from colorectal cancer. Cancer Treatment Reviews. 2009;35(8):668-75.

67. Koukourakis MI, Mavanis I, Kouklakis G, Pitiakoudis M, Minopoulos G, Manolas C, et al. Early antivascular effects of bevacizumab anti-VEGF monoclonal antibody on colorectal carcinomas assessed with functional CT imaging. American Journal of Clinical Oncology: Cancer Clinical Trials. 2007;30(3):315-8.

68. Fournier LS, Ammari S, Thiam R, Cuénod C-A. Critères de la réponse tumorale en imagerie : RECIST, mRECIST, Cheson. Journal de Radiologie Diagnostique et Interventionnelle. 1 juill 2014;95(7):678-92.

69. de Geus-Oei LF, van Laarhoven HWM, Visser EP, Hermsen R, van Hoorn BA, Kamm YJL, et al. Chemotherapy response evaluation with FDG-PET in patients with colorectal cancer. Annals of Oncology. 2008;19(2):348-52.

70. Figueiras RG, Goh V, Padhani AR, Naveira AB, Caamaño AG, Martin CV. The role of functional imaging in colorectal cancer. American Journal of Roentgenology. 2010;195(1):54-66.

71. Skougaard K, Nielsen D, Jensen BV, Pfeiffer P, Hendel HW. Early 18F-FDG-PET/CT as a predictive marker for treatment response and survival in patients with metastatic colorectal cancer treated with irinotecan and cétuximab. Acta Oncologica. 2016;55(9-10):1175-82.

72. Spectroscopie Dans I'infrarouge - Google Livres.

73. Schrader B. Infrared and Raman Spectroscopy: Methods and Applications. John Wiley & Sons; 2008. 814 p.

74. Miller LM, Bourassa MW, Smith RJ. FTIR spectroscopic imaging of protein aggregation in living cells. Biochimica et Biophysica Acta - Biomembranes. 2013;1828(10):2339-46.

75. Travo A, Piot O, Wolthuis R, Gobinet C, Manfait M, Bara J, et al. IR spectral imaging of secreted mucus: A promising new tool for the histopathological recognition of human côlonic adenocarcinomas. Histopathology. 2010;56(7):921-31.

76. Taleb I, Thiéfin G, Gobinet C, Untereiner V, Bernard-Chabert B, Heurgué A, et al. Diagnosis of hepatocellular carcinoma in cirrhotic patients: A proof-of-concept study using serum micro-raman spectroscopy. Analyst. 2013;138(14):4006-14.

77. Gauglitz G, Moore DS. Handbook of Spectroscopy. John Wiley & Sons; 2014. 2011 p.

78. Baker MJ, Trevisan J, Bassan P, Bhargava R, Butler HJ, Dorling KM, et al. Using Fourier transform IR spectroscopy to analyze biological materials. Nature Protocols. 2014;9(8):1771-91.

79. Lasch P, Naumann D. Spatial resolution in infrared microspectroscopic imaging of tissues. Biochimica et Biophysica Acta - Biomembranes. 2006;1758(7):814-29.

80. Larkin P. Infrared and Raman Spectroscopy: Principles and Spectral Interpretation. Elsevier; 2017. 288 p.

81. Baranska M. Optical Spectroscopy and Computational Methods in Biology and Medicine. 2013. 540 p.

82. Kazarian SG, Chan KLA. ATR-FTIR spectroscopic imaging: Recent advances and applications to biological systems. Analyst. 2013;138(7):1940-51.

83. Palombo F, Shen H, Benguigui LES, Kazarian SG, Upmacis RK. Micro ATR-FTIR spectroscopic imaging of atherosclerosis: An investigation of the contribution of inducible nitric oxide synthase to lesion composition in ApoE-null mice. Analyst. 2009;134(6):1107-18.

84. Beć KB, Grabska J, Huck CW. Biomolecular and bioanalytical applications of infrared spectroscopy – A review. Anal Chim Acta. 2020;1133:150-77.

85. Martens H, Nielsen JP, Engelsen SB. Light scattering and light absorbance separated by extended multiplicative signal correction. Application to near-infrared transmission analysis of powder mixtures. Analytical Chemistry. 2003;75(3):394-404.

86. Afseth NK, Kohler A. Extended multiplicative signal correction in vibrational spectroscopy, a tutorial. Chemometrics and Intelligent Laboratory Systems. 1 août 2012;117:92-9.

87. Mohlenhoff B, Romeo M, Diem M, Wood BR. Mie-type scattering and non-Beer-Lambert absorption behavior of human cells in infrared microspectroscopy. Biophysical Journal. 2005;88(5):3635-40.

88. Bassan P, Kohler A, Martens H, Lee J, Byrne HJ, Dumas P, et al. Resonant Mie Scattering (RMieS) correction of infrared spectra from highly scattering biological samples. Analyst. 2010;135(2):268-77.

89. Konevskikh T, Lukacs R, Kohler A. An improved algorithm for fast resonant Mie scatter correction of infrared spectra of cells and tissues. J Biophotonics. 2018;11(1).

90. Ly E, Piot O, Wolthuis R, Durlach A, Bernard P, Manfait M. Combination of FTIR spectral imaging and chemometrics for tumour detection from paraffin-embedded biopsies. Analyst. 2008;133(2):197-205.

91. Amigo JM. Practical issues of hyperspectral imaging analysis of solid dosage forms. Anal Bioanal Chem. 2010;398(1):93-109.

92. Trevisan J, Angelov PP, Carmichael PL, Scott AD, Martin FL. Extracting biological information with computational analysis of Fourier-transform infrared (FTIR) biospectroscopy datasets: Current practices to future perspectives. Analyst. 2012;137(14):3202-15.

93. Nallala J, Piot O, Diebold M-D, Gobinet C, Bouché O, Manfait M, et al. Infrared imaging as a cancer diagnostic tool: Introducing a new concept of spectral barcodes for identifying molecular changes in côlon tumors. Cytometry Part A. 2013;83 A(3):294-300.

94. Peter Lasch, Juergen Schmitt, Dieter Naumann. Colorectal adenocarcinoma diagnosis by FT-IR microspectrometry. In 2000.

95. Ly E, Piot O, Durlach A, Bernard P, Manfait M. Differential diagnosis of cutaneous carcinomas by infrared spectral micro-imaging combined with pattern recognition. Analyst. 2009;134(6):1208-14.

96. Kaznowska E, Depciuch J, Szmuc K, Cebulski J. Use of FTIR spectroscopy and PCA-LDC analysis to identify cancerous lesions within the human côlon. Journal of Pharmaceutical and Biomedical Analysis. 2017;134:259-68.

97. Nallala J, Gobinet C, Diebold M-D, Untereiner V, Bouché O, Manfait M, et al. Infrared spectral imaging as a novel approach for histopathological recognition in côlon cancer diagnosis. Journal of Biomedical Optics. 2012;17(11).

98. Piqueras S, Duponchel L, Tauler R, De Juan A. Resolution and segmentation of hyperspectral biomedical images by Multivariate Curve Resolution-Alternating Least Squares. Analytica Chimica Acta. 2011;705(1-2):182-92.

99. Olmos V, Benítez L, Marro M, Loza-Alvarez P, Piña B, Tauler R, et al. Relevant aspects of unmixing/resolution analysis for the interpretation of biological vibrational hyperspectral images. TrAC Trends in Analytical Chemistry. 1 sept 2017;94:130-40.

100. MacKanos MA, Hargrove J, Wolters R, Du CB, Friedland S, Soetikno RM, et al. Use of an endoscope-compatible probe to detect côlonic dysplasia with Fourier transform infrared spectroscopy. Journal of Biomedical Optics. 2009;14(4).

101. Nguyen TNQ, Jeannesson P, Groh A, Piot O, Guenot D, Gobinet C. Fully unsupervised inter-individual IR spectral histology of paraffinized tissue sections of normal côlon. J Biophotonics. 2016;9(5):521-32.

102. Farah I, Nguyen TNQ, Groh A, Guenot D, Jeannesson P, Gobinet C. Development of a memetic clustering algorithm for optimal spectral histology: Application to FTIR images of normal human côlon. Analyst. 2016;141(11):3296-304.

103. Bird B, Miljkovi M, Remiszewski S, Akalin A, Kon M, Diem M. Infrared spectral histopathology (SHP): A novel diagnostic tool for the accurate classification of lung cancer. Laboratory Investigation. 2012;92(9):1358-73.

104. Lasch P, Haensch W, Naumann D, Diem M. Imaging of colorectal adenocarcinoma using FT-IR microspectroscopy and cluster analysis. Biochim Biophys Acta Mol Basis Dis. 2004;1688(2):176-86.

105. Kwak JT, Kajdacsy-Balla A, Macias V, Walsh M, Sinha S, Bhargava R. Improving prediction of prostate cancer recurrence using chemical imaging. Sci Rep. 2015;5.

106. Khanmohammadi M, Bagheri Garmarudi A, Samani S, Ghasemi K, Ashuri A. Application of linear discriminant analysis and attenuated total reflectance fourier transform infrared microspectroscopy for diagnosis of côlon cancer. Pathol Oncol Res. 2011;17(2):435-41.

107. Nallala J, Diebold M-D, Gobinet C, Bouché O, Sockalingum GD, Piot O, et al. Infrared spectral histopathology for cancer diagnosis: A novel approach for automated pattern recognition of côlon adenocarcinoma. Analyst. 2014;139(16):4005-15.

108. Lasch P. Diagnostic Potentials of FT-IR Microspectrometry in the Examination of Colorectal Adenocarcinomas. 21 juin 2007;

109. Mayerich DM, Walsh M, Kadjacsy-Balla A, Mittal S, Bhargava R. Breast histopathology using random decision forests-based classification of infrared spectroscopic imaging data. In 2014.

110. Bird B, Miljkovi M, Remiszewski S, Akalin A, Kon M, Diem M. Infrared spectral histopathology (SHP): A novel diagnostic tool for the accurate classification of lung cancer. Laboratory Investigation. 2012;92(9):1358-73.

111. Kallenbach-Thieltges A, Großerüschkamp F, Mosig A, Diem M, Tannapfel A, Gerwert K. Immunohistochemistry, histopathology and infrared spectral histopathology of côlon cancer tissue sections. J Biophotonics. 2013;6(1):88-100.

112. Lasch P, Haensch W, Naumann D, Diem M. Imaging of colorectal adenocarcinoma using FT-IR microspectroscopy and cluster analysis. Biochimica et Biophysica Acta - Molecular Basis of Disease. 2004;1688(2):176-86.

113. Marzec KM, Wrobel TP, Rygula A, Maslak E, Jasztal A, Fedorowicz A, et al. Visualization of the biochemical markers of atherosclerotic plaque with the use of Raman, IR and AFM. Journal of Biophotonics. 2014;7(9):744-56.

114. Severcan F, Bozkurt O, Gurbanov R, Gorgulu G. FT-IR spectroscopy in diagnosis of diabetes in rat animal model. Journal of Biophotonics. 2010;3(8-9):621-31.

115. Pallua JD, Pezzei C, Zelger B, Schaefer G, Bittner LK, Huck-Pezzei VA, et al. Fourier transform infrared imaging analysis in discrimination studies of squamous cell carcinoma. Analyst. 2012;137(17):3965-74.

116. Kochan K, Heraud P, Kiupel M, Yuzbasiyan-Gurkan V, McNaughton D, Baranska M, et al. Comparison of FTIR transmission and transfection substrates for canine liver cancer detection. Analyst. 2015;140(7):2402-11.

117. Zawlik I, Kaznowska E, Cebulski J, Kolodziej M, Depciuch J, Vongsvivut J, et al. FPA-FTIR Microspectroscopy for Monitoring Chemotherapy Efficacy in Triple-Negative Breast Cancer. Sci Rep. 2016;6.

118. Augustyniak K, Chrabaszcz K, Jasztal A, Smeda M, Quintas G, Kuligowski J, et al. High and ultra-high definition of infrared spectral histopathology gives an insight into chemical environment of lung metastases in breast cancer. J Biophotonics. 2019;12(4).

119. Akalin A, Mu X, Kon MA, Ergin A, Remiszewski SH, Thompson CM, et al. Classification of malignant and benign tumors of the lung by infrared spectral histopathology (SHP). Laboratory Investigation. 2015;95(4):406-21.

120. Großerueschkamp F, Kallenbach-Thieltges A, Behrens T, Brüning T, Altmayer M, Stamatis G, et al. Marker-free automated histopathological annotation of lung tumour subtypes by FTIR imaging. Analyst. 2015;140(7):2114-20.

121. Wolthuis R, Travo A, Nicolet C, Neuville A, Gaub M-P, Guenot D, et al. IR spectral imaging for histopathological characterization of xenografted human côlon carcinomas. Anal Chem. 2008;80(22):8461-9.

122. Dong L, Sun X, Chao Z, Zhang S, Zheng J, Gurung R, et al. Evaluation of FTIR spectroscopy as diagnostic tool for colorectal cancer using spectral analysis. Spectrochim Acta Part A Mol Biomol Spectrosc. 2014;122:288-94.

123. Baker MJ, Gazi E, Brown MD, Shanks JH, Clarke NW, Gardner P. Investigating FTIR based histopathology for the diagnosis of prostate cancer. Journal of Biophotonics. 2009;2(1-2):104-13.

124. Kuepper C, Großerueschkamp F, Kallenbach-Thieltges A, Mosig A, Tannapfel A, Gerwert K. Label-free classification of côlon cancer grading using infrared spectral histopathology. Faraday Discuss. 2016;187:105-18.

125. Gaydou V, Polette M, Gobinet C, Kileztky C, Angiboust J-F, Birembaut P, et al. New insights into spectral histopathology: infrared-based scoring of tumour aggressiveness of squamous cell lung carcinomas. Chemical Science. 2019;10(15):4246-58.

126. Mazza C, Gaydou V, Eymard J-C, Birembaut P, Untereiner V, Côté J-F, et al. Identification of neoadjuvant chemotherapy response in muscle-invasive bladder cancer by fourier-transform infrared micro-imaging. Cancers. 14(1).

127. Tiwari S, Bhargava R. Extracting knowledge from chemical imaging data using computational algorithms for digital cancer diagnosis. Yale J Biol Med. 2015;88(2):131-43.

128. Histological Classification of Raman Spectra of Human Coronary Artery Atherosclerosis Using Principal Component Analysis.

129. Ly E, Piot O, Wolthuis R, Durlach A, Bernard P, Manfait M. Combination of FTIR spectral imaging and chemometrics for tumour detection from paraffin-embedded biopsies. Analyst. 2008;133(2):197-205.

130. Sebiskveradze D, Vrabie V, Gobinet C, Durlach A, Bernard P, Ly E, et al. Automation of an algorithm based on fuzzy clustering for analyzing tumoral heterogeneity in human skin carcinoma tissue sections. Laboratory Investigation. 2011;91(5):799-811.

131. Nallala J, Gobinet C, Diebold M-D, Untereiner V, Bouché O, Manfait M, et al. Infrared spectral imaging as a novel approach for histopathological recognition in côlon cancer diagnosis. Journal of Biomedical Optics. 2012;17(11).

132. Husnain M, Missen MMS, Mumtaz S, Luqman MM, Coustaty M, Ogier J-M. Visualization of high-dimensional data by pairwise fusion matrices using t-SNE. Symmetry. 2019;11(1).

133. Van Der Maaten L, Hinton G. Visualizing data using t-SNE. Journal of Machine Learning Research. 2008;9:2579-625.

134. Kobak D, Berens P. The art of using t-SNE for single-cell transcriptomics. Nature Communications. 2019;10(1).

135. Hughes C, Iqbal-Wahid J, Brown M, Shanks JH, Eustace A, Denley H, et al. FTIR microspectroscopy of selected rare diverse sub-variants of carcinoma of the urinary bladder. J Biophotonics. 2013;6(1):73-87.

136. Wang W, Zhang Y. On fuzzy cluster validity indices. Fuzzy Sets and Systems. 2007;158(19):2095-117.

137. Pakhira MK, Bandyopadhyay S, Maulik U. Validity index for crisp and fuzzy clusters. Pattern Recognition. 2004;37(3):487-501.

138. Zanaty EA. Determining the number of clusters for kernelized fuzzy C-means algorithms for automatic medical image segmentation. Egyptian Informatics Journal. 2012;13(1):39-58.

139. Wang X-Y, Garibaldi JM, Bird B, George MW. A novel fuzzy clustering algorithm for the analysis of axillary lymph node tissue sections. Applied Intelligence. 2007;27(3):237-48.

140. Nguyen TNQ, Jeannesson P, Groh A, Guenot D, Gobinet C. Development of a hierarchical double application of crisp cluster validity indices: A proof-of-concept study for automated FTIR spectral histology. Analyst. 2015;140(7):2439-48.

141. Fatima A, Cyril G, Vincent V, Stéphane J, Olivier P. Towards normalization selection of Raman data in the context of protein glycation: Application of validity indices to PCA processed spectra. Analyst. 2020;145(8):2945-57.

142. Davies DL, Bouldin DW. A Cluster Separation Measure. IEEE Transactions on Pattern Analysis and Machine Intelligence. 1979;PAMI-1(2):224-7.

143. Xie XL, Beni G. A validity measure for fuzzy clustering. IEEE Transactions on Pattern Analysis and Machine Intelligence. août 1991;13(8):841-7.

144. Arbelaitz O, Gurrutxaga I, Muguerza J, Pérez JM, Perona I. An extensive comparative study of cluster validity indices. Pattern Recognition. 2013;46(1):243-56.

145. Vendramin L, Campello RJGB, Hruschka ER. On the Comparison of Relative Clustering Validity Criteria. In: Proceedings of the 2009 SIAM International Conference on Data Mining. Society for Industrial and Applied Mathematics; 2009. p. 733-44.

146. Holland JH. Adaptation in Natural and Artificial Systems: An Introductory Analysis with Applications to Biology, Control, and Artificial Intelligence. MIT Press; 1992. 228 p.

147. Hughes C, Gaunt L, Brown M, Clarke NW, Gardner P. Assessment of paraffin removal from prostate FFPE sections using transmission mode FTIR-FPA imaging. Analytical Methods. 2014;6(4):1028-35.

148. Faoláin EÓ, Hunter MB, Byrne JM, Kelehan P, Lambkin HA, Byrne HJ, et al. Raman spectroscopic evaluation of efficacy of current paraffin wax section dewaxing agents. Journal of Histochemistry and Cytochemistry. 2005;53(1):121-9.

149. Nallala J, Lloyd GR, Stone N. Evaluation of different tissue de-paraffinization procedures for infrared spectral imaging. Analyst. 2015;140(7):2369-75.

150. Ly E, Piot O, Wolthuis R, Durlach A, Bernard P, Manfait M. Combination of FTIR spectral imaging and chemometrics for tumour detection from paraffin-embedded biopsies. Analyst. 2008;133(2):197-205.

151. Gobinet C, Sebiskveradze D, Vrabie V, Tfayli A, Piot O, Manfait M. Digital dewaxing of Raman spectral images of paraffin-embedded human skin biopsies based on ICA and NCLS. In: 2008 16th European Signal Processing Conference. 2008. p. 1-5.

152. Meksiarun P, Ishigaki M, Huck-Pezzei VAC, Huck CW, Wongravee K, Sato H, et al. Comparison of multivariate analysis methods for extracting the paraffin component from the paraffin-embedded cancer tissue spectra for Raman imaging. Scientific Reports. 2017;7.

153. Olmos V, Benítez L, Marro M, Loza-Alvarez P, Piña B, Tauler R, et al. Relevant aspects of unmixing/resolution analysis for the interpretation of biological vibrational hyperspectral images. TrAC - Trends in Analytical Chemistry. 2017;94:130-40.

154. Boutegrabet W, Guenot D, Bouché O, Boulagnon-Rombi C, Marchal Bressenot A, Piot O, et al. Automatic identification of paraffin pixels on FTIR images acquired on FFPE human samples. Anal Chem. 2 mars 2021;93(8):3750-61.

155. Byrne HJ, Knief P, Keating ME, Bonnier F. Spectral pre and post processing for infrared and Raman spectroscopy of biological tissues and cells. Chem Soc Rev. 2016;45(7):1865-78.

156. Piqueras S, Krafft C, Beleites C, Egodage K, von Eggeling F, Guntinas-Lichius O, et al. Combining multiset resolution and segmentation for hyperspectral image analysis of biological tissues. Analytica Chimica Acta. 2015;881:24-36.

157. Chrabaszcz K, Kochan K, Fedorowicz A, Jasztal A, Buczek E, Leslie LS, et al. FT-IR- and Raman-based biochemical profiling of the early stage of pulmonary metastasis of breast cancer in mice. Analyst. 30 avr 2018;143(9):2042-50.

158. Colin-Pierre C, Untereiner V, Sockalingum GD, Berthélémy N, Danoux L, Bardey V, et al. Hair Histology and Glycosaminoglycans Distribution Probed by Infrared Spectral Imaging: Focus on Heparan Sulfate Proteoglycan and Glypican-1 during Hair Growth Cycle. Biomolecules. févr 2021;11(2):192.

159. Truong JXM, Spotbeen X, White J, Swinnen JV, Butler LM, Snel MF, et al. Removal of optimal cutting temperature (O.C.T.) compound from embedded tissue for MALDI imaging of lipids. Analytical and Bioanalytical Chemistry. 2021;413(10):2695-708.

160. Trevisan J, Park J, Angelov PP, Ahmadzai AA, Gajjar K, Scott AD, et al. Measuring similarity and improving stability in biomarker identification methods applied to Fourier-transform infrared (FTIR) spectroscopy. Journal of Biophotonics. 2014;7(3-4):254-65.

161. Hassani S, Martens H, Qannari EM, Hanafi M, Borge GI, Kohler A. Analysis of -omics data: Graphical interpretation- and validation tools in multi-block methods. Chemometrics and Intelligent Laboratory Systems. 2010;104(1):140-53.

162. Yang P, Hwa Yang Y, B. Zhou B, Y. Zomaya A. A Review of Ensemble Methods in Bioinformatics. Current Bioinformatics. 1 déc 2010;5(4):296-308.

163. Gajjar K, Trevisan J, Owens G, Keating PJ, Wood NJ, Stringfellow HF, et al. Fourier-transform infrared spectroscopy coupled with a classification machine for the analysis of blood plasma or serum: A novel diagnostic approach for ovarian cancer. Analyst. 2013;138(14):3917-26.

164. Bird B, Rowlette J. High definition infrared chemical imaging of colorectal tissue using a Spero QCL microscope. Analyst. 2017;142(8):1381-6.

165. Fung AA, Shi L. Mammalian cell and tissue imaging using Raman and coherent Raman microscopy. WIREs Systems Biology and Medicine. 2020;12(6):e1501.

166. Mittal S, Yeh K, Leslie LS, Kenkel S, Kajdacsy-Balla A, Bhargava R. Simultaneous cancer and tumor microenvironment subtyping using confocal infrared microscopy for all-digital molecular histopathology. Proc Natl Acad Sci USA. 19 juin 2018;115(25):E5651-60.

Université de Strasbourg

Warda BOUTEGRABET

Histopathologie spectrale du cancer colique: Développements d'outils chimiométriques pour le traitement automatisé des images tissulaires en spectroscopie moyen-infrarouge

Résumé

Les cliniciens ont un besoin crucial de paramètres pour identifier en routine clinique, un patient dont la tumeur sera sensible ou non à un traitement. L'histopathologie spectrale par imagerie infrarouge mesure la composition biochimique des structures tissulaires sans coloration ni marquage. Bien que les études aient montré le potentiel diagnostique de cette technique, elle reste un outil de recherche sans être intégrée dans les services d'anatomopathologie. Ce travail de doctorat a porté sur des développements chimiométriques visant à rendre l'approche totalement automatisée pour un déploiement en routine clinique. Une méthode permettant d'éliminer les pixels parasites d'une image spectrale infrarouge a été élaborée ; ce développement est applicable aux tissus paraffinés et aux tissus congelés dans un gel biologique. Puis un algorithme de sélection non-supervisée de variables a été mis au point en vue d'optimiser l'exploitation du signal infrarouge. Ce travail trouve un intérêt dans l'identification de marqueurs spectroscopiques pour des applications médicales telle que la prédiction de la réponse au traitement des cancers du côlon métastasés.

Mots clés :
Histopathologie spectrale, imagerie infrarouge, chimiomtrie, cancer de côlon.

Résumé en anglais

Clinicians have a crucial need for parameters to identify in clinical routine, a patient whose tumor will be sensitive or not to a treatment. Spectral histopathology by infrared imaging measures the biochemical composition of tissue structures without staining or labelling. Although studies have shown the diagnostic potential of this technique, it remains a research tool without being integrated into pathology departments. This doctoral work focused on chemometric developments aimed at making the approach fully automated for deployment in clinical routine. A method for removing outliers pixels from an infrared spectral image has been developed; this development is applicable to paraffin-embedded tissues and to tissues frozen in a biological gel. Then, an algorithm of variable unsupervised selection was developed in order to optimize the exploitation of the infrared signal. This work finds an interest in the identification of spectroscopic markers for medical applications such as the prediction of the response to treatment of metastasized côlon cancers.