UNIVERSITÉ DE REIMS CHAMPAGNE-ARDENNE

ÉCOLE DOCTORALE SCIENCES TECHNOLOGIE SANTE (547)

THÈSE

Pour obtenir le grade de

DOCTEUR DE L'UNIVERSITÉ DE REIMS CHAMPAGNE-ARDENNE

Discipline : ***PHYSIQUE***

Spécialité : Bio-spectroscopie

Présentée et soutenue publiquement par

Thi Nguyet Que NGUYEN

Le 27 janvier 2016

NOUVEAUX DÉVELOPPEMENTS EN HISTOLOGIE SPECTRALE IR : APPLICATION AU TISSU COLIQUE

Thèse dirigée par **Pierre JEANNESSON**

JURY

Mme. Dominique GUENOT,	, Directeur de Recherche,	à l'Université de Strasbourg 1 Louis Pasteur,	**, Président**
M. Ludovic DUPONCHEL,	, Professeur,	à l'Université de Lille 1 Sciences et Technolog,	**, Rapporteur**
M. Didier WOLF,	, Professeur,	INP de Nancy,	**, Rapporteur**
M. Cyril GOBINET,	, Maître de Conférences,	à l'Université Reims Champagne-Ardenne,	**, Examinateur**
M. Pierre JEANNESSON,	, Professeur,	à l'Université Reims Champagne-Ardenne,	**, Examinateur**
M. Olivier PIOT,	, Professeur,	à l'Université Reims Champagne-Ardenne,	**, Examinateur**

(Thích Nhất Hạnh)

Remerciements

A Messieurs le Professeur Michel Manfait et le Professeur Olivier Piot,

Je vous adresse mes remerciements respectueux pour m'avoir accueillie dans votre équipe afin de réaliser ce travail de thèse. Je vous exprime également toute ma gratitude pour m'avoir donné l'occasion de présenter mes travaux dans de nombreux congrès nationaux et internationaux.

Monsieur le Professeur Ludovic Duponchel,

Je vous remercie sincèrement d'avoir accepté d'être rapporteur scientifique de ce manuscrit de thèse et de faire partie du jury de soutenance. Veuillez trouver toute ma reconnaissance, et soyez assuré de mon respect et de ma gratitude.

A Monsieur le Professeur Didier Wolf,

Je vous adresse mes remerciements respectueux pour l'honneur que vous me faites en acceptant d'être rapporteur scientifique et de juger ce travail de thèse. Veuillez accepter mes remerciements sincères ; soyez également assuré de mon respect et de ma gratitude.

A Madame le Docteur Dominique Guenot,

Je vous exprime toute ma reconnaissance et vous adresse mes remerciements les plus sincères pour m'avoir donné l'opportunité de collaborer avec votre équipe, et pour avoir accepté de faire partie de mon jury de thèse. Je vous remercie également pour votre ac-

cueil à de nombreuses reprises, pour vos conseils scientifiques, votre gentillesse et votre sympathie à mon égard.

A Monsieur le Professeur Pierre Jeannesson,,

Monsieur Jeannesson, une seule ligne de remerciement pour vous ne peut pas être suffisante. Vous m'avez appris à être critique, à prendre du recul, à valoriser les travaux, à simplifier les choses et à organiser mes idées. Vos critiques m'ont souvent permis de me "réveiller" et éviter ainsi de faire des mauvais choix. Ça fait mal, mais, sans ces remarques, il m'aurait été impossible de progresser. J'ai eu beaucoup de chance d'être encadré par vous, mon premier Papa de thèse. Votre soutien, vos encouragements et vos conseils ont donné du sucre à mes travaux pendant ces trois ans de thèse ; et bien sûr, je n'oublierai pas votre humour que parfois j'ai du mal à comprendre. Vous me manquerez.

Je vous adresse mes remerciements les plus respectueux, Papa !

A Monsieur le Docteur Cyril Gobinet,

Cyril, c'est difficile pour moi de savoir combien de remerciements je te dois (it's uncountable !). Au cours de mes années de thèse, tu m'as apporté beaucoup de connaissances en traitement de l'information, en programmation, et surtout beaucoup d'expérience pour "écrire et raconter une histoire". Tu m'as donné des conseils, des encouragements, et tu m'as orientée dans les bonnes directions non seulement pour le développement, mais aussi pour la recherche de méthodes que je n'avais jamais eu l'occasion d'étudier. Je n'oublierai pas mes premières leçons en EMSC, en algorithme génétique, et aussi en tennis :). Tu as toujours été généreux, sympathique, ouvert et tu m'as permis de progresser.

Merci infiniment pour tout, mon deuxième Papa de thèse !

A Monsieur le Docteur Ihsen Farah,

Ihsen, je tiens à te remercier sincèrement pour toute l'aide que tu m'as apportée au cours de ma dernière année de thèse. Grâce à toi, j'ai pu enrichir mes connaissances en métaheuristiques. Merci encore Ihsen, de m'avoir apporté ton soutien, tes encouragements et surtout des explications très claires et compréhensibles pour développer notre méthode d'algorithme mémétique.

A Madame Audrey Groh,

Audrey, je t'adresse mes remerciements sincères pour ton aide durant mes années de thèse. Je suis ravie d'avoir travaillé en collaboration avec toi tout au long de ces années. Merci pour ta générosité, ton accueil au sein du laboratoire, et ta disponibilité pour faire et refaire des coupes tissulaires, et mettre ainsi en valeur nos projets.

Je tiens également à remercier

Monsieur le Professeur Ganesh Sockalingum pour vos conseils méthodologiques et linguistiques pour que je puisse surmonter les difficultés au cours de mes années de thèse.

Monsieur le Docteur Jean-François Angiboust pour vos conseils informatiques et logistiques pendant mes années de thèse.

Laurence, Christiane, Christine et **Valérie** pour les conseils, les explications constructives sur l'utilisation des appareils et les techniques de laboratoire.

Mon bon ami de bureau, **monsieur le Docteur Mohammed Essendoubi**, qui m'a apporté beaucoup de conseils et d'encouragements pendant ces dernières années. Merci encore Momo pour ta gentillesse, ta générosité, ta sympathie et également ton ouverture d'esprit.

Fabricio, mon ami brésilien, pour la création des machines virtuelles. Grâce à toi, j'ai pu terminer à temps les traitements de mes données. Je tiens à te remercier également pour tes conseils, tes encouragements et ton amitié généreuse pendant ton séjour en France.

Christophe, merci pour tes encouragements, tes conseils et surtout le soutien que tu m'as apportés pendant ces années difficiles. Je me souviendrai des bons moments passés pendant les congrès et au laboratoire grâce à ta bonne humeur et ta générosité. Je te remercie pour tout, mon camarade de thèse !

Ma bande de copains, **Leila, David, Shawn, Louis, Michael, Charles, Hassan, Jaafar, Hadrien, Teddy, Lucie** et **Vincent** pour les bons moments passés ensemble.

Je remercie également mes collègues de labo que j'aime beaucoup, **Hamid, Jennifer, Céline, Nathalie M., Nathalie L., Marie-Pierre, Émilie, Lila, Pascaline, Aurélie, Marie, Joanne, Mathieu, Gouttam, Jaya**, et **Irène** pour votre sympathie au cours des années de thèse.

Je tiens à remercier sincèrement **Brigitte Perron** pour tes corrections grammaticales du manuscrit. Merci aussi pour ta gentillesse et ta bonne humeur.

A ma moitié, **Loïc** ! Je te remercie profondément pour tout ce que tu m'as donné, ta patience et ton soutien qui m'ont permis de surmonter les moments difficiles.

Bé Yến xin chân thành cảm ơn anh chị Hai, anh chị Ba và anh chị Tư lúc nào cũng mang đến cho Bé Yến nhiều niềm vui, lúc nào cũng động viên Bé Yến, nhất là từ những ngày đầu tiên Bé Yến đi học nơi xứ lạ quê người.

Lời cảm ơn cuối cùng, Bé Yến xin kính gởi đến đấng sinh thành của mình. Con cảm ơn Tía Má đã cho con tất cả để con có được ngày hôm nay !

Je dédie cette thèse à Madame Huỳnh Thị Tiết
et Monsieur Nguyễn Thành Khiết

Liste des communications

Publications dans des journaux internationaux

Nguyen, T. N. Q., 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 an automated FTIR spectral histology"*. **Analyst**, 2015, 140, (2439-2448).

Nguyen, T. N. Q., Jeannesson, P., Groh, A., Piot, O., Guenot, D., & Gobinet, C., *"Fully unsupervised inter-individual spectral histology of paraffinized tissue sections of normal colon"* - soumis à **Journal of Biophotonics**.

Farah, I., Nguyen, T. N. Q., Jeannesson, P., Groh, A., Guenot, D., & Gobinet, C., *"Development of a memetic clustering algorithm for optimal spectral histology : application to FTIR images of normal human colon"* - soumis à **Analyst**.

Publications dans des proceedings de conférences nationales

Nguyen, T. N. Q., Jeannesson, P., Groh, A., Guenot, D., & Gobinet, C., *"Fourier-transform infrared imaging and clustering : toward an automated histology of normal colon"*. Journées RiTS 2015, 2015, (146-147).

Farah, I., Nguyen, T. N. Q., Jeannesson, P., Groh, A., Guenot, D., & Gobinet, C., *"Optimal*

Spectral Histology of Human Normal Colon by Genetic Algorithm". Journées RiTS 2015, 2015, (178-179).

Communications

• Présentations orales

ECSM 2015 - 16th European Conference on the Spectroscopy of Biological Molecules , à Bochum, Allemagne (Sept. 2015).

"Fourier-transform infrared imaging and clustering : toward an automated histology of normal colon", Nguyen, T. N. Q., Jeannesson, P., Groh, A., Guenot, D., & Gobinet, C..

"Optimal Spectral Histology of Human Normal Colon by Genetic Algorithm", Farah, I., Nguyen, T. N. Q., Jeannesson, P., Groh, A., Guenot, D., & Gobinet, C..

GFSV 2015 - 21ème Journée du Groupe Français de Spectroscopie Vibrationnelle , à Reims (Juin 2015).

"Imagerie IRTF et classification non-supervisée : vers une histologie spectrale automatisée du colon normal" , Nguyen, T. N. Q., Jeannesson, P., Groh, A., Guenot, D., & Gobinet, C..

JJC 2015 - Journée des jeunes chercheurs, à Reims (Mars 2015).

"Histologie spectrale multi-images automatisée : application au tissu colique humain", Nguyen, T. N. Q., Jeannesson, P., Groh, A., Guenot, D., & Gobinet, C..

Séminaire dans l'unité MÉDyC, CNRS UMR 7369, à l'Université de Reims Champagne-Ardenne, Reims (Déc. 2014).

"Histologie spectrale infrarouge automatisée : application au colon normal".

Séminaire invité, dans l'équipe Progression tumorale et microenvironnement, Approches translationnelles et Epidémiologie, EA 3430, Fédération de Médecine Translationnelle de Strasbourg (FMTS), à l'Université de Strasbourg, Strasbourg (Nov. 2014).

"Histologie spectrale infrarouge automatisée : application au colon normal".

• Présentations par poster

RiTS 2015 - Recherche en Imagerie et Technologies pour la Santé , à Dourdan (Mars 2015).

"Fourier-transform infrared imaging and clustering : toward an automated histology of normal colon", Nguyen, T. N. Q., Jeannesson, P., Groh, A., Guenot, D., & Gobinet, C..

"Optimal Spectral Histology of Human Normal Colon by Genetic Algorithm", Farah, I., Nguyen, T. N. Q., Jeannesson, P., Groh, A., Guenot, D., & Gobinet, C..

SPEC 2014 Shedding New Light on Disease, à Crakovie, Pologne (Août 2014).

"Automatic spectral histology of normal human colon tissues by infrared microimaging and cluster validity indices", Nguyen, T. N. Q., Jeannesson, P., Groh, A., Guenot, D., & Gobinet, C..

ECDP 2014 - 12th European Congress of Digital Pathology à Paris (Juin 2014).

"Automatic spectral histology of human colon tissues by infrared microimaging and cluster validity indices", Nguyen, T. N. Q., Jeannesson, P., Groh, A., Guenot, D., & Gobinet, C..

Table des matières

Liste des abréviations **20**

Table des figures **21**

Liste des tableaux **25**

I Introduction **27**

II Approche méthodologique **33**

II.1 - Histologie conventionnelle du côlon normal 34

II.1.1 - Côlon humain 34

II.1.2 - Côlon murin 37

II.2 - Échantillons biologiques 39

II.2.1 - Préparation des échantillons 39

II.2.2 - Echantillons utilisés 39

II.3 - Histologie spectrale IR 40

II.3.1 - Acquisition des images spectrales IR 40

II.3.2 - Traitements numériques multivariés des images spectrales 44

II.3.3 - Affectation des classes aux structures histologiques 48

III Développement d'une histologie spectrale automatisée par indices de validité 51

III.1 - Préambule 52

III.2 - Article #1 : *"Development of a hierarchical double application of crisp cluster validity indices : a proof-of-concept study for automated FTIR spectral histology"* 56

III.3 - Résultats supplémentaires 67

III.3.1 - Application des indices de validité sur un jeu de données hiérarchiques multi-dimensionnelles 67

III.3.2 - Application des 28 indices de validité supplémentaires sur les images spectrales 70

III.3.3 - Confirmation des résultats sur 10 nouveaux patients 73

IV Développement d'une histologie spectrale multi-images 77

IV.1 - Préambule 78

IV.2 - Article #2 : *"Fully unsupervised inter-individual spectral histology of paraffinized tissue sections of normal colon"* 81

IV.3 - Résultats supplémentaires 108

IV.3.1 - Confirmation de l'efficacité du protocole non-automatisé à l'échelle intra- et inter-individuelle chez l'Homme 108

IV.3.2 - Impact des paramètres de l'EMSC sur le nombre de classes estimé par le protocole automatisé 110

V Développement d'une histologie spectrale optimale par la métaheuristique 113

V.1 - Préambule 114

V.2 - Article #3 : "*Development of a memetic clustering algorithm for optimal spectral histology : application to FTIR images of normal human colon*" . 117

V.3 - Résultats supplémentaires 152

V.3.1 - Confirmation de l'efficacité de CAM 152

V.3.2 - Modification du protocole automatisé par inclusion de CAM . . . 152

VI Conclusion et perspectives **155**

Bibliographie **163**

Liste des abréviations

AG	Algorithme Génétique
ANN	Artificial Neural Network
ATR	Attenuated Total Reflectance
CaF_2	Fluorure de calcium
CAM	Clustering par Algorithme Mémétique
EMSC	Extended Multiplicative Signal Correction
FCM	Fuzzy C-Means
FPA	Focal Plane Array
GABC	Genetic Algorithm-Based Clustering
GKA	Genetic KM Algorithm
HCA	Hierarchical Cluster Analysis
HE	Hématoxiline-Eosine
IRTF	InfraRouge à transformée de Fourier
IR	InfraRouge
KM	*k*-Means
LDA	Linear Discriminant Analysis
MCT	Mercure Cadmium Telluride
MNF	Minimum Noise Fraction

MSC Multiplicative Signal Correction

MW-U Mann–Whitney-U

PCA Principal Component Analysis

QCL Quantum Cascade Lasers

RF Random Forest

RMieSC Resonant Mie Scattering Correction

SG Savitzky-Golay

SIMCA Soft Independent Modelling by Class Analogy

SVM Support Vector Machines

Table des figures

I.1 **Différentes étapes du développement de l'instrumentation en spectroscopie moyen IR** . ATR, Attenuated Total Reflection. MCT, Mercury Cadmium Telluride. FPA, Focal Plane Arrays. QCL, Quantum Cascade Lasers. Sources des images utilisés (respectivement de gauche à droite) : herschel.cf.ac.uk, www.nasonline.org,[47],[27], synchrotron.org.au et[3]. . . . 28

I.2 **Diagramme des principales étapes du traitement numérique des images spectrales IR tissulaires**. EMSC, Extended Multiplicative Signal Correction. RMieSC, Resonant Mie Scattering Correction. MNF, Minimum Noise Fraction. SG, Savitzky-Golay. PCA, Principal Component Analysis. KM, *k*-Means. FCM, Fuzzy C-Means. HCA, Hierarchical Cluster Analysis. AG, Algorithme Génétique. MW-U, Mann–Whitney-U. SVM, Support Vector Machines. ANN, Artificial Neural Network. LDA, Linear Discriminant Analysis. SIMCA, Soft Independent Modelling by Class Analogy. RF, Random Forest. 30

II.1 **Anatomie du côlon humain**. 34

II.2 **Structure de la paroi du côlon humain**. 35

II.3 **Structure de la muqueuse du côlon humain après coloration par Hématoxyline-Eosine (HE)**. La muqueuse comprend différentes structures histologiques : (1) épithélium, (2) cryptes, (3) chorion, (4) musculaire muqueuse, et (5) îlot lymphoïde. (6) Sous-muqueuse. Échelle, 100 μm. 36

II.4 **Anatomie du côlon murin[78]**. Il commence par le cæcum en forme de J, et se continue par le côlon proximal, médian et distal. 38

II.5 **Structure histologique du côlon murin après coloration par HE**. Échelle, 100 μm. 38

II.6 **Exemple de modes de vibration du groupement CH_2.** Les modes de vibration peuvent être divisés en deux catégories : les modes d'élongation symétrique et asymétrique qui font varier la longueur des liaisons, et les modes de déformation faisant varier les angles de ces liaisons (cisaillement, balancement, rotation pure et torsion). 41

II.7 **Bandes d'absorption dans le moyen IR caractéristiques des biomolécules**. Acides nucléiques (820 cm^{-1} et 1720 cm^{-1}), carbohydrates (1000-1200 cm^{-1}), protéines (1500-1700 cm^{-1}), lipides (2700-3000 cm^{-1}). . . 42

II.8 **Micro-imageur spectral IR.** Le système est composé (A) d'un système d'imagerie (Micro Imager Spotlight 300, Perkin Elmer) et (B) d'un spectromètre IRTF (Spectrum One, Perkin Elmer). 43

II.9 **Exemple d'une image spectrale IR.** Cette image équivaut à un cube de données, où X et Y représentent les coordonnées spatiales de chaque spectre, et Λ les différents nombres d'onde de la gamme spectrale IR. . . 43

II.10 **Exemples de spectres IR d'une image acquise sur une coupe tissulaire paraffinée**. 45

II.11 **Effet du prétraitement par EMSC**. (A) Spectres IR acquis sur une coupe tissulaire, (B) spectres prétraités par EMSC. 47

II.12 **Attribution des classes d'une image pseudo-couleur aux structures histologiques du tissu**. (A) Image pseudo-couleur avec $k = 13$ classes, (B) coloration HE et (C) identification des structures histologiques. Échelles, 100 μm. 49

III.1 **Illustration du jeu de données synthétiques composé de 451 dimensions**. (A) Données synthétiques générées. (B) Représentation 2D de la structure hiérarchique des données synthétiques. (C) Représentation 2D des données synthétiques pour deux dimensions non informatives. 69

IV.1 **Comparaison entre l'histologie conventionnelle et l'histologie spectrale multi-images à l'échelle intra-individuelle pour le patient #1**. (A) Coloration HE. (B) Images en pseudo-couleurs reconstruites pour un nombre de classes $k = 11$ choisi empiriquement. Les images ont été acquises sur différentes zones d'un même échantillon. Échelles, 100 μm. . . 108

IV.2 **Comparaison entre l'histologie conventionnelle et l'histologie spectrale multi-images à l'échelle inter-individuelle pour le lot de 72 images des 15 patients**. Le nombre de classes $k = 9$ a été choisi empiriquement. Échelles, 100 μm. 109

IV.3 **Protocole automatisé avec application successive du modèle de paraffine $\mathbf{I_X}$ de chaque patient**. 110

IV.4 **Protocole automatisé avec application successive du spectre de référence $\hat{\mathbf{s}}_\mathbf{X}$ de chaque patient**. 111

VI.1 **Exemple d'une identification imparfaite de structures histologiques estimées par double application hiérarchique de PBM**. Les structures du chorion et de la musculaire muqueuse sont représentées par une seule et même classe. Échelles, 100 μm. 156

VI.2 **Résultat d'une troisième application d'indice PBM sur une classe sélectionnée**. (A) L'application de l'indice PBM sur la classe sélectionnée, représentant une mixture du chorion et de la musculaire muqueuse. (B) L'image en pseudo-couleur est construite en combinant les estimations des applications de l'indice PBM. Échelles, 100 μm. 156

VI.3 **Exemple de l'identification des sous-structures d'un îlot lymphoïde par double application hiérarchique de PBM**. Échelles, 100 μm. . . . 157

VI.4 **Exemple d'histologie spectrale inter-espèce par le protocole non-automatisé, avec $k = 7$ classes**. Échelles, 100 μm. 159

VI.5 **Extraction non-supervisée de bio-marqueurs par algorithme génétique**. (A) Coloration HE. (B) Image en pseudo-couleurs reconstruite après application de KM avec $k = 15$, sur le jeu de données réduit aux 30 variables sélectionnées par l'algorithme génétique. (C) Localisation sur les centroïdes, des variables sélectionnées (en magenta). Échelles, 100 μm. 161

Liste des tableaux

II.1 **Caractéristiques des échantillons coliques étudiés**. M, masculin. F, féminin. 40

III.1 **Nombre optimal de classes k_{opt} estimé par la simple et la double application d'indices de validité sur le jeu de données synthétiques de 451 dimensions**. Les valeurs grisées représentent le nombre estimé de classes induisant la partition la plus proche de la partition originale. . . . 70

III.2 **Liste des indices de validité supplémentaires**. Pour chaque indice, son nom, son acronyme et sa référence bibliographique sont donnés. 71

III.3 **Nombre optimal de classes $\mathbf{k_{opt}}$ et taux de précision estimés par la simple et la double application des indices de validité supplémentaires**. Les valeurs grisées représentent les partitions optimales qui retrouvent les structures histologiques principales du côlon humain normal. 72

III.4 **Nombre optimal de classes $\mathbf{k_{opt}}$ et taux de précision estimés par la simple application de l'ensemble des indices de validité sur les 10 patients supplémentaires**. Les valeurs grisées représentent les partitions optimales qui retrouvent les structures histologiques principales du côlon humain normal. 74

III.5 **Nombre optimal de classes k_{opt} et taux de précision estimés par la double application de l'ensemble des indices de validité sur les 10 patients supplémentaires**. Les valeurs grisées représentent les partitions optimales qui retrouvent les structures histologiques principales du côlon humain normal. 75

IV.1 **Nombre de classes k_{opt} estimé par le protocole automatisé pour (A) un modèle de paraffine variable et (B) un spectre de référence variable.** 111

V.1 **Moyenne $\overline{f}$ et écart-type σ de la fonction objectif de KM calculés sur 100 répétitions de KM et 10 répétitions de CAM, GAK et GABC**. Pour chaque patient, la valeur en gras représente la plus petite valeur de $\overline{f}$ parmi les quatre méthodes de clustering. 152

V.2 **Nombre de classes optimaux k_{opt} estimé par les deux protocoles automatisés basés sur CAM et KM**. La similarité entre les partitions estimées par ces deux méthodes a été calculée par le taux de bonne classification τ. 153

Chapitre I

Introduction

En histopathologie conventionnelle, les échantillons tissulaires issus de biopsies sont analysés par des anatomo-pathologistes au moyen de différentes techniques de coloration ou d'immuno-marquage, leur permettant ainsi d'émettre un diagnostic qui sera confronté aux données cliniques, biologiques et d'imagerie médicale. Actuellement, les micro-spectroscopies vibrationnelles Raman et IR apparaissent de plus en plus comme de nouveaux outils complémentaires de caractérisation tissulaire et d'aide au diagnostic applicables à terme au milieu clinique[3,36].

La micro-spectroscopie vibrationnelle d'absorption IR repose sur l'interaction lumière-matière et permet d'analyser un échantillon cellulaire ou tissulaire de manière rapide, non-destructive et sans marquage préalable. Les informations obtenues sont enregistrées sous forme de spectres représentant l'absorbance en fonction du nombre d'onde. Ensuite, le traitement statistique de ces informations spectrales permettra de sonder la composition et/ou la structure des grandes classes de biomolécules présentes dans l'échantillon notamment les protéines, les carbohydrates et les acides nucléiques.

Les grandes phases de l'évolution de la spectroscopie vibrationnelle d'absorption IR sont présentées dans la Figure I.1. Adaptés à l'IR dans les années 70[73], les détecteurs de type MCT (Mercury Cadmium Telluride)[16] ont été appliqués pour la première fois à l'imagerie spectrale biomédicale au début des années 2000[21,49,51]. Grâce à ce type de détecteur, des cartographies spectrales précises de grandes zones tissulaires hétérogènes ont pu être acquises, avec une taille de pixel d'environ 5x5 μm^2 pour un temps d'acquisition d'environ

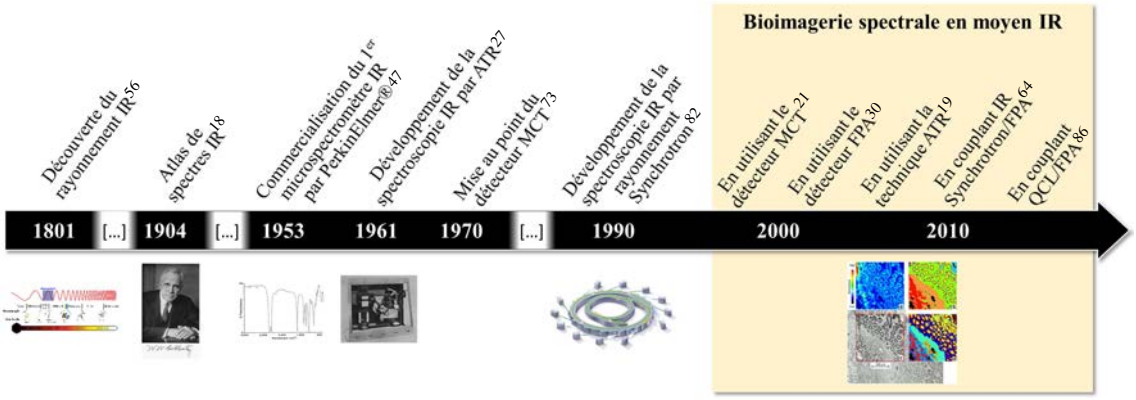


FIGURE I.1 – Différentes étapes du développement de l'instrumentation en spectroscopie moyen IR . ATR, Attenuated Total Reflection. MCT, Mercury Cadmium Telluride. FPA, Focal Plane Arrays. QCL, Quantum Cascade Lasers. Sources des images utilisés (respectivement de gauche à droite) : herschel.cf.ac.uk, www.nasonline.org,[47],[27], synchrotron.org.au et[3].

0,7s/pixel. Récemment, d'autres types de technologies ont été développées tel que les détecteurs FPA (Focal Plane Array)[30,38,52], l'ATR (Attenuated Total Reflectance)[14,19,70], les QCL (Quantum Cascade Lasers)[41,86] et les sources IR Synchrotron[57,64] afin d'améliorer différents paramètres comme les résolutions spectrale et spatiale, le temps d'acquisition et le rapport signal sur bruit des spectres.

Les images spectrales obtenues sont constituées d'un nombre très important de spectres (de quelques dizaines de milliers à plusieurs millions de spectres par mm^2), hautement multi-dimensionnels (typiquement 451 nombre d'ondes dans la gamme spectrale allant de 900 à 1800cm^{-1}, pour une résolution spectrale de 4cm^{-1}). De plus, il est à noter que l'information biologique d'intérêt est "noyée" au sein de ces données. L'extraction optimale de cette information biologique exige donc le développement et l'application de méthodes chimiométriques spécifiques à ces données spectrales IR.

Les principales étapes de ces traitements chimiométriques sont les suivantes :

i) prétraitement des spectres, dans le but d'éliminer les effets parasites tout en conservant les informations spectrales biomoléculaires ;

ii) extraction et/ou sélection de variables (non-supervisée ou supervisée), afin de réduire la dimension des données spectrales et ainsi la complexité des calculs, et d'identifier des marqueurs spectraux spécifiques du processus biologique étudié ;

iii) classification non-supervisée (ou clustering), pour explorer la structure des données spectrales en les partitionnant en groupes homogènes ;

iv) classification supervisée (ou algorithme d'apprentissage), permettant de construire un modèle capable de prédire automatiquement le statut d'un spectre.

Appliquée à des coupes tissulaires, l'association de l'analyse numérique des données et de l'imagerie spectrale IR a permis l'émergence de l'histologie spectrale IR[71,76]. La Figure I.2 résume les principales méthodes numériques utilisées pour traiter les données spectrales et parvenir à ce nouveau type d'histologie. Cependant, les méthodes de classification non-supervisée, couramment utilisées comme KM (*k*-Means) et FCM (Fuzzy C-Means)[21,42,49,61], présentent certaines limites. En effet, leur nombre de classes est choisi de façon empirique. De plus, ce sont des méthodes de recherche locale qui dépendent de leur initialisation[32,37].

Par ailleurs, le prétraitement des spectres et leur classification non-supervisée sont appliqués image par image, avec pour conséquence une correction des spectres image par image et une attribution des couleurs aux clusters qui s'avère différente entre les images[53,61]. Ces traitements individuels compliquent donc très fortement, pour l'expérimentateur et le pathologiste, l'interprétation des résultats d'histologie spectrale.

L'objectif général de notre thèse a été de proposer des solutions à ces limites de l'histologie spectrale IR non-supervisée. L'efficacité de nos approches a été éprouvée sur des images spectrales IR provenant d'un tissu biologique complexe, le côlon, à la fois chez la Souris et chez l'Homme.

Dans une première partie, nous présentons le développement d'une méthode d'histologie spectrale automatisée. La double application hiérarchique d'indices de validité a permis de déterminer le nombre optimal de classes nécessaire à une caractérisation complète des structures histologiques. Ces travaux ont fait l'objet d'une publication dans Analyst en 2015.

Dans une deuxième partie, nous avons étendu la méthode précédente à l'échelle inter-individuelle, chez la Souris et chez l'Homme, par couplage d'un prétraitement par EMSC (Extended Multiplicative Signal Correction) et d'une classification non-supervisée KM ; ce couplage étant appliqué conjointement à toutes les images spectrales IR. Ce travail fait l'objet d'un article soumis dans Journal of Biophotonics.

Dans une dernière partie, face à l'essor des métaheuristiques et à leur capacité à résoudre

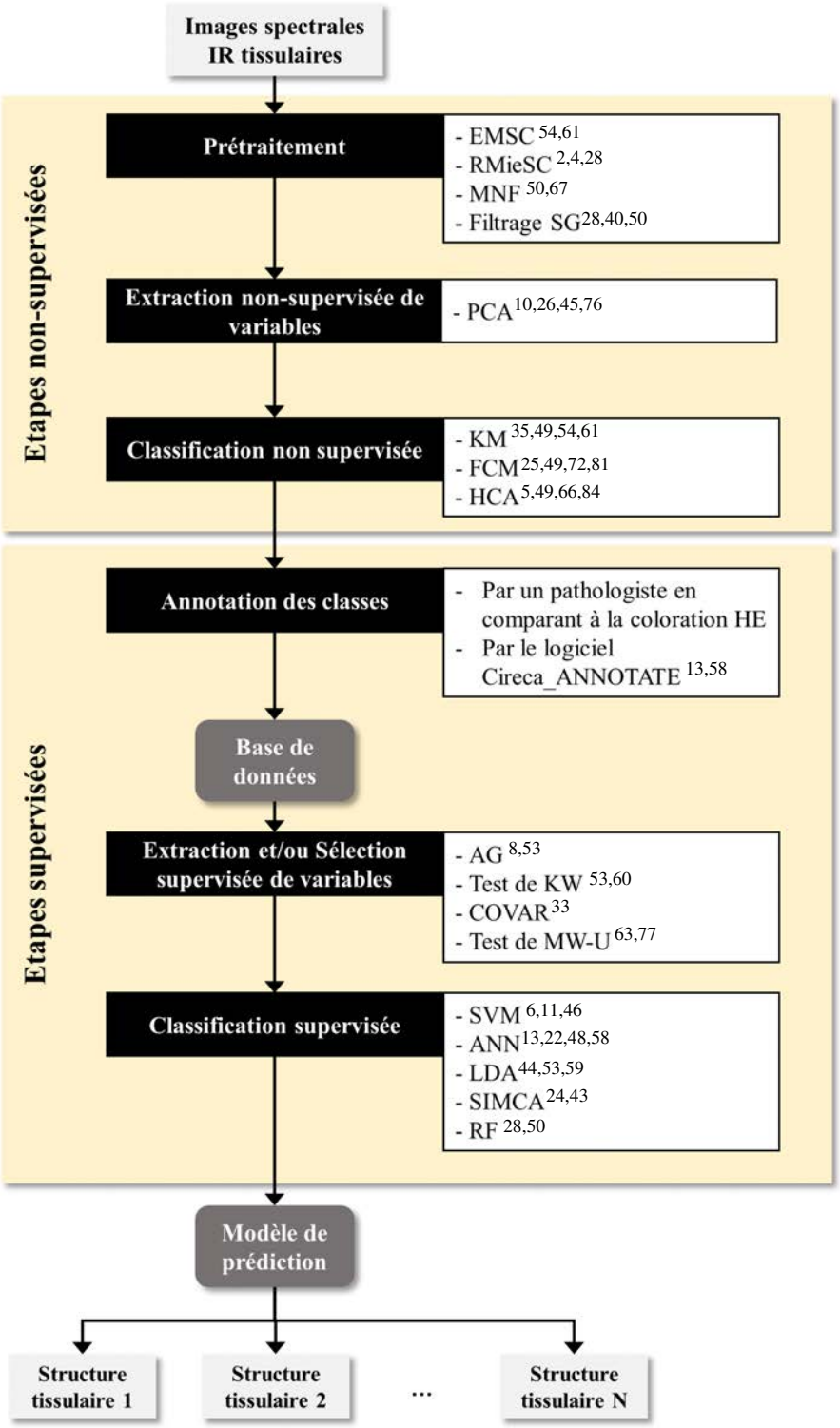


FIGURE I.2 – Diagramme des principales étapes du traitement numérique des images spectrales IR tissulaires. EMSC, Extended Multiplicative Signal Correction. RMieSC, Resonant Mie Scattering Correction. MNF, Minimum Noise Fraction. SG, Savitzky-Golay. PCA, Principal Component Analysis. KM, *k*-Means. FCM, Fuzzy C-Means. HCA, Hierarchical Cluster Analysis. AG, Algorithme Génétique. MW-U, Mann–Whitney-U. SVM, Support Vector Machines. ANN, Artificial Neural Network. LDA, Linear Discriminant Analysis. SIMCA, Soft Independent Modelling by Class Analogy. RF, Random Forest.

des problèmes complexes d'optimisation numérique, nous avons transposé un algorithme mémétique aux données spectrales IR. Cet algorithme mémétique se compose d'un algorithme génétique et d'un raffinement par classification non-supervisée KM. Comparé aux méthodes classiques de clustering, cet algorithme mémétique appliqué aux images spectrales IR, a permis de réaliser une classification non-supervisée optimale et indépendante de l'initialisation. Ces résultats font l'objet d'un article soumis à Analyst.

Chapitre II

Approche méthodologique

Sommaire

II.1 - Histologie conventionnelle du côlon normal 34
II.1.1 - Côlon humain 34
II.1.2 - Côlon murin 37
II.2 - Échantillons biologiques 39
II.2.1 - Préparation des échantillons 39
II.2.2 - Echantillons utilisés 39
II.3 - Histologie spectrale IR 40
II.3.1 - Acquisition des images spectrales IR 40
II.3.2 - Traitements numériques multivariés des images spectrales . . . 44
II.3.3 - Affectation des classes aux structures histologiques 48

II.1 - Histologie conventionnelle du côlon normal

Le côlon des mammifères s'étend du cæcum jusqu'au rectum et constitue la partie terminale de l'intestin, appartenant à l'appareil digestif. Les fonctions principales du côlon sont la déshydratation du bol alimentaire et l'évacuation des déchets de la digestion.

II.1.1 - Côlon humain

Du point de vue anatomique, le côlon humain mesure environ 1m de longueur pour un diamètre de 8cm et peut être divisé en plusieurs segments anatomiques (Figure II.1) : côlon ascendant, côlon transverse, côlon descendant et côlon sigmoïde. La notion de côlon droit inclut successivement l'appendice, le cæcum, le côlon ascendant et la première moitié droite du transverse ; et la notion de côlon gauche, la deuxième moitié gauche du côlon transverse, et les côlons descendant et sigmoïde[78].

Sur le plan histologique, la paroi du côlon comprend quatre couches tissulaires : la muqueuse, la sous-muqueuse, la musculeuse et la séreuse (Figure II.2).

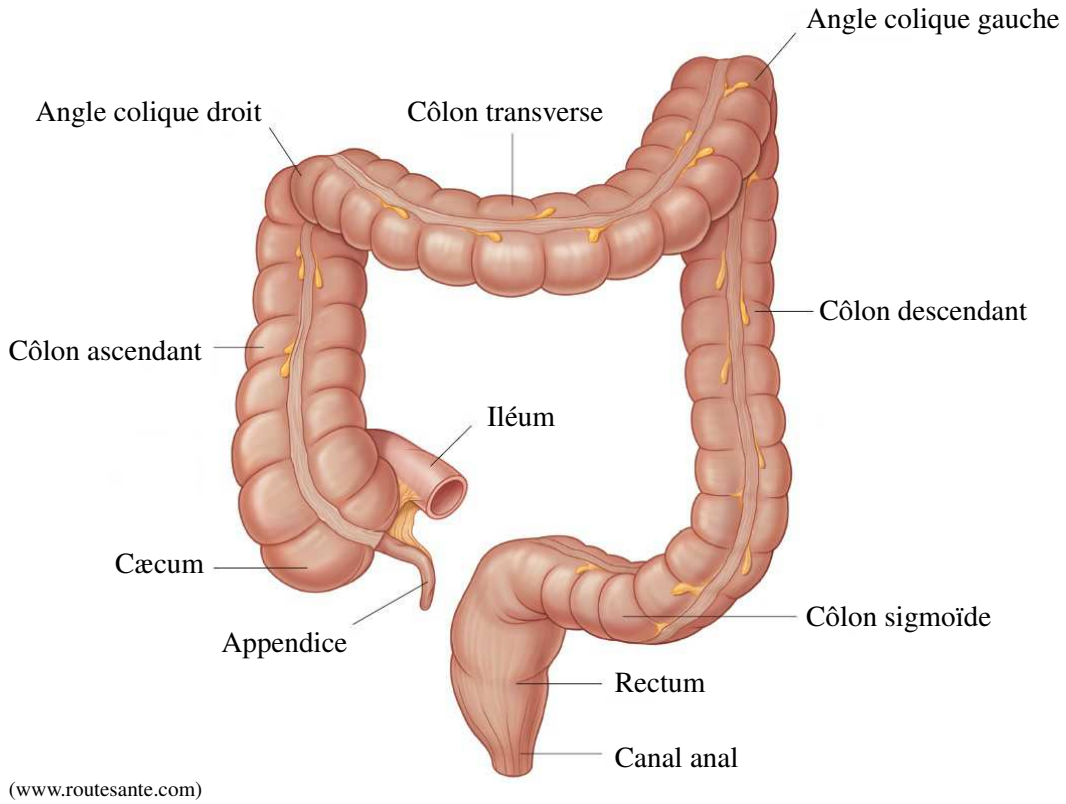
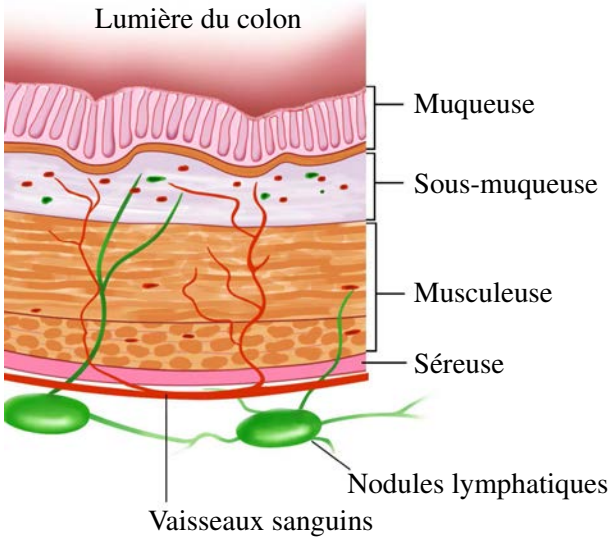


FIGURE II.1 – Anatomie du côlon humain.



(suncoastsurgicalassociates.com)

FIGURE II.2 – Structure de la paroi du côlon humain.

II.1.1.a. La muqueuse

Contrairement à l'intestin grêle, la muqueuse du côlon est dépourvue de plis et de villosités. Elle est constituée d'un épithélium bordant la lumière intestinale, d'un chorion, et de la couche musculaire muqueuse (Figure II.3).

L'épithélium : La surface de la muqueuse est constituée d'un épithélium cylindrique simple qui présente de nombreuses invaginations appelées cryptes de Lieberkühn. Ces glandes tubulaires rectilignes sont orientées perpendiculairement à l'axe du côlon et sont connues comme les unités fonctionnelles et structurales de la muqueuse du côlon. Leur revêtement est constitué des types cellulaires suivants :

- Les entérocytes qui sont situés au niveau de la partie apicale des cryptes, et qui possèdent de courtes microvillosités apicales pour former une bordure en brosse ; ils participent au transport de l'eau et des ions.
- Les cellules caliciformes qui représentent le type cellulaire prédominant ; elles sécrètent du mucus permettant ainsi de lubrifier et de protéger la surface de la muqueuse .
- Les cellules entéro-endocrines peu nombreuses qui sécrètent des hormones peptidiques.

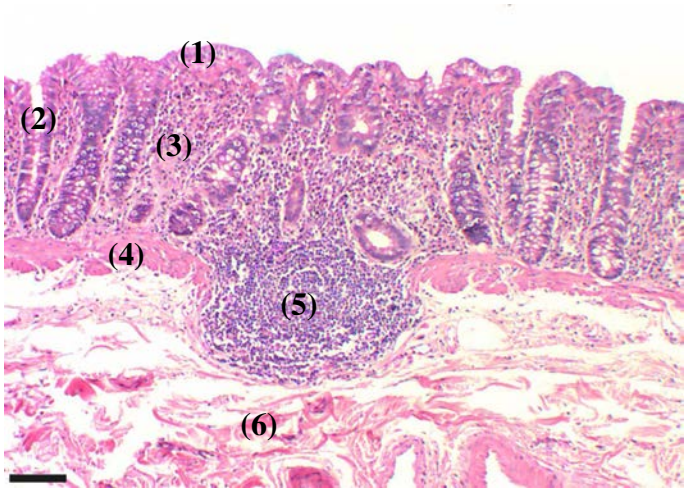


FIGURE II.3 – Structure de la muqueuse du côlon humain après coloration par Hématoxyline-Eosine (HE). La muqueuse comprend différentes structures histologiques : (1) épithélium, (2) cryptes, (3) chorion, (4) musculaire muqueuse, et (5) îlot lymphoïde. (6) Sous-muqueuse. Échelle, 100 μm.

– Les cellules souches qui contribuent au renouvellement de l'épithélium.

Ces différents types cellulaires de l'épithélium sont répartis le long des cryptes, sauf les cellules souches qui se trouvent à la base des cryptes.

Le chorion : Il correspond au tissu conjonctif de soutien sous-jacent de l'épithélium. Il est composé d'une matrice extra-cellulaire et de nombreux types cellulaires : macrophages, mastocytes, neutrophiles, éosinophiles, lymphocytes, fibroblastes, cellules musculaires lisses, fibres nerveuses, vaisseaux sanguins et lymphatiques.

La musculaire muqueuse : Elle est située entre la muqueuse et la sous-muqueuse. C'est une fine couche tissulaire composée de cellules musculaires lisses disposées de façon concentrique dont la fonction est peu connue.

Les îlots lymphoïdes : Ces amas peu nombreux de cellules lymphoïdes s'étendent jusqu'à la sous-muqueuse et jouent un rôle dans l'immunité de la muqueuse.

II.1.1.b. La sous-muqueuse

La sous-muqueuse est principalement constituée de fibroblastes, de macrophages, de mastocytes, de basophiles et d'éosinophiles incorporés dans un tissu conjonctif vascularisé. Les fibres nerveuses du plexus sous-muqueux sont également présentes.

II.1.1.c. La musculeuse

Entourant la sous-muqueuse, la musculeuse comprend une couche musculaire circulaire interne et une couche longitudinale externe composées de cellules musculaires lisses.

II.1.1.d. La séreuse

La séreuse est une couche de tissu conjonctif dense, vascularisée présentant des poches de tissu adipeux. Elle se termine par un mésothélium qui est le feuillet viscéral du péritoine.

II.1.2 - Côlon murin

Chez la Souris, le côlon a une longueur d'environ 14 cm. Il contient quatre parties principales : le cæcum, le côlon proximal, médian et distal (Figure II.4).

Comme chez l'Homme, la structure histologique du côlon murin contient 4 couches : la muqueuse, la sous-muqueuse, la musculeuse et la séreuse (Figure II.5). Au niveau du cæcum et du côlon proximal, la muqueuse présente des plis transversaux. Elle devient plate au niveau du côlon médian et possède des plis longitudinaux dans le côlon distal.

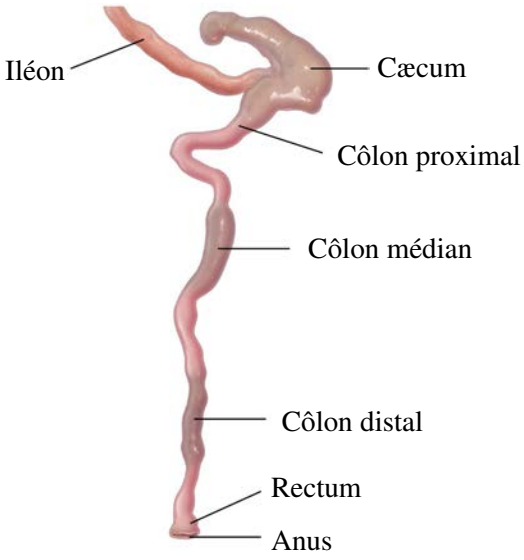


FIGURE II.4 – Anatomie du côlon murin[78]. Il commence par le cæcum en forme de J, et se continue par le côlon proximal, médian et distal.

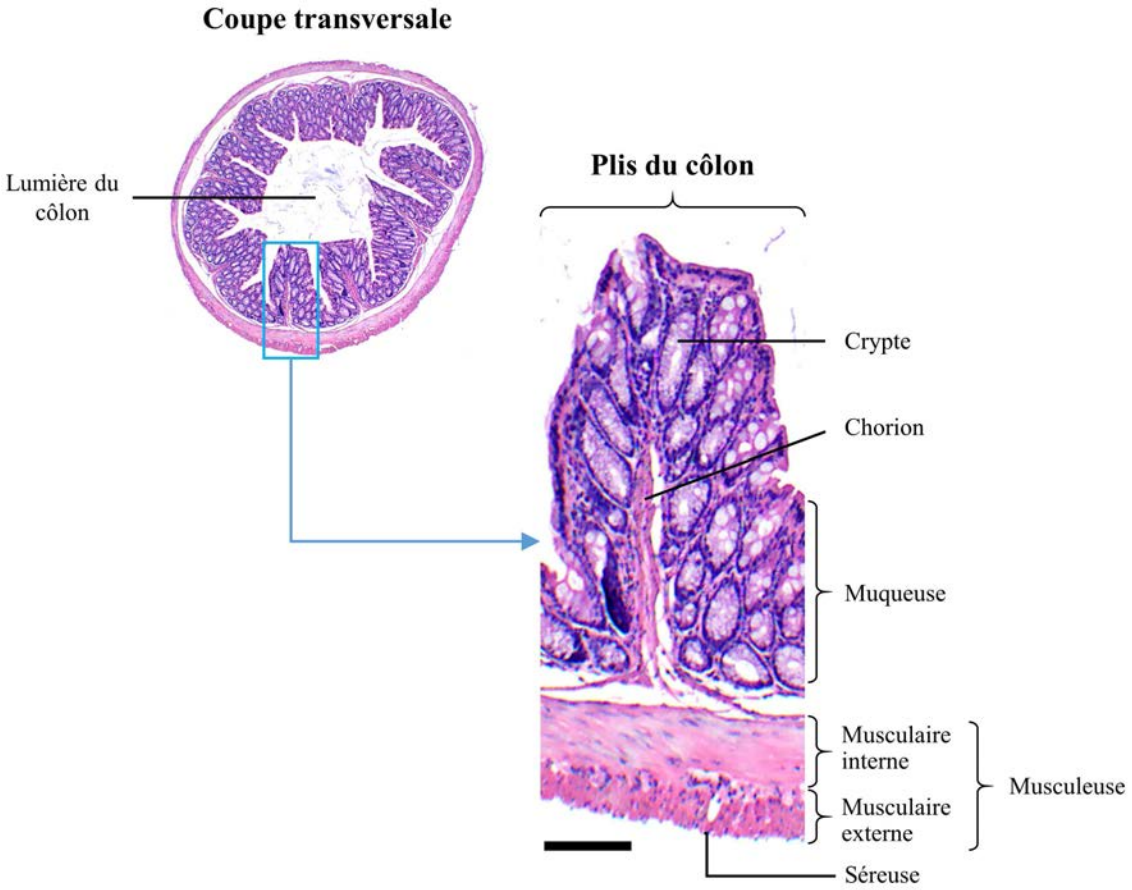


FIGURE II.5 – Structure histologique du côlon murin après coloration par HE. Échelle, 100 μm.

II.2 - Échantillons biologiques

II.2.1 - Préparation des échantillons

Les échantillons utilisés pour ce projet ont été fournis par Dr. D. Guenot (EA 3430, Progression tumorale et microenvironnement, Approches translationnelles et Épidémiologie, Fédération de Médecine Translationnelle de Strasbourg), dans le cadre d'une collaboration au sein du Cancéropôle Grand-Est. Ces échantillons tissulaires ont été fixés au formol et inclus en paraffine. Pour chaque échantillon, deux coupes consécutives de 6μm d'épaisseur ont été préparées et déposées, pour l'une sur une lame de verre (SuperfrostTM Plus) et pour l'autre sur un support en fluorure de calcium (CaF_2) pour l'analyse en imagerie spectrale IR. Pour ce faire, la coupe est déposée sur le support sans utiliser d'agent adhérent et au moyen d'une goutte d'eau distillée. Ce support est ensuite placé sur une plaque chauffante jusqu'à évaporation complète de l'eau et liquéfaction de la paraffine. L'intérêt du support CaF_2 est qu'il n'absorbe pas dans l'IR, et permet ainsi d'effectuer les mesures en mode de transmission afin d'analyser des échantillons tissulaires de faible épaisseur. Il faut souligner que pour la suite des expériences de spectroscopie, cette coupe n'a été ni déparaffinée chimiquement, ni colorée. La coupe sur lame de verre fait l'objet d'une coloration HE, qui permet une identification des structures histologiques par des anatomopathologistes. Les informations seront utilisées pour attribuer les classes spectrales à ces structures.

II.2.2 - Echantillons utilisés

Dans ce projet, toutes les études ont été réalisées sur de la muqueuse saine colique.

Chez l'Homme, les échantillons ont été obtenus à partir des résections chirurgicales des tumeurs primaires. Ces segments de côlon sont composés des tumeurs et de muqueuse saine associée. Les coupes tissulaires ont été réalisées sur des zones identifiées comme saines (Tableau II.1), prélevées à distance (5 cm) de la tumeur.

Pour le modèle murin, les échantillons sont issus de côlons de souris C57BL/6.

TABLE II.1 – Caractéristiques des échantillons coliques étudiés. M, masculin. F, féminin.

#	Age	Sexe	Zone du côlon
1	47	M	Caecum
2	68	M	Côlon droit
3	47	M	Recto-sigmoïde
4	74	F	Sigmoïde
5	73	F	Côlon droit
6	71	M	Sigmoïde
7	76	F	Sigmoïde
8	62	F	Caecum
9	72	F	Recto-sigmoïde
10	76	F	Recto-sigmoïde
11	71	M	Côlon gauche
12	82	F	Sigmoïde
13	75	F	Côlon droit
14	57	M	Recto-sigmoïde
15	74	M	Caecum

II.3 - Histologie spectrale IR

Le protocole d'histologie spectrale suivi dans ce travail comporte trois étapes principales :

- l'acquisition des images spectrales ;
- leur traitement numérique ;
- et l'affectation des classes aux structures histologiques.

II.3.1 - Acquisition des images spectrales IR

II.3.1.a. Principe de la spectroscopie d'absorption infrarouge

La spectroscopie vibrationnelle d'absorption IR permet de caractériser et d'identifier la composition biochimique d'un tissu fixé et paraffiné, sans marquage extrinsèque et sans coloration. Elle est basée sur la mesure des fréquences vibrationnelles spécifiques des molécules lorsqu'elles sont sondées par un rayonnement excitateur.

Une molécule polyatomique de N atomes possède 3N-6 degrés de liberté pour les modes de vibration des liaisons atomiques. En fonction de la fréquence du rayonnement incident,

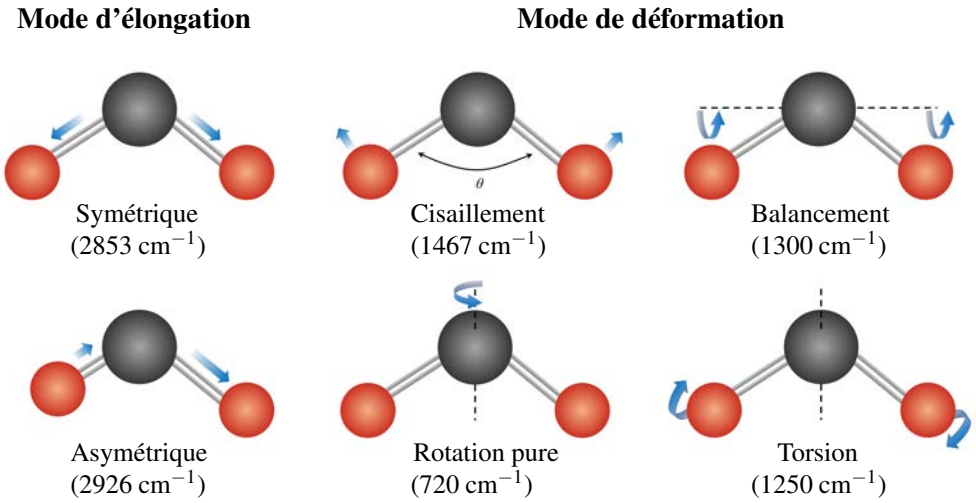


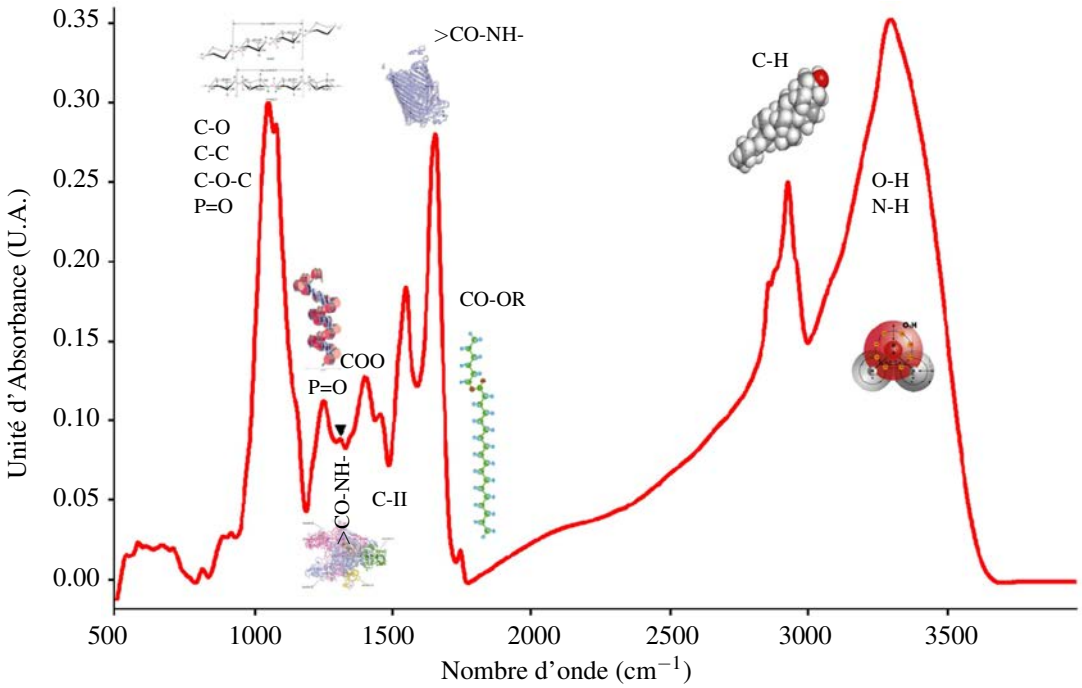
FIGURE II.6 – Exemple de modes de vibration du groupement CH_2. Les modes de vibration peuvent être divisés en deux catégories : les modes d'élongation symétrique et asymétrique qui font varier la longueur des liaisons, et les modes de déformation faisant varier les angles de ces liaisons (cisaillement, balancement, rotation pure et torsion).

ces liaisons vont subir différentes modes de vibration : élongation, déformation d'angle ou déformation hors du plan (Figure II.6).

Théoriquement, la spectroscopie vibrationnelle d'absorption IR consiste à mesurer les variations de l'énergie du rayonnement incident et du rayonnement après l'interaction avec l'échantillon. Elle permet, par la suite, d'obtenir les variations d'intensité de chacun de ces modes de vibration. Ces variations sont donc représentées sous forme d'un spectre, qui est donc considéré comme une empreinte digitale contenant les informations biomoléculaires spécifiques de l'échantillon.

Le domaine du rayonnement IR comporte trois régions : le proche IR (12500 - 4000 cm^{-1}), le moyen IR (4000 - 400 cm^{-1}) et le lointain IR (400 - 10 cm^{-1}). Le moyen IR correspond au domaine de fréquence de vibration des liaisons de la majorité des molécules et représente ainsi la région la plus adaptée à l'analyse des biomolécules.

Lorsque l'énergie du rayonnement lumineux traversant l'échantillon est égale à l'énergie de vibration des liaisons atomiques d'une de ses molécules, une absorption partielle du rayonnement a lieu engendrant une diminution de son intensité. Un spectre IR contient donc des valeurs d'absorbance A (ou transmittance T) en fonction du nombre d'onde σ (exprimé en cm^{-1}), ou de la longueur d'onde $\lambda = \frac{1}{\sigma}$. La Figure II.7 présente les principales bandes d'absorption dans le moyen IR caractéristiques des grandes classes de biomolécules.



(Thèse de D. Sebiskveradze, 2011)

FIGURE II.7 – Bandes d'absorption dans le moyen IR caractéristiques des biomolécules. Acides nucléiques (820 cm^{-1} et 1720 cm^{-1}), carbohydrates (1000-1200 cm^{-1}), protéines (1500-1700 cm^{-1}), lipides (2700-3000 cm^{-1}).

II.3.1.b. Instrumentation

Les images spectrales IR sont enregistrées par un imageur microscopique Micro Imager Spotlight 300 couplé à un spectromètre IR Spectrum One (Perkin Elmer, Courtabœuf, France) (Figure II.8). Ce système utilise une source lumineuse polychromatique de type Globar qui est focalisée par un objectif de type Cassegrain. La partie micro-imageur est équipée d'un détecteur de type MCT de 16 éléments, refroidi à l'azote liquide. Un système de purge est installé afin de minimiser la contribution de la vapeur d'eau du CO_2.

Pour acquérir une image spectrale IR, la source lumineuse traverse les différent points de l'échantillon selon un balayage en 2 dimensions. Chaque image est enregistrée avec une taille de pixel de 6,25 x 6,25 μ m^2. A chaque pixel, un spectre d'absorption IR de gamme spectrale 400-4000 cm^{-1} a été enregistré. Chaque spectre a une résolution spectrale de 4 cm^{-1} et une accumulation spectrale de 16 scans. En mode de transmission, un spectre référence a été enregistré au préalable à chaque acquisition sur une zone propre du support CaF_2, avec une résolution spectrale de 4 cm^{-1} et 240 accumulations.

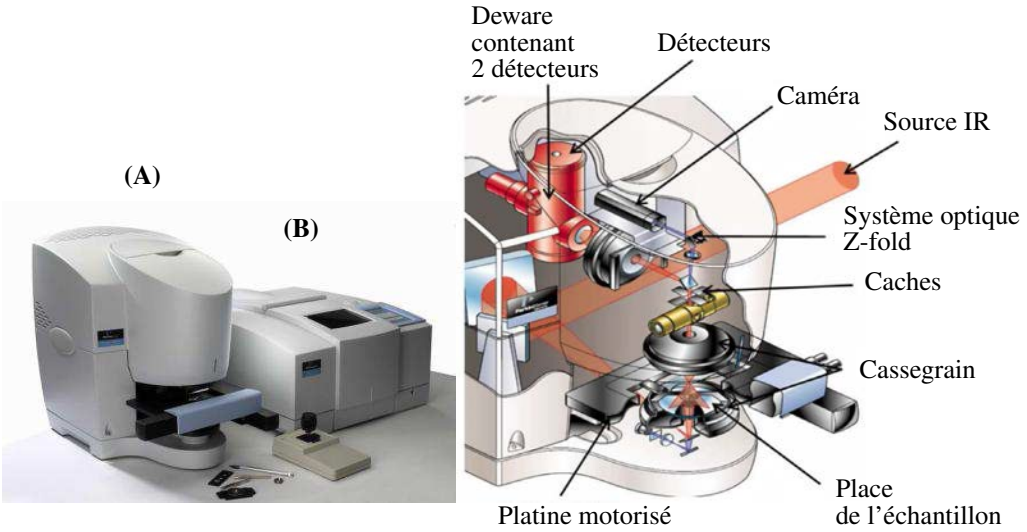


FIGURE II.8 – Micro-imageur spectral IR. Le système est composé (A) d'un système d'imagerie (Micro Imager Spotlight 300, Perkin Elmer) et (B) d'un spectromètre IRTF (Spectrum One, Perkin Elmer).

Une image spectrale, acquise via le logiciel SpectrumImage (Perkin Elmer, Courtabœuf, France), représente donc une image tridimensionnelle ou un cube de données (Figure II.9). Les deux premières dimensions X et Y sont les coordonnées géographiques de chaque spectre. La troisième dimension Λ correspond aux nombres d'onde de la gamme spectrale moyen IR (c'est-à-dire, chaque pixel est représenté par un spectre IR).

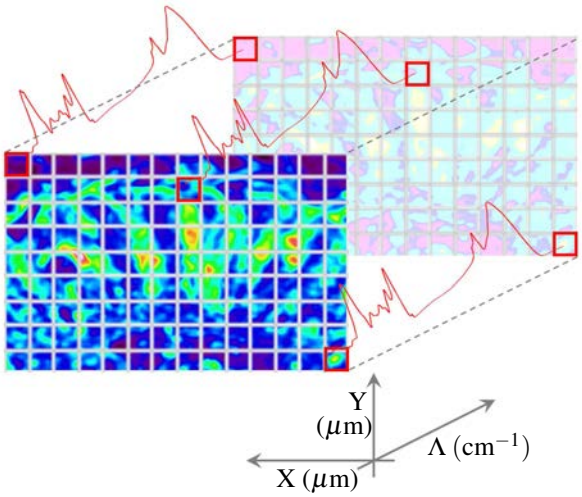


FIGURE II.9 – Exemple d'une image spectrale IR. Cette image équivaut à un cube de données, où X et Y représentent les coordonnées spatiales de chaque spectre, et Λ les différents nombres d'onde de la gamme spectrale IR.

II.3.2 - Traitements numériques multivariés des images spectrales

Les différentes étapes de traitement ont été mises en œuvre via des programmes écrits en langage Matlab (The Mathworks, Natick, MA, USA), sauf indication contraire.

II.3.2.a. Prétraitements

Au sein des spectres bruts, les informations biomoléculaires sont altérées par des signaux parasites liés à des facteurs physiques et environnementaux. Il est donc indispensable de corriger ces spectres en appliquant des méthodes de prétraitements numériques.

Correction atmosphérique

Pour chaque image spectrale, les contributions des contaminants atmosphériques (vapeur d'eau et CO_2) sont corrigées par soustraction du spectre de référence enregistré sur le support de CaF_2, grâce au logiciel SpectrumImage.

Sélection de la gamme spectrale

La gamme spectrale de 900 à 1800 cm^{-1} a été sélectionnée car elle est considérée comme la plus informative pour les échantillons biologiques[79]. Elle est nommée région de l'empreinte digitale car elle contient les bandes spectrales spécifiques des principales classes de biomolécules constituant les échantillons tissulaires.

EMSC

Les spectres acquis sur les coupes tissulaires paraffinées sont pollués par la diffusion de la lumière due à l'échantillon et par le signal de la paraffine. Ces effets se traduisent par des bandes spectrales intenses et une ligne de base diffuse (Figure II.10) qui diffèrent d'un spectre à un autre.

La diffusion peut être corrigée par EMSC[1] ou par dérivée du premier ordre[9,49], et le signal de paraffine par déparaffinage chimique[2,62,76], ou par l'exclusion des régions spectrales[7].

Dans ce travail, ces deux effets ont été corrigés simultanément par l'EMSC selon le protocole que notre laboratoire a récemment décrit[83] pour l'étude du tissu de côlon normal et cancéreux.

Considérons une image spectrale $\mathbf{S} \in \mathbb{R}^{N \times D}$ composée de N spectres. Chaque spectre $\mathbf{s_i} \in \mathbb{R}^{1 \times D}$, $1 \leq i \leq N$, contient D valeurs d'absorbance mesurées aux D nombres d'onde $W = \{w_1, \ldots, w_D\}$. L'EMSC modélise linéairement chaque spectre $\mathbf{s_i}$ par :

$$\mathbf{s_i} = a_i \hat{\mathbf{s}} + \mathbf{b_i I} + \mathbf{c_i P} + \mathbf{e_i} \tag{II.1}$$

où :

- $\hat{\mathbf{s}} \in \mathbb{R}^{1 \times D}$ est un spectre de référence de l'image spectrale analysée. Généralement, $\hat{\mathbf{s}}$ est choisi comme le spectre moyen de l'image : $\hat{\mathbf{s}} = \frac{1}{N} \sum_{i=1}^{N} \mathbf{s_i}$
- $\mathbf{I} \in \mathbb{R}^{M \times D}$ est une matrice d'interférence composée de M composantes qui modélisent la contribution de la paraffine.
- $$\mathbf{P} = \begin{bmatrix} w_1^0 & \ldots & w_1^p \\ \vdots & \ddots & \vdots \\ w_D^0 & \ldots & w_D^p \end{bmatrix}^T \in \mathbb{R}^{(p+1) \times D}$$

  est la transposée de la matrice de Vandermonde du vecteur W ; cette matrice est utilisée pour calculer $\mathbf{c_i P}$ qui représente une fonction polynomiale d'ordre p modélisant les effets de diffusion de la lumière.

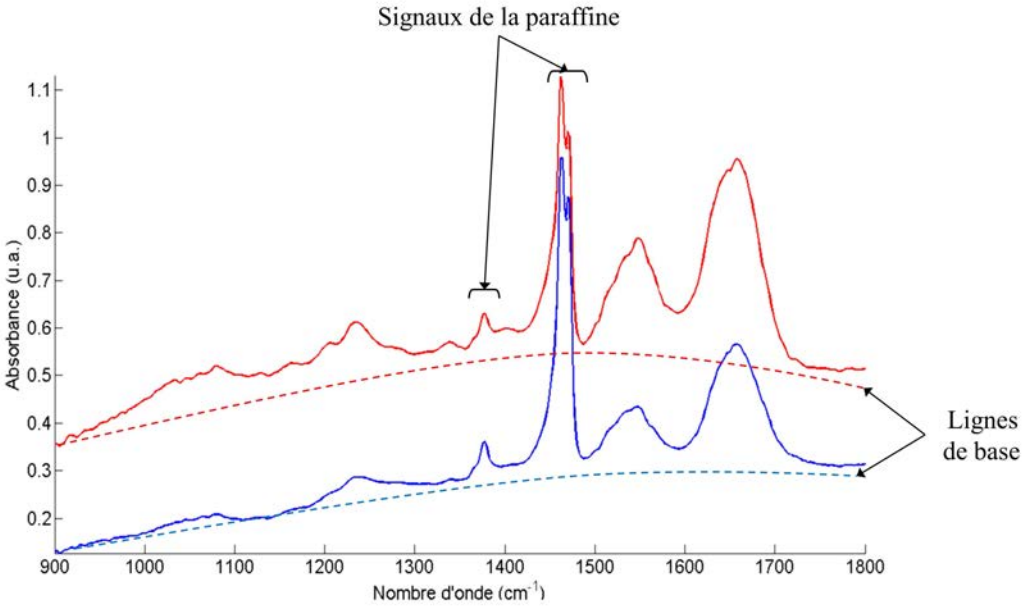


FIGURE II.10 – Exemples de spectres IR d'une image acquise sur une coupe tissulaire paraffinée.

- $\mathbf{e_i} \in \mathbb{R}^{1\times D}$ est le vecteur d'erreur de modélisation.
- a_i est le coefficient de régression de $\hat{\mathbf{s}}$ sur $\mathbf{s_i}$.
- $\mathbf{b_i} \in \mathbb{R}^{1\times M}$ est le vecteur des coefficients de régression de $\mathbf{I}$ sur $\mathbf{s_i}$.
- $\mathbf{c_i} \in \mathbb{R}^{1\times(p+1)}$ est le vecteur des coefficients de régression de $\mathbf{P}$ sur $\mathbf{s_i}$ et représente donc le vecteur des coefficients de la fonction polynomiale d'ordre p.

L'estimation de a_i, $\mathbf{b_i}$ et $\mathbf{c_i}$ est réalisée par la méthode des moindres carrés ordinaires afin de minimiser la somme des erreurs de modélisation $\sum_{j=1}^{D}(e_{ij})^2$.

Le spectre $\mathbf{s_i}$ corrigé de la contribution de la paraffine et de la ligne de base est calculé par :

$$\mathbf{s_i^{corr}} = a_i\hat{\mathbf{s}} + \mathbf{e_i}. \tag{II.2}$$

$\mathbf{s_i^{corr}}$ peut être vu comme une approximation du spectre $\mathbf{s_i}$ par le spectre de référence $\hat{\mathbf{s}}_\mathbf{i}$. L'erreur $\mathbf{e_i}$ est donc la partie informative de l'équation II.2 puisqu'elle représente les différences biomoléculaires entre le spectres moyen $\hat{\mathbf{s}}_\mathbf{i}$ et le spectre analysé $\mathbf{s_i}$.

Ensuite, une normalisation est appliquée sur $\mathbf{s_i^{corr}}$ afin de le corriger d'effets indésirables comme la longueur effective du trajet optique :

$$\bar{\bar{\mathbf{s}}}_\mathbf{i} = \hat{\mathbf{s}} + \frac{\mathbf{e_i}}{a_i} = \frac{\mathbf{s_i^{corr}}}{a_i}. \tag{II.3}$$

La matrice d'interférence $\mathbf{I}$ est construite par le spectre moyen et les 10 premières composantes principales calculées à partir de spectres acquis sur de la paraffine pure. La ligne de base a été modélisée par une fonction polynomiale d'ordre 4.

Après traitement par EMSC, les variabilités de la ligne de base et du signal de la paraffine sont neutralisées et les spectres sont normalisés sur le spectre de référence (Figure II.11). De plus, un seuillage des a_i et des erreurs de modélisation normalisées $\sum_{j=1}^{D}(\frac{e_{ij}}{a_i})^2$ permet de détecter dans l'image spectrale les spectres pures de paraffine et de les éliminer des prochaines analyses numériques. Le nombre de spectres restant est noté R.

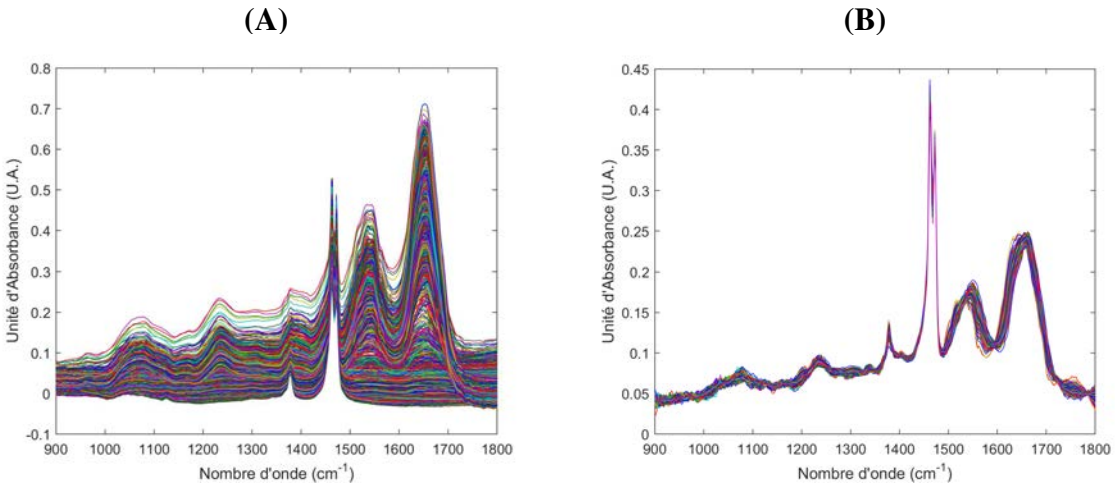


FIGURE II.11 – Effet du prétraitement par EMSC. (A) Spectres IR acquis sur une coupe tissulaire, (B) spectres prétraités par EMSC.

II.3.2.b. Classification non-supervisée : le clustering *k*-Means

Après correction par EMSC, les spectres sont soumis à la méthode de classification non-supervisée KM[55] afin de mettre en évidence les structures histologiques présentes dans l'échantillon analysé.

KM est une méthode de clustering permettant de partitionner un jeu de données $\bar{\bar{\mathbf{S}}} = \{\bar{\bar{\mathbf{s}}}_\mathbf{i}, 1 \leq i \leq R\}$ composé de R spectres de D dimensions en k classes (ou clusters), en minimisant la variation intra-cluster définie par :

$$F = \sum_{j=1}^{k} \sum_{i=1}^{R_j} \|\mathbf{c_j} - \bar{\bar{\mathbf{s}}}_\mathbf{i}\|^2 \tag{II.4}$$

où :

- R_j est le nombre de spectres appartenant à la $j^{\text{ème}}$ classe ;
- $\mathbf{c_j}$ est le barycentre de la $j^{\text{ème}}$ classe, aussi appelé centroïde ;
- $\|.\|$ représente la distance Euclidienne.

Les clusters sont itérativement estimés par l'algorithme suivant :

Étape 1 : Initialisation des k centroïdes $\mathbf{c_j}$ en choisisant aléatoirement k spectres du jeu de données $\bar{\bar{\mathbf{S}}}$.

Étape 2 : Attribution de chaque spectre $\bar{\bar{\mathbf{s}}}_\mathbf{i}$ à la classe j dont la distance Euclidienne $\|\mathbf{c_j} - \bar{\bar{\mathbf{s}}}_\mathbf{i}\|$ est la plus petite.

Étape 3 : Mise à jour de chaque centroïde $\mathbf{c_j}$ en moyennant les spectres appartenant à la $j^{\text{ème}}$ classe.

Étape 4 : Répétition des étapes 2 et 3 jusqu'à ce que plus aucun spectre ne change de classe.

Pour limiter la dépendance de KM à l'étape d'initialisation, KM a été relancé 20 fois sur chaque image spectrale. Le résultat KM générant la variation intra-cluster F la plus petite est retenu.

II.3.3 - Affectation des classes aux structures histologiques

Le KM permet d'estimer une partition composée de k classes auxquelles l'algorithme attribue aléatoirement des couleurs. La Figure II.12A présente l'exemple d'une image en pseudo-couleur d'un tissu de côlon comportant 13 classes. A l'aide de la coloration HE de la coupe adjacente Figure II.12B, il est possible d'attribuer spécifiquement chaque couleur à une structure histologique donnée (Figure II.12C). Les pixels blancs correspondent aux spectres identifiés comme acquis sur la paraffine pure et éliminés par EMSC.

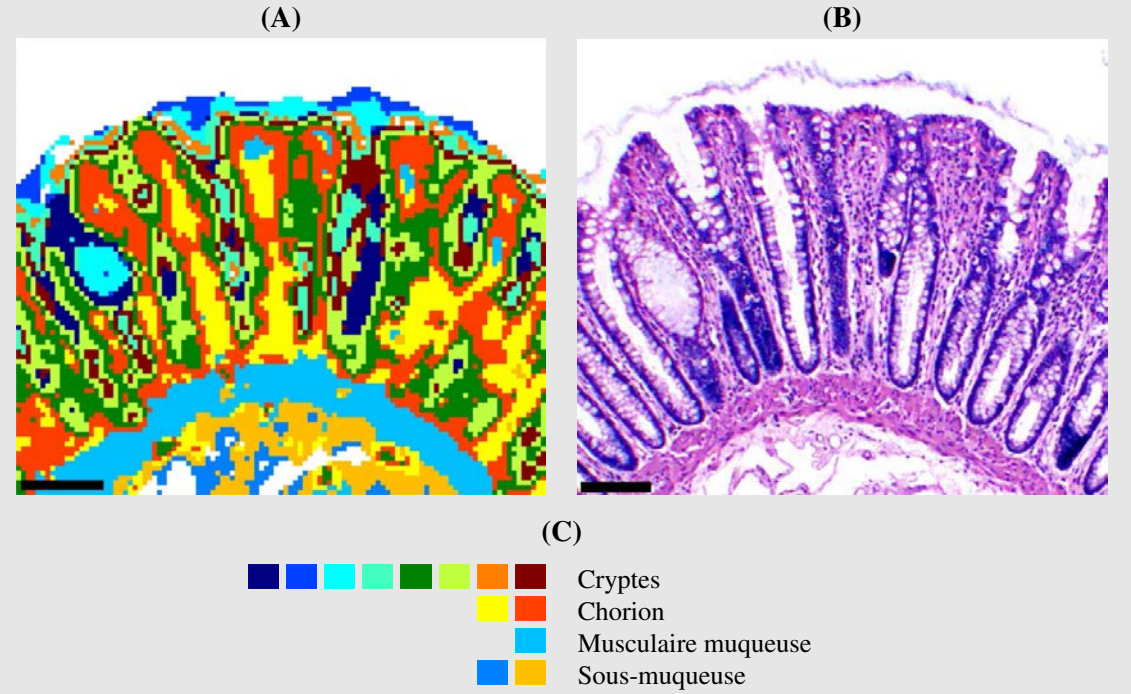


FIGURE II.12 – Attribution des classes d'une image pseudo-couleur aux structures histologiques du tissu. (A) Image pseudo-couleur avec $k = 13$ classes, (B) coloration HE et (C) identification des structures histologiques. Échelles, 100 μm.

Chapitre III

Développement d'une histologie spectrale automatisée par indices de validité

Sommaire

III.1 - Préambule 52

III.2 - Article #1 : *"Development of a hierarchical double application of crisp cluster validity indices : a proof-of-concept study for automated FTIR spectral histology"* 56

III.3 - Résultats supplémentaires 67

III.3.1 - Application des indices de validité sur un jeu de données hiérarchiques multi-dimensionnelles 67

III.3.2 - Application des 28 indices de validité supplémentaires sur les images spectrales 70

III.3.3 - Confirmation des résultats sur 10 nouveaux patients 73

III.1 - Préambule

Cette partie du travail a fait l'objet de l'article suivant :
T.N.Q. Nguyen, P. Jeannesson, A. Groh, D. Guenot, C. Gobinet.*"Development of a hierarchical double application of crisp cluster validity indices : a proof-of-concept study for automated FTIR spectral histology"*, *Analyst*, 140, pp. 2439-2448, **2015**

Contexte et objectif du projet

En biophotonique, des études récentes ont montré que la combinaison de l'imagerie spectrale IR avec des méthodes de classification non-supervisée, comme KM, permet de réaliser une histologie spectrale des tissus humains. Grâce à cette application, les structures histologiques présentes dans une image spectrale IR sont différenciées. Elle rend la comparaison entre l'histologie spectrale et conventionnelle possible.

Cependant, cette méthode dépend d'un nombre de classes k qui doit être fixé par l'utilisateur. Soit k est choisi empiriquement suivant une connaissance *a priori* de la structure des données étudiées. Soit il est déterminé par l'approche par essais et erreurs (trial-and-error) en incrémentant le nombre de classes k jusqu'à l'obtention d'une partition histologiquement cohérente. Le choix du nombre de classes k devient donc extrêmement difficile lorsque des données multidimensionnelles sont acquises sur un échantillon complexe, tel qu'un échantillon tissulaire.

Il est possible d'automatiser le choix du nombre de classes k en utilisant les indices de validité. De nombreux indices de validité ont été développés pour le clustering KM, mais, à notre connaissance, ils n'ont jamais été appliqués à l'imagerie spectrale IR.

Dans ce travail, l'objectif a été d'automatiser la classification KM par indices de validité afin de réaliser une histologie spectrale objective et automatisée.

Indices de validité

Un indice de validité est une fonction mathématique permettant de mesurer la qualité d'une partition (estimée par un algorithme de classification non-supervisée, tel que KM)

en calculant le rapport entre :

i) la compacité des classes calculée à partir des distances entre les points appartenant à une même classe ;

ii) et la séparation entre les classes calculée à partir des distances entre des points appartenant à des classes différentes.

Appliqué sur des partitions estimées sur un même jeu de données pour un nombre de classes k variant de 2 à k_{max}, un indice de validité permet de trouver le nombre optimal de classes k_{opt} générant la partition présentant les classes les plus compactes et les plus séparées.

Dans la littérature, de nombreux indices de validité ont été décrits. Dans ce projet, 37 indices ont été testés. Cependant, seuls les 9 indices les plus représentatifs sont présentés dans l'article :

- les 2 indices donnant les meilleurs résultats sur nos données : PBM (Pakhira-Bandyopadhyay-Maulik)[65], SI (Sym-Index)[69] ;
- les 4 indices les plus populaires : Dunn[23], DB (Davies-Bouldin)[20], SWC (Silhouette-Width-Criterion)[68] et XB (Xie-Beni)[85] ;
- les 3 indices les plus récents : COP (Context-independent Optimality and Partiality)[29], SV (Separation-Variance)[87] et OS (Overlap-Separation)[87]

Les 28 indices restants sont mentionnés dans les résultats supplémentaires de ce chapitre.

Résultats

Les indices de validité ont été appliqués sur les partitions estimées par KM ($2 \leqslant k \leqslant 20$) sur les images spectrales acquises sur des coupes tissulaires de côlon normal de patients. Les résultats montrent que les indices présentent deux comportements extrêmes (Tableau 3 de l'article, et Tableaux III.3 et III.4 des résultats supplémentaires). Quelques indices surestiment très largement le nombre de classes ($k_{opt} \approx 20$), rendant difficile l'attribution des clusters aux structures histologiques. La majorité des indices sous-estiment le nombre

de classes ($k_{opt} \approx 4$), induisant un mélange de plusieurs structures histologiques dans un même cluster.

Pour résoudre ce problème de sous-estimation, nous avons développé une double application hiérarchique d'indices de validité selon les étapes suivantes :

i) L'indice de validité est appliqué sur les partitions estimées par KM sur une image spectrale, pour $2 \leqslant k \leqslant 20$. Le nombre optimal de classes estimé par l'indice de validité représente le nombre de classes principales k_{main} composant le jeu de données.

ii) Pour chaque classe principale, KM est appliqué sur les spectres appartenant à cette classe pour $2 \leqslant k \leqslant 20$. L'indice de validité est appliqué une seconde fois sur ces nouvelles sous-partitions KM. Le nombre optimal de sous-clusters composant cette classe est ainsi estimé.

iii) Le nombre optimal de clusters k_{opt} est calculé par la somme des nombres optimaux de sous-clusters, et la partition optimale est obtenue en assemblant les k_{main} sous-partitions optimales.

L'objectif de cette approche étant d'estimer le nombre optimal de classes de jeux de données présentant une structure hiérarchique, nous l'avons tout d'abord testée sur deux jeux de données synthétiques. Le premier, bidimensionnel et présenté dans l'article, a été généré à des fins pédagogiques. Le deuxième, composé de 451 dimensions et présenté dans les résultats supplémentaires, a pour but de simuler des données présentant les caractéristiques des spectres IR.

Appliqué sur ces jeux de données synthétiques, notre méthodologie est capable d'accéder à toutes les sous-structures des classes, contrairement à l'application simple des indices de validité (Tableau 2 de l'article et Tableau III.1 des résultats supplémentaires).

Appliqués sur des images spectrales IR des cinq patients présentées dans l'article, seules les doubles applications hiérarchiques de PBM et SI sont capables de retrouver toutes les structures et sous-structures histologiques du côlon normal, à savoir les cryptes, le chorion, la musculaire muqueuse, la sous-muqueuse, les îlots lymphoïdes et le mucus secrété (Figure 7 et Tableau 4 de l'article, et Tableau III.3 des résultats supplémentaires).

Étendus aux dix patients supplémentaires présentés dans le paragraphe II.2.2 - , ces ré-

sultats confirment l'efficacité de la double application de PBM. Par contre, la double application de SI surestime le nombre de classes nécessaires pour décrire complètement les structures histologiques. L'ensemble de ces résultats est présenté dans le Tableau III.5.

III.2 - Article #1 : *''Development of a hierarchical double application of crisp cluster validity indices : a proof-of-concept study for automated FTIR spectral histology''*

T.N.Q. Nguyen, P. Jeannesson, A. Groh, D. Guenot, C. Gobinet.

Analyst, **2015**, 140, pp. 2439-2448

Analyst

PAPER

Cite this: *Analyst*, 2015, **140**, 2439

Development of a hierarchical double application of crisp cluster validity indices: a proof-of-concept study for automated FTIR spectral histology†

Thi Nguyet Que Nguyen,[a,b] Pierre Jeannesson,[a,b] Audrey Groh,[c] Dominique Guenot[c] and Cyril Gobinet*[a,b]

Received 23rd October 2014,
Accepted 15th January 2015

DOI: 10.1039/c4an01937g

www.rsc.org/analyst

Fourier-transform infrared (FTIR) spectral imaging is currently used as a non-destructive and label-free method for analyzing biological specimens. However, to highlight the different tissue regions, unsupervised clustering methods are commonly used leading to a subjective choice of the number of clusters. Here, we develop a hierarchical double application of 9 selected crisp cluster validity indices (CCVIs) using K-Means clustering. This approach when tested first on an artificial dataset showed that the indices Pakhira–Bandyopadhyay–Maulik (PBM) and Sym-Index (SI) perfectly estimated the expected 9 sub-clusters. Then, the concept was applied to a real dataset consisting of FTIR spectral images of normal human colon tissue samples originating from 5 patients. PBM and SI were revealed to be the most efficient indices that correctly identified the different colon histological components including crypts, lamina propria, muscularis mucosae, submucosa, and lymphoid aggregates. In conclusion, these results strongly suggest that the hierarchical double CCVI application is a promising method for automated and informative spectral histology.

1. Introduction

Fourier-transform infrared (FTIR) microspectroscopy is a vibrational spectroscopy based on the absorption, by a sample, of an incident infrared (IR) light. This technique permits to acquire spectra from IR active molecules, giving global molecular information about the sample, in a non-destructive and non-invasive manner without requiring any staining or labeling. When applied to biological samples, this technology may reveal structural and metabolic changes at the tissular and even at the cellular level.[1] Therefore, FTIR microspectroscopy actually appears as a helpful tool for the diagnosis of disease states in oncology.[2]

In the 80s, FTIR microspectroscopy, which is a point-wise technology, was extended to FTIR imaging leading to two-dimensional scans of tissue sections. Data cubes, named FTIR images, are thus automatically and rapidly acquired with two spatial dimensions and one spectral dimension. Each pixel of a FTIR image contains a full IR spectrum specific to the tissue molecular composition at that point. A data cube can be composed of several thousand pixels, and each spectrum can be recorded in several hundred wavelengths.

Analyzing this data cube is complex due to its high volume and dimensionality. Furthermore, the studied biomolecular changes induce usually a weak and subtle spectral response. An objective and rapid analysis of such data requires the development of sophisticated chemometrical and computational tools. For spectral histology, FTIR imaging is usually associated with clustering methods such as K-Means (KM),[3] Hierarchical Clustering Analysis (HCA)[4] or Fuzzy C-Means (FCM)[5] to reveal the different tissue structures based on their different biomolecular spectral signatures.[6]

As these methods are unsupervised, they depend on the number of clusters k that must be fixed by the user. Usually, the choice of k is driven by *a priori* knowledge of the structure of the studied dataset. A trial-and-error procedure can also be used by iteratively increasing the number of clusters until obtaining a coherent partition of the studied phenomenon. However, this subjective adjustment of k parameter is extremely difficult when multidimensional datasets are acquired on complex samples.

[a]*Université de Reims Champagne-Ardenne, Equipe MéDIAN-Biophotonique et Technologies pour la Santé, UFR de Pharmacie, 51 rue Cognacq-Jay, 51096 Reims, France. E-mail: cyril.gobinet@univ-reims.fr*

[b]*CNRS UMR 7369, Matrice Extracellulaire et Dynamique Cellulaire (MEDyC), Reims, France*

[c]*Université de Strasbourg (UdS), EA 3430 Progression tumorale et microenvironnement. Approches translationnelles et Epidémiologie. Fédération de Médecine Translationnelle de Strasbourg (FMTS), Bâtiment U1113, 3 Avenue Molière, 67200 Strasbourg, France*

†This article includes work which was presented at SPEC 2014 Shedding New Light on Disease held in Krakow, Poland, August 17–22, 2014.

To automatically estimate the optimal k, validity indices[7–9] have been developed for crisp (such as KM or HCA) or fuzzy (such as FCM) clustering methods. Fuzzy validity indices have already been applied to FTIR imaging for the automatic spectral histology of axillary lymph nodes[10] and skin cancers.[11] However, due to the failure of these indices, new methods have been developed based on a cluster merging scheme[10] or on an inter-cluster redundancy measure.[11] To our knowledge, crisp validity indices have never been applied in FTIR imaging for automated spectral histology.

In this study, we evaluated the performance of nine well-known crisp cluster validity indices (CCVIs) on KM partitions estimated on an artificial dataset and FTIR images acquired on healthy human colon tissue sections. We demonstrated that a hierarchical double application of CCVIs is necessary to match the main tissue structures encountered in the colon tissue. Indeed, this approach permitted automatic estimation of clusters specific to crypts, lamina propria, muscularis mucosae, submucosa, and lymphoid aggregates.

2. Methodologies for estimating the number of clusters

First, we describe the well-known KM clustering and nine classically used CCVIs. Then, we propose a new methodology based on the hierarchical double application of CCVIs on KM partitions.

2.1. KM clustering

In data analysis, KM[3] is one of the most popular unsupervised classification methods. It aims to separate a set of N unlabeled points of d dimensions into k clusters. Each cluster is represented by its barycenter, named centroid.

KM algorithm starts by choosing randomly k initial points as centroids. Then, it runs iteratively following these steps:

i. Based on a chosen distance metric, each point is assigned to the cluster whose centroid is the nearest.

ii. The centroids are updated according to the new partition computed in step (i).

Steps (i) and (ii) are repeated until point assignment is stabilized. The KM algorithm thus converges to a local minimum of the objective function defined as the within-cluster sum of point-to-centroid distances. Traditionally, the point-to-centroid distance is computed as the Euclidean distance, such as in our study.

2.2. CCVIs

A CCVI is a mathematical function that measures the quality of a partition. By performing a KM clustering for different values of k, $2 \leq k \leq k_{\max}$, a CCVI value can be computed for each partition. The number of clusters giving the optimal CCVI value is defined as the optimal number of clusters k_{opt} for the considered dataset. In this study, the partition estimated by KM using k_{opt} clusters is called the optimal single CCVI partition.

During the last few decades, most of the existing CCVIs were defined as the combination of compactness and separation measures of clusters, where compactness (or within-cluster distance) indicates the closeness of points in the same cluster, and separation (or between-clusters distance) indicates the isolation of one cluster from another. The optimal partition maximizes the separation between clusters and minimizes the cluster compactness.

Among the numerous CCVIs described in the literature,[7,8] the nine CCVIs used here include: four popular CCVIs such as Dunn,[12] Davies–Bouldin (DB),[13] Silhouette-Width-Criterion (SWC)[14] and Xie–Beni (XB);[15] five recent CCVIs such as Pakhira–Bandyopadhyay–Maulik (PBM),[16] Sym-Index (SI),[17] Context-independent Optimality and Partiality (COP),[18] Separation-Variance (SV)[19] and Overlap-Separation (OS).[19]

Let $\mathbf{X} = \{\mathbf{x}_i,\ i = 1, \ldots, N\}$ be a dataset composed of N points, where each point $\mathbf{x}_i \in \mathbb{R}^D$. After partitioning $\mathbf{X}$ into k clusters, $\mathbf{c}_l$ describes the centroid of the l^{th} cluster, where $l = 1, \ldots, k$, and $\mathbf{X}_l = \{\mathbf{x}_{l(j)}, j = 1, \ldots, N_l\}$ is the set of N_l points belonging to the l^{th} cluster. $||.||$ is the Euclidean distance.

2.2.1. Dunn. This index[12] is defined as the ratio between the minimal inter-cluster distance and the largest cluster diameter:

$$\text{Dunn}(k) = \frac{\min_{p,q \in \{1,\ldots,k\}, p \neq q} \delta_{p,q}}{\max_{l \in \{1,\ldots,k\}} \Delta_l} \tag{1}$$

where $\delta_{p,q}$ is the closest distance between two points across clusters p and q (*i.e.*, the separation between clusters p and q), and Δ_l is the maximum distance between two points in the l^{th} cluster (*i.e.*, the l^{th} cluster compactness).

k_{opt} is defined as the number of clusters which maximizes the Dunn's index.

2.2.2. Davies–Bouldin (DB). This index[13] is defined as the average ratio between the intra- and inter-cluster distances of each cluster with its most similar cluster:

$$\text{DB}(k) = \frac{1}{k}\sum_{l=1}^{k} R_l \tag{2}$$

where $R_l = \max_{l \neq m;\ 1 \leq l,\ m \leq k} \{R_{l,m}\}$ and $R_{l,m} = (S_l + S_m)/M_{l,m}$. The term $S_i = \left\{\frac{1}{N_i}\sum_{j=1}^{N_i} \| \mathbf{x}_{i(j)} - \mathbf{c}_i \|^q\right\}^{1/q}$ represents the dispersion measure of the i^{th} cluster, with N_i being the number of points in the i^{th} cluster. The Minkowski metric $M_{l,m} = \left\{\sum_{d=1}^{D} |\mathbf{c}_{l,d} - \mathbf{c}_{m,d}|^p\right\}^{1/p}$ is the distance between the centroids of l^{th} and m^{th} clusters, where $d = 1, \ldots, D$ is the d^{th} dimension of the centroid. In our study, $p = 2$ and $q = 1$ were chosen for Euclidean distance calculation.

k_{opt} is defined as the number of clusters which minimizes the DB index.

2.2.3. Silhouette-Width-Criterion (SWC). This index[14] is defined as the average silhouette of all points of the dataset:

$$\mathrm{SWC}(k) = \frac{1}{N}\sum_{i=1}^{k}\sum_{j=1}^{N_i} s(\mathbf{x}_{i(j)}) \quad (3)$$

where $s(\mathbf{x}_{i(j)}) = \dfrac{b_{i(j)} - a_{i(j)}}{\max(b_{i(j)}, a_{i(j)})}$ is the silhouette of the j^{th} point belonging to the i^{th} cluster. The term $a_{i(j)} = \dfrac{1}{N_i}\sum_{l=1,l\neq j}^{N_i} ||\mathbf{x}_{i(l)} - \mathbf{x}_{i(j)}||$ is the mean distance of this j^{th} point to all other points in the i^{th} cluster. The term $b_{i(j)} = \min_{m=1,\ldots,k;\, m\neq i}\left\{\dfrac{1}{N_m}\sum_{l=1}^{N_m} \| \mathbf{x}_{i(j)} - \mathbf{x}_{m(l)} \|\right\}$ is the average distance of $\mathbf{x}_{i(j)}$ to its nearest cluster.

k_{opt} is defined as the number of clusters which maximizes the SWC index.

2.2.4. Xie–Beni (XB). This index[15] is defined as the ratio between the compactness and the separation of clusters:

$$XB(k) = \frac{\pi}{s} \quad (4)$$

where $\pi = \dfrac{1}{N}\sum_{i=1}^{k}\sum_{j=1}^{N_i} ||\mathbf{x}_{i(j)} - \mathbf{c}_i||^2$ is the average compactness of the clusters, and $s = \min_{1\leq l,\ m\leq k;\ l\neq m} ||\mathbf{c}_l - \mathbf{c}_m||^2$ is the minimum separation measure between the clusters.

k_{opt} is defined as the number of clusters which minimizes the XB index.

2.2.5. Pakhira–Bandyopadhyay–Maulik (PBM). This index[16] is defined as the square ratio between the largest normalized inter-cluster distance D_{N} and the normalized sum of intra-cluster distances E_{N}:

$$\mathrm{PBM}(k) = \left(\frac{D_{\text{N}}}{E_{\text{N}}}\right)^2 \quad (5)$$

where $D_{\text{N}} = \dfrac{\max_{l,m=1,\ldots,k}||c_l - c_m||}{k}$, $E_{\text{N}} = \dfrac{\sum_{i=1}^{k}\sum_{j=1}^{N_i} ||\mathbf{x}_{i(j)} - c_i||}{\sum_{i=1}^{N} ||\mathbf{x}_i - \bar{\mathbf{x}}||}$ and $\bar{\mathbf{x}}$ is the average point of the entire dataset.

k_{opt} is defined as the number of clusters which maximizes the PBM index.

2.2.6. Sym-Index (SI). This index[17] is defined as the ratio between the largest normalized inter-cluster distance D_{N} and the total intra-cluster symmetry distance ε:

$$\mathrm{SI}(k) = \frac{D_{\text{N}}}{\varepsilon} \quad (6)$$

where $D_N = \dfrac{\max_{l,m=1,\ldots,k}||c_l - c_m||}{k}$ and $\varepsilon = \sum_{i=1}^{k}\sum_{j=1}^{N_i} d^*_{\text{PS}}(\mathbf{x}_{i(j)}, \mathbf{c}_i)$.

$d^*_{\text{PS}}(\mathbf{x}_{i(j)}, \mathbf{c}_i) = ||\mathbf{x}_{i(j)} - \mathbf{c}_i|| \times \dfrac{\sum_{l=1}^{\text{knn}} ||\mathbf{x}^*_{i(j)} - \mathbf{x}_{i(j(l))}||}{\text{knn}}$ denotes the Point Symmetry (PS) distance of point $\mathbf{x}_{i(j)}$ to its centroid $\mathbf{c}_i$, where $\mathbf{x}^*_{i(j)} = 2 \times \mathbf{c}_i - \mathbf{x}_{i(j)}$ is the symmetrical point of $\mathbf{x}_{i(j)}$ with respect to its centroid $\mathbf{c}_i$, $\mathbf{x}_{i(j(l))}$ is the l^{th} nearest neighbor of $\mathbf{x}^*_{i(j)}$ belonging to the i^{th} cluster, and knn is the number of considered nearest neighbors. In our study, knn was chosen equal to 2.

k_{opt} is defined as the number of clusters which maximizes the SI.

2.2.7. Context-independent optimality and partiality (COP). This index[18] is defined as the average ratio between the compactness and the separation of each cluster:

$$\mathrm{COP}(k) = \frac{1}{k}\sum_{i=1}^{k}\frac{\mathrm{intra}(i)}{\mathrm{inter}(i)} \quad (7)$$

where $\mathrm{intra}(i) = \sum_{j=1}^{N_i} ||\mathbf{x}_{i(j)} - \mathbf{c}_i||$ is the i^{th} intra-cluster distance, and $\mathrm{inter}(i) = \min_{j=1,\ \ldots,\ N_i}\{\max_{m=1,\ \ldots,\ k;\ m\neq i}\{\max_{l=1,\ \ldots,\ N_m}||\mathbf{x}_{i(j)} - \mathbf{x}_{m(l)}||\}\}$ is the i^{th} inter-cluster distance.

k_{opt} is defined as the number of clusters which minimizes the COP index.

2.2.8. Separation-Variance (SV). This index[19] is defined as the ratio between a separation cluster measure and an intra-cluster variance:

$$\mathrm{SV}(k) = \frac{S}{V} \quad (8)$$

where $S = \sum_{i=1}^{k} \min_{m=1,\ \ldots,\ k;\ m\neq i}||\mathbf{c}_i - \mathbf{c}_m||$ is the separation measure, and $V = \sum_{i=1}^{k} \max_{j=1,\ \ldots,\ N_i}||\mathbf{x}_{i(j)} - \mathbf{c}_i||$ represents the maximal intra-cluster variance, also named compactness.

k_{opt} is defined as the number of clusters which maximizes the SV index.

2.2.9. Overlap-Separation (OS). This index[19] is defined as the ratio between an overlap degree and a separation measure:

$$\mathrm{OS}(k) = \frac{O}{S} \quad (9)$$

where $S = \sum_{i=1}^{k} \min_{m=1,\ \ldots,\ k;\ m\neq i}||\mathbf{c}_i - \mathbf{c}_m||$ is the separation measure, and $O = \sum_{i=1}^{k}\sum_{j=1}^{N_i} O_{\mathbf{x}_{i(j)}}$ is the inter-cluster overlap degree. $O_{\mathbf{x}_{i(j)}} = \begin{cases} \dfrac{a_{\mathbf{x}_{i(j)}}}{b_{\mathbf{x}_{i(j)}}}, & \text{if } \dfrac{b_{\mathbf{x}_{i(j)}} - a_{\mathbf{x}_{i(j)}}}{a_{\mathbf{x}_{i(j)}} + b_{\mathbf{x}_{i(j)}}} < T \\ 0, & \text{otherwise} \end{cases}$ is the overlap degree between clusters computed at the $\mathbf{x}_{i(j)}$ point, with $a_{\mathbf{x}_{i(j)}} = \dfrac{1}{P \times N_i}\sum_{l=1}^{P\times N_i} ||\mathbf{x}_{i(j)} - \mathbf{x}_{i(j(l))}||$ being the average distance between $\mathbf{x}_{i(j)}$ and its $(P \times N_i)$ nearest neighbors belonging to the same cluster i, and $b_{\mathbf{x}_{i(j)}} = \dfrac{1}{P \times N_i}\sum_{l=1}^{P\times N_i} ||\mathbf{x}_{i(j)} - \mathbf{x}_{m(l)}||$, $m = 1$, ..., k, $m \neq i$, being the average distance between $\mathbf{x}_{i(j)}$ and its $(P \times N_i)$ nearest neighbors not belonging to clusters i. $0 < P \leq 1$ is used to define the number of considered nearest neighbors. T is the overlap rate. As suggested in ref. 19, in our study, P and T were fixed to 1 and 0.4, respectively.

k_{opt} is defined as the number of clusters which minimizes the OS index.

2.3. Hierarchical double application of CCVI

CCVIs are designed to estimate the number of clusters in datasets composed of well-separated compact clusters. To deal with datasets having a hierarchical structure, *i.e.*, datasets composed of clusters, themselves composed of several subclusters, we have developed a new methodology based on a hierarchical double application of CCVIs.

This method is composed of the following steps:

i. First, the CCVI is applied on the KM results as explained in section 2.2. The estimated optimal number of clusters is the number of main clusters k_{main} composing the dataset. The corresponding partition is called the main partition.

ii. Second, for each i^{th} cluster of this main partition, $i = \{1, ..., k_{main}\}$, the CCVI is applied a second time on the sub-dataset composed of the points belonging to this i^{th} cluster using the procedure described in section 2.2. The optimal number of sub-clusters k_{sub} (i) composing the i^{th} cluster is thus estimated. The optimal sub-partition of the i^{th} sub-dataset is thus obtained by the application of KM on this sub-dataset with $k = k_{sub}$ (i).

iii. Third, the final optimal number of clusters is computed as $k_{opt} = \sum_{i=1}^{k_{main}} k_{sub}(i)$, and the optimal double CCVI partition is obtained by assembling all the k_{main} estimated optimal sub-partitions together.

2.4. Accuracy rate

In this study, the performance of a given CCVI applied on spectral images has been evaluated by its accuracy rate. This value was defined as the percentage of images with a correct cluster estimation. Such images were defined as spectral images for which the optimal CCVI partition (single or double) retrieves at least the main histological structures, even additional sub-structures.

3. Datasets used for performance assessment of the proposed methodologies

We used two types of datasets, one artificial hierarchical dataset and five real FTIR imaging datasets.

3.1. Artificial hierarchical dataset

Before applying the double CCVI procedure on real-life datasets, we first validated its functionality by using a simulated hierarchical dataset. This dataset contains 3 main clusters composed of either 2, 3 or 4 overlapped sub-clusters (Fig. 1(a)). Each of these 9 sub-clusters consists of 300 two-dimensional points following a normal distribution with specific means and variances, as detailed in Table 1.

The closeness of the artificial sub-clusters is illustrated by a dendrogram computed on their barycenters (Fig. 1(b)). The horizontal axis represents the dissimilarity between clusters, and the vertical one, the nine artificial sub-clusters. This dendrogram shows that the dissimilarity between sub-clusters is low compared to that of the main clusters, revealing the hierarchical structure of the dataset.

3.2. FTIR imaging datasets

3.2.1. Sample preparation. Formalin-fixed paraffin-embedded tissue blocks of normal colon zones were obtained from the colon cancer surgery of five patients. For each tissue

Table 1 Means and variances of the normal distribution used to generate the x_1 and x_2 dimensions of each sub-cluster

Main cluster	M1		M2			M3			
Sub-cluster	S1	S2	S3	S4	S5	S6	S7	S8	S9
Mean (x_1)	290	240	225	170	115	410	350	360	385
Var (x_1)	300	300	300	300	400	200	200	300	300
Mean (x_2)	400	360	580	550	545	690	730	670	615
Var (x_2)	700	300	600	300	400	400	600	300	300

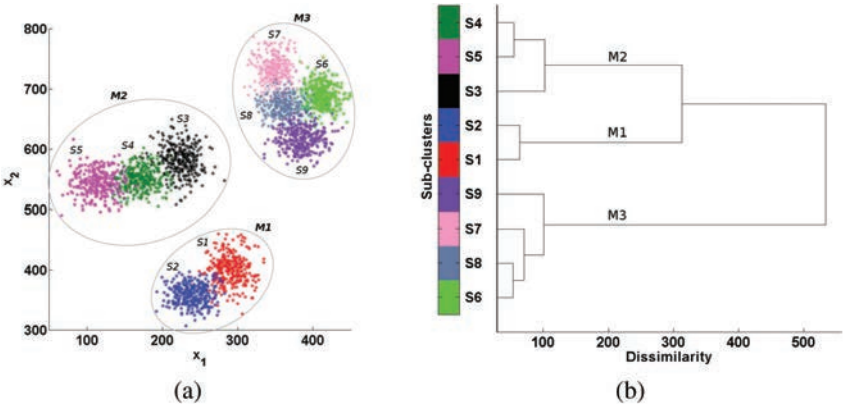


Fig. 1 Graphical representation of the hierarchical artificial dataset. (a) This dataset is composed of 3 main clusters (M1-3) and 9 sub-clusters (S1-9). (b) Dendrogram constructed from the sub-cluster barycenters.

block, two consecutive 6 μm thick slices were prepared. The first slice was mounted onto a calcium fluoride (CaF_2) window (Crystran, Dorset, UK) because this material is transparent for IR light. FTIR imaging was realized directly on this slice without any chemical staining and dewaxing. The second slice was mounted on a glass window and stained by Harris' Hematoxylin and Eosin (HE) for conventional histological analysis, by an pathologist. This HE slice is considered in our study as a reference for comparison with FTIR imaging.

3.2.2. FTIR image acquisition. On each sample slice, FTIR imaging was performed with a Spectrum Spotlight 300 FTIR imaging system coupled to a Spectrum one FTIR spectrometer (Perkin Elmer, Courtabœuf, France), equipped with a liquid nitrogen-cooled 16-element Mercury Cadmium Telluride (MCT) detector and calibrated for transmission mode.

Using the Spectrum Image software (Perkin Elmer), FTIR images were recorded with a 6.25 μm spatial resolution. Each pixel contained a full absorbance spectrum averaged over 16 scans in a mid-IR range of 750 to 4000 cm^{-1}, with a 4 cm^{-1} spectral resolution. Prior to each image acquisition, a reference spectrum was recorded with 240 accumulations from a blank area of the same CaF_2 window but outside and away from the paraffin-embedded tissue section.

For each tissue section, an FTIR image was recorded on the tissue area (between 8536 and 14 224 pixels). Then, another image was acquired on a zone of the tissue section containing only pure paraffin (between 1806 and 10 000 spectra) for numerical dewaxing (see further section 3.2.3). Knowing that the acquisition time for one pixel is 0.7 second, each complete image was recorded between 21 and 166 minutes.

3.2.3. Data processing. The raw spectra are composed not only of the informative biochemical absorptions, but also of other undesirable signals such as the atmospheric interference of water vapor and CO_2, paraffin absorption and scattering effects. The application of specific preprocessing steps is thus necessary in order to correct the spectra from these parasitic signals.

First, atmospheric correction was performed on each FTIR image by the Spectrum Image software to remove water vapor and CO_2 contributions.

The further processing steps were applied using in-house scripts written with Matlab (The Mathworks, Natick, MA, USA).

Second, the spectral range was limited to the 900–1800 cm^{-1} range which is known as the fingerprint region for biological samples.[20]

Third, the spectra were numerically corrected from paraffin signal using a method based on Extended Multiplicative Signal Correction (EMSC)[21] and described in a previous study.[22] This technique linearly models each spectrum $\mathbf{s}$ as:

$$\mathbf{s} = a\hat{\mathbf{s}} + \mathbf{bP} + \mathbf{cI} + \mathbf{e} \quad (10)$$

where $\hat{\mathbf{s}}$ is the mean image spectrum, $\mathbf{bP}$ is a polynomial function modeling the physical light scattering effects, $\mathbf{I}$ is the interference matrix composed of the first principal components estimated from the pure paraffin spectra acquired in a pure paraffin zone, and $\mathbf{e}$ is the modeling error. The coefficients a, $\mathbf{b}$, $\mathbf{c}$ are estimated by ordinary least squares in order to minimize the modeling error. The spectra are corrected by the following equation:

$$\mathbf{s}_{\mathbf{EMSC}} = \hat{\mathbf{s}} + \frac{\mathbf{e}}{a} \quad (11)$$

This processing step is efficient to remove the variance of paraffin and scattering effects, and to normalize the spectra on the mean image spectrum. The corrected FTIR images are thus finally composed only of the spectral variance of the tissue.

In our study, a fourth-order polynomial function was used to model $\mathbf{P}$. The first ten principal components of paraffin and the mean paraffin spectrum were used to construct the interference matrix $\mathbf{I}$, in order to keep at least 95% of the variance of the paraffin spectra.

3.2.4. KM pseudo-color-coded images and cluster assignment. After numerical dewaxing, KM clustering was applied on each preprocessed IR image. For each cluster of an estimated KM partition, a unique color can be attributed to its pixels. Thus, a color-coded image can be reconstructed for a rapid and simple visual analysis of KM results. The generated clusters were annotated into their corresponding histological classes by an expert pathologist using the HE-stained images as a reference.

4. Results and discussion

The efficiency of the single application of CCVIs (see section 2.2) was compared to that of the hierarchical double application (see section 2.3). To do so, both methods were first tested on a hierarchical artificial dataset previously described in section 3.1.

4.1. Limitations of a single application of CCVIs

The nine selected CCVIs were tested on a hierarchical artificial dataset using KM clustering with the number of clusters k varying from 2 to 20.

As shown in Table 2, the optimal number of clusters k_{opt} estimated by a single application of CCVIs was equal to 3 for the nine CCVIs. These 3 clusters illustrated in Fig. 2 (in green, blue and red) correspond to the 3 main clusters of the artificial dataset (Fig. 1(a)). This indicates that a single CCVI application performed well with well-separated compact clusters, which is in accordance with the mathematical definition of CCVIs. In addition, these results showed that the single CCVI application was unable to reach the sub-structures of the hierarchical dataset. To overcome this problem, a second CCVI application was performed on each cluster.

4.2. Differential efficiency of a hierarchical double application according to the CCVI type

As defined in section 2.3, the hierarchical double CCVI application involves three steps, with outcome results illustrated by the PBM index. The result of the first step is exactly the same as that of the optimal single CCVI partition in 3 main clusters (as already presented in Fig. 2). For the 1st, 2nd and 3rd main clusters, the second step estimated optimal sub-partitions with 2, 3 and 4 sub-clusters, respectively (Fig. 3(a)–3(c)). The

Table 2 Optimal number of clusters k_{opt} estimated by two types (single and double) of CCVI applications for an artificial dataset, using the nine selected CCVIs. Bold values indicate the estimated optimal number of clusters inducing KM partitions that best fit the original partition

	CCVIs								
Application types	PBM	SI	DB	COP	SWC	XB	Dunn	SV	OS
Single	3	3	3	3	3	3	3	3	3
Double	**9**	**9**	8	8	8	8	48	26	59

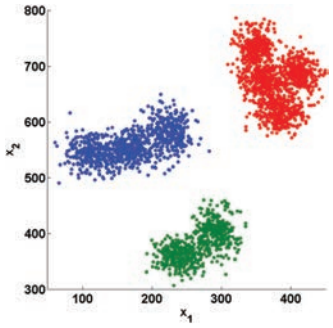


Fig. 2 Optimal single CCVI partition assessed for the artificial dataset using k_{opt} = 3.

optimal double PBM partition, as calculated in the third step, is thus composed of $k_{opt} = 2 + 3 + 4 = 9$ sub-clusters (Fig. 3(d)).

The estimation of k_{opt} for each of the nine studied CCVIs is shown in Table 2. Three CCVI groups can be deduced. The first represented by the PBM and SI indices estimated correctly the expected 9 sub-clusters (Fig. 3(d)). In contrast, the second including DB, COP, SWC and XB indices underestimated the number of sub-clusters with k_{opt} = 8 (Fig. 4(a)) and the third with Dunn, SV and OS indices overestimated it with $k_{opt} \geq 26$ (Fig. 4(b)).

Taken together, these data indicate that a hierarchical double application of CCVIs using the PBM and SI indices enabled access to the dataset sub-structures by estimating the expected partition.

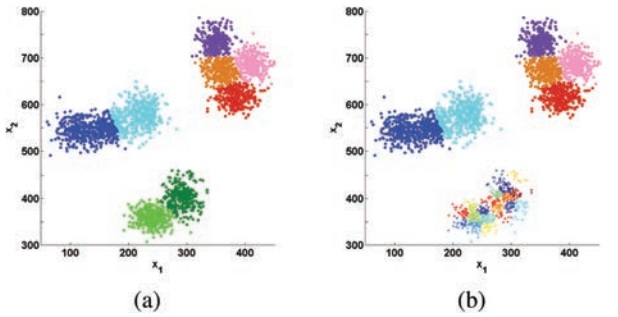


Fig. 4 Optimal double CCVI partitions estimated for the artificial dataset by (a) DB (k_{opt} = 8) and (b) SV (k_{opt} = 26) indices.

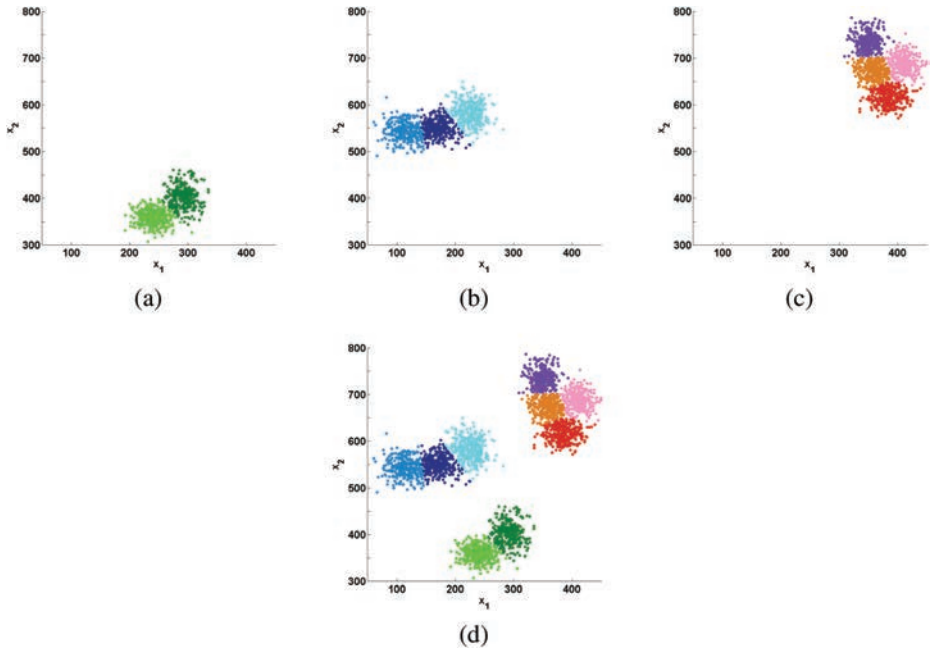


Fig. 3 Results obtained from the second and third steps of the hierarchical double PBM application for the artificial dataset. The optimal PBM sub-partition estimated by the second step for each main cluster is composed of (a) k_{sub} (1) = 2, (b) k_{sub} (2) = 3 and (c) k_{sub} (3) = 4 sub-clusters. (d) Optimal double PBM partition obtained by assembling (a), (b) and (c). The same results were obtained using a hierarchical double SI application.

4.3. Efficiency of single *versus* double application of CCVIs on FTIR imaging datasets

Here, single and double CCVI applications were performed on real-life data. To this end, FTIR spectral images were recorded from healthy areas of the human colon tissue as mentioned in section 3.2.

The KM pseudo-color-coded images reconstructed from optimal CCVI partitions were then compared to conventional histology for a morphological recognition of the tissue structures. Fig. 5(a) illustrates the four histological tissue structures characteristic of a cross-section of the human colon tract. The outer layer or mucosa is composed of the crypts of Lieberkuhn (structure 1) and of a connective tissue, the lamina propria (structure 2). Then, muscularis mucosae (structure 3) of variable thickness precede the underlying submucosa (structure 4) that contains abundant adipose tissue. In some tissue sections, we can observe an additional structure, the lymphoid aggregate (Fig. 5(b)) that can extend from the lamina propria through the muscularis mucosae and into the submucosa.[23]

Among the tissue samples analyzed in this study, four were composed of the four histological structures, and one contained in addition a lymphoid aggregate. For these samples, the k_{opt} and the CCVI accuracy rates have been evaluated by the single (Table 3) and double (Table 4) CCVI applications. As defined in section 2.4, an accuracy rate of 100% indicates that

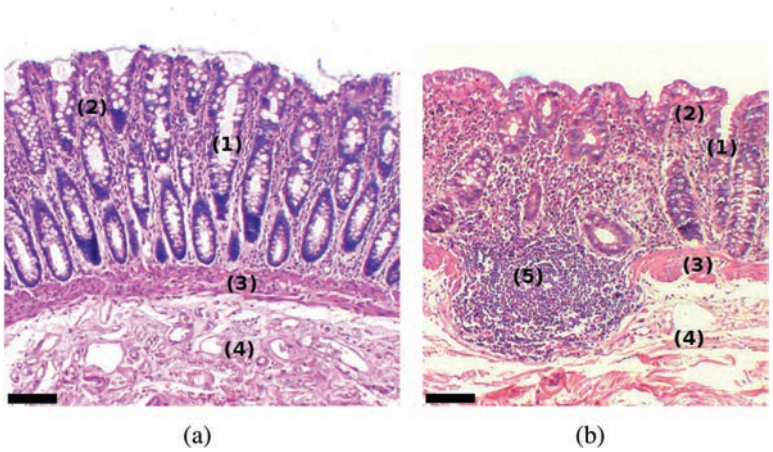


Fig. 5 Conventional histology of a normal colon tissue. (a) and (b) HE stained images of tissues where the crypts, lamina propria, muscularis mucosae, submucosa and lymphoid aggregate are annotated by (1), (2), (3), (4) and (5), respectively. Scale bar indicates 100 μm.

Table 3 Optimal number of clusters k_{opt} and accuracy rates estimated by a single CCVI application on the FTIR images. Bold values represent the optimal partitions that retrieved at least the main histological structures of a normal human colon

		CCVIs								
Patient sample	Number of histological structures	Dunn	OS	SV	PBM	SI	DB	COP	SWC	XB
1	4	**13**	3	3	4	5	2	2	2	3
2	4	4	3	3	3	3	3	2	3	3
3	4	**10**	**19**	**19**	4	5	3	2	2	3
4	4	**19**	**20**	2	4	4	2	2	2	2
5	5	2	**20**	2	4	4	2	2	2	2
Accuracy rate (%)		60	60	20	0	0	0	0	0	0

Table 4 Optimal number of clusters k_{opt} and accuracy rates estimated by a hierarchical double CCVI application on the FTIR images. Bold values represent the optimal partitions that retrieved at least the main histological structures of a normal human colon

		CCVIs								
Patient sample	Number of histological structures	PBM	SI	Dunn	OS	SV	DB	COP	SWC	XB
1	4	**12**	**13**	**198**	**60**	**39**	4	4	4	6
2	4	**9**	**9**	**74**	**60**	**56**	**7**	4	**7**	**7**
3	4	**12**	**20**	**164**	**373**	**325**	**7**	4	4	6
4	4	**12**	**12**	**312**	**392**	4	4	**7**	4	4
5	5	**10**	**11**	**12**	**393**	**39**	**8**	5	4	4
Accuracy rate (%)		100	100	100	100	80	60	20	20	20

all the histological structures were correctly identified for all the tested samples.

In Table 3, the highest accuracy rates were observed for Dunn and OS (60%), and SV (20%). The other CCVIs were totally inefficient due to an underestimation of the number of clusters. It has to be noticed that the number of clusters superior or equal to the number of histological structures do not necessarily induce a partition with a correct cluster estimation. For example, in Fig. 6, some clusters of the optimal single SI partitions, of patients 1, 3 and 4, mixed two histologi-

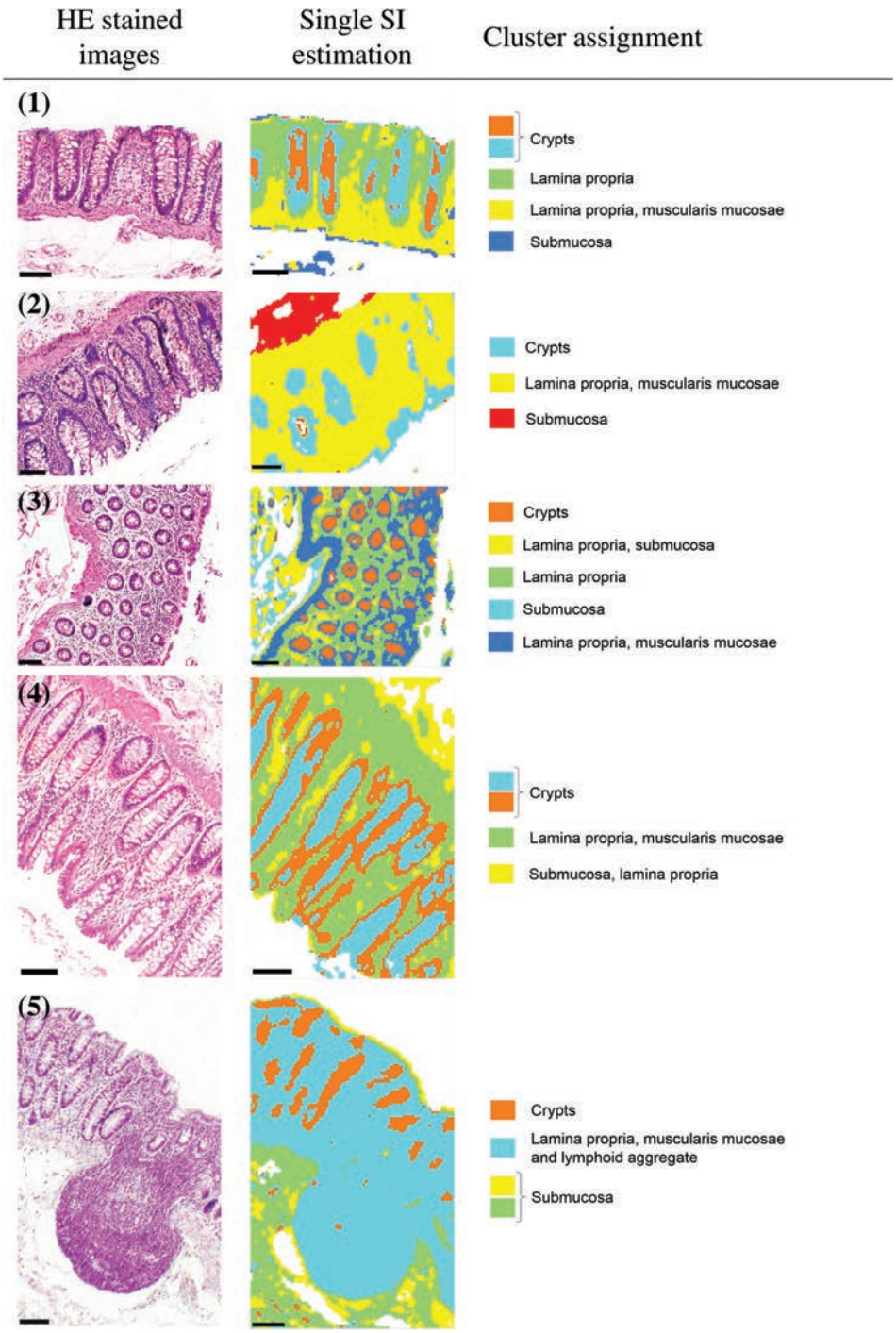


Fig. 6 Comparison between conventional histology and pseudo-color-coded images reconstructed from optimal single SI partitions. The numbers in parentheses refer to the patient sample. Scale bars indicate 100 μm.

cal structures, the muscularis mucosae and the lamina propria.

As can be seen in Table 4, PBM and SI were revealed to be the most effective indices since (i) their accuracy rates reached 100% and (ii) their k_{opt} varied between 9 and 20 while exactly matching the main colon tissue structures (Fig. 7). These observed additional sub-clusters could correspond to the histological sub-structures as previously described.[24] Indeed, in colon crypts analyzed by FTIR imaging, we have identified sub-clusters correlated to the secreted mucus and to the basal and apical parts of the mucus-secreting cells.

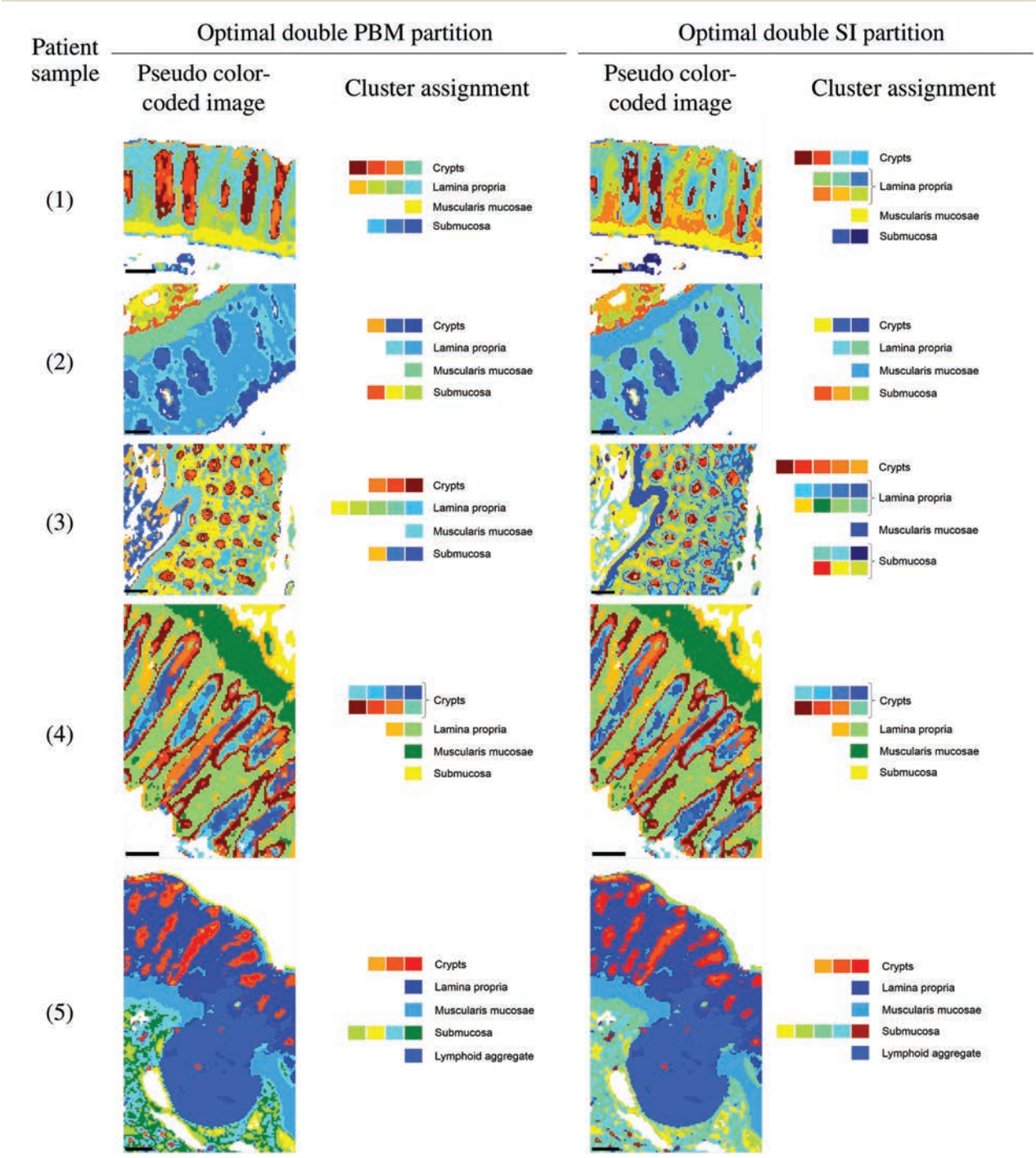


Fig. 7 Pseudo-color-coded images reconstructed from optimal double PBM and SI partitions for the five colon tissue samples. Scale bars indicate 100 μm.

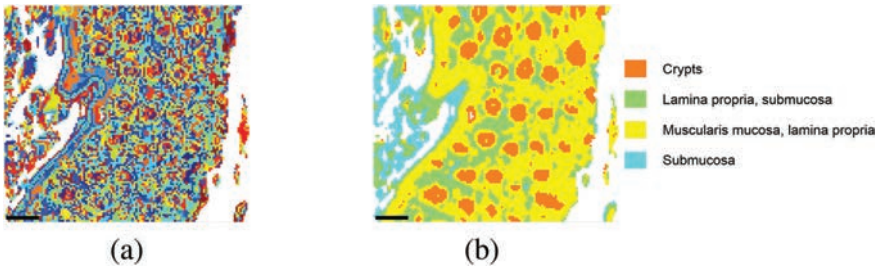


Fig. 8 Pseudo-color-coded images reconstructed by double CCVI applications. Optimal partition estimated by (a) Dunn (k_{opt} = 164) and (b) COP (k_{opt} = 4). The corresponding HE image is shown in Fig. 6 (3). Scale bars indicate 100 μm.

Although their accuracy rate ranged between 80 and 100%, Dunn, OS, and SV exhibited dramatically high k_{opt} values, thus complicating the assignment of clusters to the corresponding histological structures (Fig. 8a). In addition, these indices presented a wide range (4 to 393) of k_{opt} values. This can be explained by their wide range of k_{opt} values previously estimated after a single CCVI application (2 to 20 clusters).

In contrast, DB, COP, SWC and XB estimated low k_{opt} inducing optimal partitions which partially corresponds to the histological structures. For example, the optimal COP partition shown in Fig. 8(b), assigned the yellow cluster to both the muscularis mucosae and the lamina propria.

Applied to normal human colon tissues, PBM and SI were revealed to be the most efficient indices giving access to the main histological structures and sub-structures. FTIR microimaging coupled with a hierarchical double CCVI application represents automated histology, more informative, more discriminant, and more objective than histopathology. However, further studies are needed to develop a fully automated tissue sub-structure assignment. This will be performed by training supervised classification algorithms on spectral databases constructed from the different tissue sub-structures.

Acknowledgements

The authors are grateful for financial support from Cancéropôle Grand-Est, Ligue contre le Cancer, the URCA technological platform of cellular and tissular imaging PICT-IBiSA, Région Champagne-Ardenne, Région Alsace and Ministère de l'Enseignement Supérieur et de la Recherche.

References

1 F. L. Martin, J. G. Kelly, V. Llabjani, P. L. Martin-Hirsch, I. I. Patel, J. Trevisan, N. J. Fullwood and M. J. Walsh, *Nat. Protoc.*, 2010, **5**, 1748–1760.
2 M. Sattlecker, N. Stone and C. Bessant, *TrAC, Trends Anal. Chem.*, 2014, **59**, 17–25.
3 J. MacQueen, *Proc. fifth Berkeley symp. math. stat. probab.*, 1967, **1**, 281–297.
4 J. Ward and H. Joe, *J. Am. Stat. Assoc.*, 1963, **58**, 236–244.
5 J. C. Bezdek, *Pattern recognition with fuzzy objective function algorithms*, Plenum Press, New York, 1981.
6 J. Trevisan, P. P. Angelov, P. L. Carmichael, A. D. Scott and F. L. Martin, *Analyst*, 2012, **137**, 3202–3215.
7 L. Vendramin, R. J. G. B. Campello and E. R. Hruschka, *Stat. Anal. Data Min.: ASA Data Sci. J.*, 2010, **03**, 209–235.
8 O. Arbelaitz, I. Gurrutxaga, J. Muguerza, J. M. Pérez and I. Perona, *Pattern Recognit.*, 2013, **46**, 243–256.
9 W. Wang and Y. Zhang, *Fuzzy Sets Syst.*, 2007, **158**, 2095–2117.
10 X.-Y. Wang, J. Garibaldi, B. Bird and M. George, *Appl. Intell.*, 2007, **27**, 237–248.
11 D. Sebiskveradze, V. Vrabie, C. Gobinet, A. Durlach, P. Bernard, E. Ly, M. Manfait, P. Jeannesson and O. Piot, *Lab. Invest.*, 2011, **91**, 799–811.
12 J. C. Dunn, *J. Cybern.*, 1974, **4**, 95–104.
13 D. L. Davies and D. W. Bouldin, *IEEE Trans. Pattern Anal. Mach. Intell.*, 1979, **1**, 224–227.
14 P. J. Rousseeuw, *J. Comput. Appl. Math.*, 1987, **20**, 53–65.
15 X. L. Xie and G. Beni, *IEEE Trans. Pattern Anal. Mach. Intell.*, 1991, **13**, 841–847.
16 M. K. Pakhira, S. Bandyopadhyay and U. Maulik, *Pattern Recognit.*, 2004, **37**, 487–501.
17 S. Bandyopadhyay and S. Saha, *IEEE Trans. Knowl. Data Eng.*, 2008, **20**, 1441–1457.
18 I. Gurrutxaga, I. Albisua, O. Arbelaitz, J. I. Martín, J. Muguerza, J. M. Pérez and I. Perona, *Pattern Recognit.*, 2010, **43**, 3364–3373.
19 K. R. Žalik and B. Žalik, *Pattern Recognit. Lett.*, 2011, **32**, 221–234.
20 M. Khanmohammadi, A. B. Garmarudi, K. Ghasemi, H. K. Jaliseh and A. Kaviani, *Med. Oncol.*, 2009, **26**, 292–297.
21 A. Kohler, C. Kirschner, A. Oust and H. Martens, *Appl. Spectrosc.*, 2005, **59**, 707–716.
22 R. Wolthuis, A. Travo, C. Nicolet, A. Neuville, M.-P. Gaub, D. Guenot, E. Ly, M. Manfait, P. Jeannesson and O. Piot, *Anal. Chem.*, 2008, **80**, 8461–8469.
23 P. M. Treuting and S. M. Dintzis, *Comparative Anatomy and Histology: A Mouse and Human Atlas*, Academic Press, 2011.
24 A. Travo, O. Piot, R. Wolthuis, C. Gobinet, M. Manfait, J. Bara, M.-E. Forgue-Lafitte and P. Jeannesson, *Histopathology*, 2010, **56**, 921–931.

III.3 - Résultats supplémentaires

III.3.1 - Application des indices de validité sur un jeu de données hiérarchiques multi-dimensionnelles

Le jeu de données synthétiques utilisé dans l'article pour démontrer l'efficacité de notre approche est composé de deux dimensions. Cependant, un jeu aussi simple est difficilement comparable à une image spectrale composée, dans ce travail, de 451 nombres d'onde, donc 451 dimensions ou variables. C'est pourquoi un deuxième jeu de données artificielles a été généré de façon à mimer des données spectrales IR.

III.3.1.a. Construction du jeu de données

Ce jeu de données comprend 3 classes principales, chacune composée de 2, 3 ou 4 sous-classes qui se chevauchent. Donc, 9 sous-classes composent ce jeu de données. Chaque sous-classe contient 300 objets de 451 dimensions. Chaque dimension d ($d = \{1,2,\dots 451\}$) est générée indépendamment par l'un des trois modèles ci-dessous choisi aléatoirement selon une distribution uniforme :

Modèle 1 : La valeur de la $d^{\text{ème}}$ dimension de chaque objet suit une loi normale de moyenne M_{ij} et de variance V_{ij} avec :

- $i = \{1,2,3\}$ l'indice de la classe principale à laquelle appartient l'objet,
- $j = \{1,2,3,4\}$ l'indice de la sous-classe à laquelle appartient l'objet,
- $M_{ij} = M_i + I_{ij}$ la moyenne de la $j^{\text{ème}}$ sous-classe appartenant à la $i^{\text{ème}}$ classe principale, où :
  - M_i est la moyenne de la $i^{\text{ème}}$ classe principale et est choisie aléatoirement (suivant la loi uniforme) dans l'intervalle $[100;500]$,
  - I_{ij} est la moyenne intermédiaire de la $j^{\text{ème}}$ sous-classe appartenant à la $i^{\text{ème}}$ classe principale et est choisie aléatoirement (suivant la loi uniforme) dans l'intervalle $[-75;75]$,

- V_{ij} la variance de la $j^{\text{ème}}$ sous-classe appartenant à la $i^{\text{ème}}$ classe principale, où :
  - $V_{11} = V_{12} = V_{21} = V_{22} = V_{33} = V_{34} = 300$,
  - $V_{23} = 400$,
  - $V_{31} = V_{32} = 200$.

Modèle 2 : La valeur de la $d^{\text{ème}}$ dimension de chaque objet est générée en utilisant le modèle 1 avec :

- M_i choisie dans l'intervalle $[300;800]$,
- $V_{11} = 700$,
- $V_{12} = V_{22} = V_{33} = V_{34} = 300$,
- $V_{21} = V_{32} = 600$,
- $V_{23} = V_{31} = 400$.

Modèle 3 : Les valeurs de la $d^{\text{ème}}$ dimension des objets suivent la loi normale de moyenne M et de variance V choisies aléatoirement (suivant la loi uniforme) respectivement dans les intervalles $[100;800]$ et $[0;700]$.

Les modèles 1 et 2 ont pour but de générer la structure hiérarchique du jeu de données. Le modèle 3 permet de générer des variables non informatives.

De plus, un bruit Gaussien de moyenne nulle et de variance choisie aléatoirement (suivant la loi uniforme) dans l'intervalle $[0;100]$ a été ajouté à chaque valeur générée par les modèles 1 et 2. Enfin, pour obtenir des intensités similaires aux différences spectrales observées entre les tissus coliques, les valeurs de chaque variable ont été normalisées par un facteur choisi aléatoirement (suivant la loi uniforme) dans l'intervalle $\left[-10^{-5};10^{-5}\right]$.

Ces conditions permettent donc de construire un jeu de données artificielles qui peut être comparé à nos données spectrales en termes d'échelle, de multi-dimensionnalité, de rapport signal sur bruit et de variables non informatives (Figure III.1)

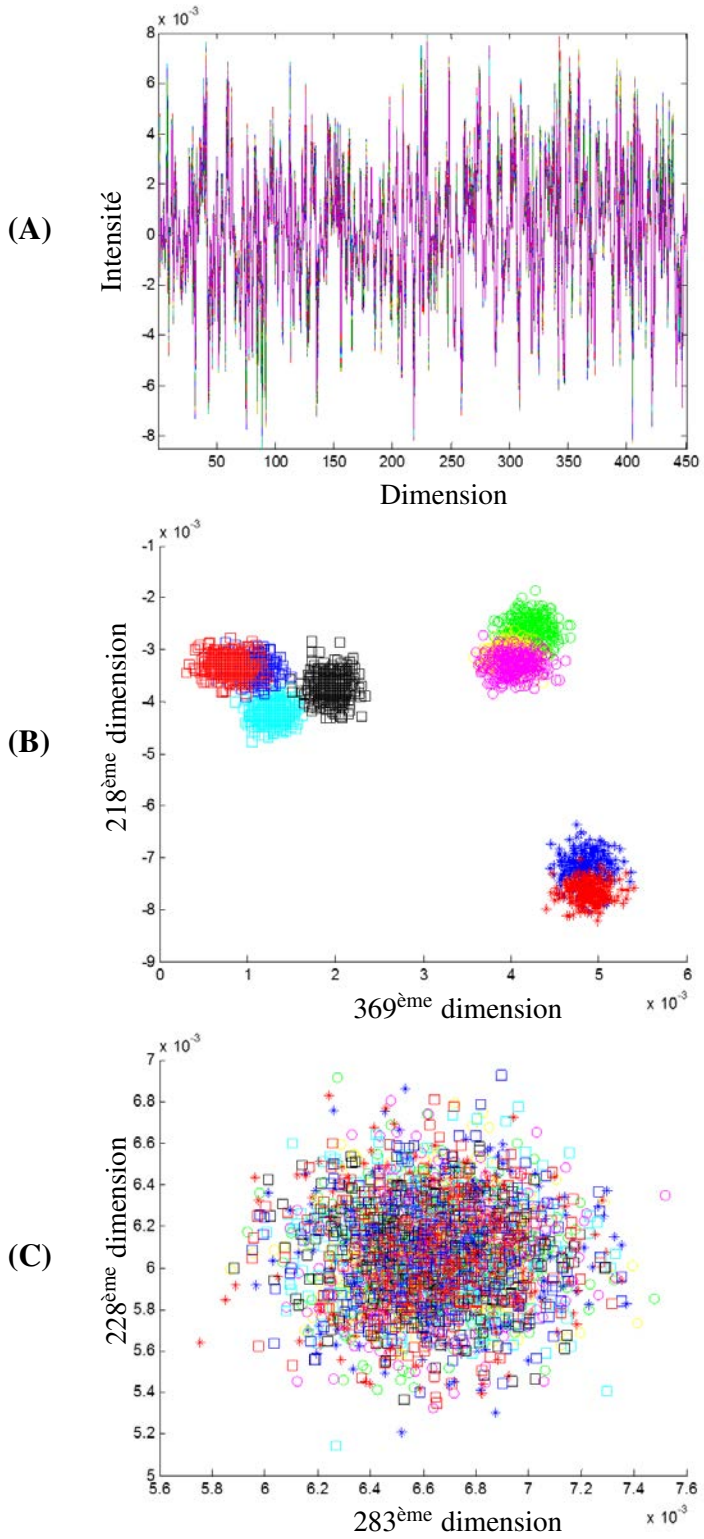


FIGURE III.1 – Illustration du jeu de données synthétiques composé de 451 dimensions. (A) Données synthétiques générées. (B) Représentation 2D de la structure hiérarchique des données synthétiques. (C) Représentation 2D des données synthétiques pour deux dimensions non informatives.

III.3.1.b. Limite des indices de validité et efficacité de la double application hiérarchique sur le jeu de données synthétiques

Tout d'abord, les indices de validité ont simplement été appliqués sur ce nouveau jeu de données synthétiques. Les résultats présentés sur la deuxième ligne du Tableau III.1 montrent que le nombre de classes estimé correspond aux nombre de classes principales.

Malgré la complexité de ce jeu de données hiérarchiques, l'application double des indices de validité permet d'accéder à toutes ses sous-structures, comme présenté sur la dernière ligne du Tableau III.1. Seule la double application d'OS surestime le nombre de sous-classes.

L'ensemble de ces résultats démontre l'efficacité de notre approche à retrouver la structure hiérarchique d'un jeu de données, quelle que soit sa complexité.

TABLE III.1 – Nombre optimal de classes k_{opt} estimé par la simple et la double application d'indices de validité sur le jeu de données synthétiques de 451 dimensions. Les valeurs grisées représentent le nombre estimé de classes induisant la partition la plus proche de la partition originale.

Types d'application	Indices de validité								
	PBM	SI	DB	COP	SWC	XB	Dunn	SV	OS
Simple	3	3	3	3	3	3	3	3	3
Double	9	9	9	9	9	9	9	9	26

III.3.2 - Application des 28 indices de validité supplémentaires sur les images spectrales

III.3.2.a. Présentation de ces indices

Le nom, l'acronyme et la référence dans laquelle ces indices de validité ont été décrits sont répertoriés dans le Tableau III.2.

CI, CS, S_Dbw et VRC sont des indices de validité originaux basés sur des concepts différents de ceux des indices présentés dans l'article. Les 24 autres indices sont des extensions ou des généralisations de ceux présentés dans notre article :

TABLE III.2 – Liste des indices de validité supplémentaires. Pour chaque indice, son nom, son acronyme et sa référence bibliographique sont donnés.

Indice de validité	Acronyme	Référence
Alternative Simplified Silhouette Width Criterion	ASSWC	[80]
Alternative Silhouette Width Criterion	ASWC	[80]
C-Index	CI	[34]
Chou-Su	CS	[17]
Davies-Bouldin*	DB*	[39]
Davies-Bouldin**	DB**	[39]
Indices Dunn généralisés	$Dunn_{12}$, $Dunn_{13}$, $Dunn_{21}$, $Dunn_{22}$, $Dunn_{23}$, $Dunn_{31}$, $Dunn_{32}$, $Dunn_{33}$, $Dunn_{41}$, $Dunn_{42}$, $Dunn_{43}$, $Dunn_{51}$, $Dunn_{52}$, $Dunn_{53}$, $Dunn_{61}$, $Dunn_{62}$, $Dunn_{63}$	[12]
S_Dbw	S_Dbw	[31]
Simplified Silhouette Width Criterion	SSWC	[80]
Variance Ratio Criterion	VRC	[15]
Xie-Beni*	XB*	[39]
Xie-Beni**	XB**	[39]

- SSWC, ASWC et ASSWC sont des variantes de SWC. Ils diffèrent dans la définition de la dissimilarité moyenne entre un objet et un cluster voisin et dans la définition de la silhouette d'un objet.
- DB*, DB** et XB*, XB** sont dérivés respectivement de DB et XB par de nouvelles définitions de la compacité des classes.
- $Dunn_{ij}$ avec $i = \{1,2,\dots 6\}$ et $j = \{1,2,3\}$ sont des généralisations de Dunn, sachant que $Dunn_{11}$ est l'indice original. L'indice i fait référence à l'une des 6 définitions données dans [12] pour calculer la séparation entre les classes. L'indice j permet de choisir l'une des 3 définitions possibles de la compacité des classes selon [12].

III.3.2.b. Échec de l'application de ces indices de validité pour les 5 patients étudiés dans l'article

Le Tableau III.3 présente le nombre optimal de classes k_{opt} et le taux de précision évalués par ces indices supplémentaires appliqués aux images spectrales acquises pour ces 5 patients. L'inefficacité de ces indices est évidente puisque le nombre estimé de classes est soit plus petit, soit beaucoup plus grand (proche de $k_{max} = 20$) que le nombre de structures histologiques attendues.

Seuls CI, S_Dbw, XB** et CS possèdent un taux de précision de 100% lors de leur double

TABLE III.3 – Nombre optimal de classes k_{opt} et taux de précision estimés par la simple et la double application des indices de validité supplémentaires. Les valeurs grisées représentent les partitions optimales qui retrouvent les structures histologiques principales du côlon humain normal.

		Application simple d'indices						Application double d'indices					
#		1	2	3	4	5	Taux de	1	2	3	4	5	Taux de
Nombre de structures histologiques		4	4	4	4	5	précision (%)	4	4	4	4	5	précision (%)
Indice	CI	18	20	18	18	14	100	308	364	309	317	238	100
	S_Dbw	20	19	18	19	18	100	386	365	328	343	295	100
	XB**	15	3	18	12	14	80	214	50	281	199	205	100
	CS	3	3	2	2	2	0	50	50	22	22	34	100
	$Dunn_{13}$	6	3	2	3	2	0	35	8	7	6	5	60
	$Dunn_{12}$	2	3	2	3	2	0	4	8	7	6	5	40
	ASWC	3	3	2	2	2	0	7	7	4	4	4	20
	ASSWC	3	2	2	2	2	0	6	4	4	4	4	0
	DB*	2	2	2	2	2	0	4	4	4	4	4	0
	DB**	2	2	2	2	2	0	6	4	4	5	4	0
	$Dunn_{21}$	2	2	2	2	2	0	4	4	4	4	4	0
	$Dunn_{22}$	2	2	2	2	2	0	4	4	4	4	4	0
	$Dunn_{23}$	2	2	2	2	2	0	4	4	4	4	4	0
	$Dunn_{31}$	2	2	2	2	2	0	4	4	4	4	4	0
	$Dunn_{32}$	2	2	2	2	2	0	4	4	4	4	4	0
	$Dunn_{33}$	2	2	2	2	2	0	4	4	4	4	4	0
	$Dunn_{41}$	2	2	2	2	2	0	4	4	4	4	4	0
	$Dunn_{42}$	2	2	2	2	2	0	4	4	4	4	4	0
	$Dunn_{43}$	2	2	2	2	2	0	4	4	4	4	4	0
	$Dunn_{51}$	2	2	2	2	2	0	4	4	4	4	4	0
	$Dunn_{52}$	2	2	2	2	2	0	4	4	4	4	4	0
	$Dunn_{53}$	2	2	2	2	2	0	4	4	4	4	4	0
	$Dunn_{61}$	2	2	2	2	2	0	4	4	4	4	5	0
	$Dunn_{62}$	2	2	2	2	2	0	4	4	4	4	4	0
	$Dunn_{63}$	2	2	2	2	2	0	4	4	4	4	4	0
	SSWC	3	2	2	2	2	0	6	4	4	4	4	0
	VRC	3	2	2	2	3	0	7	4	4	4	6	0
	XB*	2	2	2	2	2	0	4	4	4	4	4	0

application hiérarchique. Cependant, leurs valeurs de k_{opt} sont très largement surestimées (entre 22 et 386 classes). L'interprétation des partitions correspondantes est donc impossible par un anatomo-pathologiste. Les autres indices sous-estiment k_{opt}, ne permettant pas de retrouver toutes les structures histologiques principales des coupes analysées.

Ces résultats confirment les conclusions de l'article, à savoir que seule une double application hiérarchique de PBM ou SI est capable de retrouver la structure de données IR acquises sur des coupes tissulaires de côlon.

III.3.3 - Confirmation des résultats sur 10 nouveaux patients

Afin de vérifier les résultats précédents, des images spectrales ont été acquises sur des échantillons coliques provenant de 10 nouveaux patients. Sur ces données, une simple application et une double application hiérarchique des 37 indices présentés dans ce chapitre (9 dans l'article et 28 dans les résultats supplémentaires) ont été réalisées.

L'incapacité de la simple application à retrouver les structures tissulaires est confirmée dans le Tableau III.4.

Pour la double application hiérarchique (Tableau III.5), PBM et SI restent les seuls indices capables d'accéder à la structure complexe des données avec un nombre estimé de classes relativement petit. Cependant, ces indices ne sont pas infaillibles puisqu'ils n'ont pas été capables d'élucider la structure des données du patient #14. Afin de surmonter cette limite, une stratégie est proposée dans le chapitre VI.

Dans la suite de ce travail, seule la double application de PBM sera utilisée car :

i) PBM estime un nombre de classes inférieur à SI, facilitant ainsi l'affectation des classes aux structures histologiques par un anatomo-pathologiste,

ii) en terme de temps d'exécution, PBM est plus rapide que SI.

TABLE III.4 – Nombre optimal de classes k_{opt} et taux de précision estimés par la simple application de l'ensemble des indices de validité sur les 10 patients supplémentaires. Les valeurs grisées représentent les partitions optimales qui retrouvent les structures histologiques principales du côlon humain normal.

#		6	7	8	9	10	11	12	13	14	15	Taux de
Nombre de structures histologiques		4	4	4	4	4	4	4	4	4	4	précision (%)
Indice	CI	18	20	20	20	20	18	20	20	18	20	100
	S_Dbw	20	11	20	20	20	19	19	20	20	19	100
	OS	20	20	20	20	20	20	18	20	20	20	100
	Dunn	2	15	19	20	19	19	18	11	2	20	80
	XB**	18	3	17	17	13	12	19	4	18	8	70
	CS	2	13	2	3	4	2	18	20	2	20	30
	SV	2	13	2	3	4	2	18	20	2	20	30
	ASSWC	2	3	2	2	2	2	3	2	2	3	0
	ASWC	2	3	2	2	2	2	3	3	2	3	0
	COP	2	3	2	2	2	2	2	2	2	2	0
	DB	3	3	3	2	2	2	3	3	3	3	0
	DB*	2	3	2	2	2	2	2	2	2	2	0
	DB**	2	2	2	2	2	2	3	2	2	2	0
	$Dunn_{12}$	2	4	5	3	2	2	3	3	2	2	0
	$Dunn_{13}$	2	4	5	3	2	2	3	3	2	2	0
	$Dunn_{21}$	2	2	2	2	2	2	2	2	2	2	0
	$Dunn_{22}$	2	2	2	2	2	2	2	2	2	2	0
	$Dunn_{23}$	2	2	2	2	2	2	2	2	2	2	0
	$Dunn_{31}$	2	2	2	2	2	2	2	2	2	2	0
	$Dunn_{32}$	2	2	2	2	2	2	2	2	2	2	0
	$Dunn_{33}$	2	2	2	2	2	2	2	2	2	2	0
	$Dunn_{41}$	2	2	2	2	2	2	2	2	2	2	0
	$Dunn_{42}$	2	2	2	2	2	2	2	2	2	2	0
	$Dunn_{43}$	2	2	2	2	2	2	2	2	2	2	0
	$Dunn_{51}$	2	2	2	2	2	2	2	2	2	2	0
	$Dunn_{52}$	2	2	2	2	2	2	2	2	2	2	0
	$Dunn_{53}$	2	2	2	2	2	2	2	2	2	2	0
	$Dunn_{61}$	2	2	2	2	2	2	2	2	2	2	0
	$Dunn_{62}$	2	2	2	2	2	2	2	2	2	2	0
	$Dunn_{63}$	2	2	2	2	2	2	2	2	2	2	0
	PBM	4	3	3	5	4	4	5	4	4	6	0
	SI	4	7	3	5	4	4	5	4	4	6	0
	SSWC	2	3	2	2	2	2	3	2	2	3	0
	SWC	2	3	2	2	2	2	3	3	2	2	0
	VRC	2	3	2	2	2	4	5	2	2	2	0
	XB	2	3	2	2	2	2	3	3	2	2	0
	XB*	2	2	2	2	2	2	2	2	2	2	0

TABLE III.5 – Nombre optimal de classes k_{opt} et taux de précision estimés par la double application de l'ensemble des indices de validité sur les 10 patients supplémentaires. Les valeurs grisées représentent les partitions optimales qui retrouvent les structures histologiques principales du côlon humain normal.

	#	6	7	8	9	10	11	12	13	14	15	Taux de
	Nombre de structures histologiques	4	4	4	4	4	4	4	4	4	4	précision (%)
Indice	CI	359	348	338	363	357	332	326	341	319	324	100
	Dunn	124	258	264	328	329	267	258	159	19	323	100
	OS	98	385	390	387	393	389	321	396	390	390	100
	S_Dbw	360	211	344	376	373	362	356	390	370	357	100
	XB**	216	44	305	312	227	201	255	59	279	102	100
	PBM	12	16	12	19	9	12	20	14	11	20	90
	SI	20	30	13	18	13	24	20	22	14	33	90
	CS	47	179	36	53	33	22	310	334	4	348	90
	SV	137	220	37	46	59	34	304	343	4	355	90
	$Dunn_{12}$	49	8	14	17	12	6	8	18	4	4	60
	$Dunn_{13}$	57	8	14	17	12	6	8	24	4	4	60
	DB	6	11	12	7	4	4	8	7	9	12	20
	VRC	8	11	4	5	4	12	14	4	4	5	20
	ASSWC	5	9	5	4	4	4	6	4	4	41	0
	ASWC	4	8	4	4	4	4	6	6	4	6	0
	COP	6	7	5	5	4	7	6	4	5	5	0
	DB*	5	7	5	4	4	4	4	4	4	4	0
	DB**	4	6	17	4	4	6	9	4	4	4	0
	$Dunn_{21}$	5	4	4	5	4	4	5	4	4	4	0
	$Dunn_{22}$	5	4	4	4	4	4	4	4	4	4	0
	$Dunn_{23}$	5	4	4	4	4	4	4	4	4	4	0
	$Dunn_{31}$	5	5	4	5	4	4	4	4	4	4	0
	$Dunn_{32}$	5	4	4	4	4	4	4	4	4	4	0
	$Dunn_{33}$	5	4	4	4	4	4	4	4	4	4	0
	$Dunn_{41}$	5	7	4	5	4	4	4	4	4	4	0
	$Dunn_{42}$	5	4	4	4	4	4	4	4	4	4	0
	$Dunn_{43}$	5	4	4	4	4	4	4	4	4	4	0
	$Dunn_{51}$	5	5	4	5	4	4	4	4	4	4	0
	$Dunn_{52}$	5	4	4	4	4	4	4	4	4	4	0
	$Dunn_{53}$	5	4	4	4	4	4	4	4	4	4	0
	$Dunn_{61}$	5	4	8	5	4	4	4	5	4	4	0
	$Dunn_{62}$	5	4	4	4	4	4	4	5	4	4	0
	$Dunn_{63}$	5	4	4	4	4	4	4	5	4	4	0
	SSWC	5	8	5	4	4	4	6	4	4	6	0
	SWC	5	8	4	4	4	4	6	7	4	4	0
	XB	5	9	5	5	4	4	6	7	4	4	0
	XB*	4	4	4	4	4	4	4	4	4	4	0

Chapitre IV

Développement d'une histologie spectrale multi-images

Sommaire

IV.1 - Préambule 78

IV.2 - Article #2 : *"Fully unsupervised inter-individual spectral histology of paraffinized tissue sections of normal colon"* 81

IV.3 - Résultats supplémentaires 108

IV.3.1 - Confirmation de l'efficacité du protocole non-automatisé à l'échelle intra- et inter-individuelle chez l'Homme 108

IV.3.2 - Impact des paramètres de l'EMSC sur le nombre de classes estimé par le protocole automatisé 110

IV.1 - Préambule

Ce travail fait l'objet d'un manuscrit en révision dans *Journal of Biophotonics* :
T.N.Q. Nguyen, P. Jeannesson, A. Groh, O. Piot, D. Guenot, C. Gobinet, *"Fully unsupervised inter-individual spectral histology of paraffinized tissue sections of normal colon"*.

Contexte et objectif du projet

L'histologie spectrale sur échantillons paraffinés est classiquement effectuée image par image. C'est-à-dire que chaque image est traitée indépendamment par EMSC pour conduire à un déparaffinage numérique, puis par classification non-supervisée KM afin d'identifier les différentes structures tissulaires ; une couleur étant affectée aléatoirement à chaque classe.

Comme le montre la Figure 2 de l'article, cette approche image par image entraîne des inconvénients majeurs puisqu'une même structure histologique peut être représentée d'une image à une autre par :

- des couleurs différentes,
- des nombres de classes différents.

En physiopathologie, ces limitations compliquent fortement l'analyse comparative inter-images que ce soit à l'échelle d'un même individu ou entre plusieurs individus.

Dans ce travail, nous avons développé une histologie spectrale multi-images basée sur une analyse numérique conjointe de lots d'images spectrales. L'étude a été appliquée à des coupes tissulaires paraffinées provenant de côlon de souris et de patients. Cette approche a permis d'obtenir un code couleur commun à toutes les images. De plus, chaque structure histologique a pu être définie par un nombre de classes constant pour toutes les images.

Protocole développé pour une histologie spectrale multi-images

L'histologie spectrale multi-images a été réalisée en deux étapes successives sur des coupes tissulaires de côlon normal de souris et de patients. La séquence de ces deux

étapes est présentée sous forme de schéma dans la Figure 3 de l'article :

i) l'ensemble des images IR brutes est d'abord déparaffiné numériquement par une EMSC conjointe spécifiquement développée dans ce travail. Grâce à cette méthode, les spectres de ces images sont simultanément corrigés de la variation de la ligne de base et du signal de la paraffine, en utilisant un spectre de référence et un modèle de paraffine communs. Par conséquent, les spectres corrigés contiennent uniquement des informations d'ordre biomoléculaire spécifiques des tissus analysés ;

ii) ensuite, l'application de façon conjointe à toutes les images d'une classification non-supervisée KM, permet à toute structure histologique d'être identifiée par une même couleur au niveau de tous les échantillons étudiés. Pour cette méthode KM, le nombre de classes a été choisi de façon conventionnelle, c'est-à-dire empiriquement. Dans la suite du travail, ce protocole sera qualifié de "protocole non-automatisé".

Résultats

Le protocole non-automatisé a été appliqué à l'échelle intra-individuelle, sur un lot d'images spectrales IR acquises sur des zones tissulaires différentes, provenant d'une coupe paraffinée d'un individu donné. Contrairement aux résultats de l'histologie spectrale classique (Figure 2 de l'article), une même structure tissulaire de côlon murin (voire même une sous-structure) apparait avec la même couleur sur toutes les images (Figure 5 de l'article). Ces résultats ont été confirmés chez l'homme (Figure IV.1 des résultats supplémentaires).

Des résultats aussi concluants ont été obtenus à l'échelle inter-individuelle aussi bien chez la Souris (Figure 6 de l'article) que chez l'Homme (Figure IV.2 des résultats supplémentaires).

Afin d'automatiser cette histologie spectrale multi-images, la double application hiérarchique d'indices de validité développée dans le Chapitre III a été couplée à ce protocole. Pour ce faire, nous avons utilisé l'indice PBM précédemment identifié comme le plus performant pour l'histologie spectrale, et nous l'avons appliqué chez la Souris et chez l'Homme à l'échelle intra- et inter-individuelles (Figures 7-9 de l'article). Toutes

les structures histologiques sont parfaitement retrouvées. De plus, il est intéressant de constater que cette approche fournit des données biomoléculaires supplémentaires non accessibles en histologie conventionnelle. En effet, les centroïdes des classes des cryptes sont capables de donner des informations sur la nature des produits de sécrétion de ces structures glandulaires tels que les mucus (Figure 10 de l'article).

Les deux protocoles présentés dépendent de deux paramètres susceptibles d'affecter l'attribution des pixels aux classes :

i) le spectre de référence de l'EMSC sur lequel tous les spectres acquis seront "fittés". Dans ce travail, ce spectre a été calculé comme le spectre moyen soit d'une image choisie au hasard, soit de l'ensemble des images considérées.

ii) le modèle de paraffine utilisé pour neutraliser la variabilité de son signal IR au niveau des spectres acquis. Dans ce travail, ce modèle a été généré à partir des spectres de paraffine pure acquis soit sur une coupe tissulaire choisie au hasard, soit sur l'ensemble des coupes analysées.

Les résultats du Tableau 2 de l'article montrent que le protocole automatisé s'avère insensible au choix du spectre de référence. En revanche, il apparaît être dépendant du modèle de paraffine en raison de la forte variabilité du signal IR de la paraffine d'un échantillon à un autre. Afin de corriger cette variabilité, ce modèle doit être généré à partir des spectres de paraffine acquis sur toutes les coupes.

Enfin, dans la section des résultats supplémentaires de ce chapitre (page 110), nous montrons que le nombre de classes estimé par ce protocole automatisé est dépendant du modèle de paraffine et indépendant du spectre de référence.

IV.2 - Article #2 : *"Fully unsupervised inter-individual spectral histology of paraffinized tissue sections of normal colon"*

T.N.Q. Nguyen, P. Jeannesson, A. Groh, O. Piot, D. Guenot, C. Gobinet.

Article soumis à *Journal of Biophotonics*, le 26-10-2015.

Journal of Biophotonics

Fully unsupervised inter-individual IR spectral histology of paraffinized tissue sections of normal colon

--Manuscript Draft--

Manuscript Number:	
Full Title:	Fully unsupervised inter-individual IR spectral histology of paraffinized tissue sections of normal colon
Article Type:	Full Article
Section/Category:	Image Processing in Biomedical Diagnosis
Keywords:	FTIR imaging; automatic multi-image spectral histology; clustering; normal colon tissue
Corresponding Author:	Cyril Gobinet, Ph.D Universite de Reims Champagne-Ardenne Reims, FRANCE
Corresponding Author Secondary Information:	
Corresponding Author's Institution:	Universite de Reims Champagne-Ardenne
Corresponding Author's Secondary Institution:	
First Author:	Thi Nguyet Que Nguyen
First Author Secondary Information:	
Order of Authors:	Thi Nguyet Que Nguyen
	Pierre Jeannesson
	Audrey Groh
	Olivier Piot
	Dominique Guenot
	Cyril Gobinet, Ph.D
Order of Authors Secondary Information:	
Abstract:	In label-free FTIR histology, spectral images are individually recorded from tissue sections, preprocessed and clustered. Each single resulting color-coded image is annotated by a pathologist to obtain the best possible match with tissue structures revealed after Hematoxylin-Eosin staining. However, the main limitations of this approach are the empirical choice of the number of clusters and the marked color heterogeneity between the clustered spectral images. Here, using normal murine and human colon tissues, we developed an automatic multi-image spectral histology to simultaneously analyze a set of spectral images (8 images mice samples and 72 images human ones). This procedure consisted of a joint Extended Multiplicative Signal Correction (EMSC) to numerically deparaffinize the tissue sections, followed by an automated joint K-Means (KM) clustering using the hierarchical double application of PBM validity index. Using this procedure, the main murine and human colon histological structures were correctly identified at both the intra- and the inter-individual levels. Here, we show that batched multi-image spectral histology procedure is insensitive to the reference spectrum but highly sensitive to the paraffin model of joint EMSC. In conclusion, combining joint EMSC and joint KM clustering by double PBM application allows to achieve objective and automated batched multi-image spectral histology.
Additional Information:	
Question	**Response**

Please submit a plain text version of your cover letter here. Please note, if you are submitting a revision of your manuscript, there is an opportunity for you to provide your responses to the reviewers later; please do not add them to the cover letter.	Reims, 26th October 2015 Dear Editor, Following the invitation of the Guest Editors Robert KOPROWSKI and Thomas BOCKLITZ, we are pleased to submit for consideration in the special issue on “IMAGE PROCESSING IN BIOMEDICAL DIAGNOSIS” of the Journal of Biophotonics our manuscript entitled “Fully unsupervised inter-individual IR spectral histology of paraffinized tissue sections of normal colon” and co-authored by Thi Nguyet Que NGUYEN, Pierre JEANNESSON, Audrey GROH, Olivier PIOT, Dominique GUENOT, and Cyril GOBINET. This original research article describes the development of an automatic multi-image spectral histology able to simultaneously analyse a set of IR spectral images. This procedure is based on a joint EMSC to numerically dewax the studied tissue samples and an automated joint KM using the hierarchical double application of the PBM validity index. This batched multi-image spectral histology has been successfully applied on murine and human colon tissues. Original Article Statement: •This manuscript is the authors' original work and has not been published nor has it been submitted simultaneously elsewhere. •All authors have checked the manuscript and have agreed to the submission. Thank you for the kind attention you will give to this submission. Yours sincerely, Prof. Olivier PIOT & Dr. Cyril GOBINET

October 26, 2015

Abstract In label-free Fourier-transform infrared histology, spectral images are individually recorded from tissue sections, pre-processed and clustered. Each single resulting color-coded image is annotated by a pathologist to obtain the best possible match with tissue structures revealed after Hematoxylin-Eosin staining. However, the main limitations of this approach are the empirical choice of the number of clusters and the marked color heterogeneity between the clustered spectral images. Here, using normal murine and human colon tissues, we developed an automatic multi-image spectral histology to simultaneously analyze a set of spectral images (8 images mice samples and 72 images human ones). This procedure consisted of a joint Extended Multiplicative Signal Correction (EMSC) to numerically deparaffinize the tissue sections, followed by an automated joint K-Means (KM) clustering using the hierarchical double application of Pakhira-Bandyopadhyay-Maulik (PBM) validity index. Using this procedure, the main murine and human colon histological structures were correctly identified at both the intra- and the inter-individual levels, especially the crypts, secreted mucus, lamina propria and submucosa. Here, we show that batched multi-image spectral histology procedure is insensitive to the reference spectrum but highly sensitive to the paraffin model of joint EMSC. In conclusion, combining joint EMSC and joint KM clustering by double PBM application allows to achieve objective and automated batched multi-image spectral histology.

Fully unsupervised inter-individual IR spectral histology of paraffinized tissue sections of normal colon

Thi Nguyet Que Nguyen[1,2], Pierre Jeannesson[1,2], Audrey Groh[3], Olivier Piot[1,2], Dominique Guenot[3], and Cyril Gobinet[1,2,*]

1. Introduction

Currently, spectral histology appears increasingly as an alternative method for studying human tissues whether normal or pathological. Fourier-transform infrared (FTIR) spectral imaging, which allows characterization of histological structures

[1] Université de Reims Champagne-Ardenne, Equipe MéDIAN-Biophotonique et Technologies pour la Santé, UFR de Pharmacie, Reims (France) [2] CNRS UMR7369, Matrice Extracellulaire et Dynamique Cellulaire (MEDyC), Reims, (France) [3] Progression tumorale et microenvironnement, Approches translationnelles et Epidémiologie, EA 3430, Fédération de Médecine Translationnelle de Strasbourg (FMTS), Université de Strasbourg (UdS), Strasbourg (France)

[*] Corresponding author: e-mail: cyril.gobinet@univ-reims.fr

of a tissue, is an innovative, non-destructive, non-invasive and label-free diagnostic method. In oncology, this method can be used for discriminating between normal and cancerous tissues from different organs, such as brain [1], skin [2], colon [3, 4], lung [5, 6], breast [7], prostate [8] or cervix [9].

FTIR spectral histology is based on digital analysis of spectral images acquired from standard paraffinized or frozen tissue sections. An unsupervised classification such as K-Means (KM) is commonly used to identify the different tissue structures. A color-coded image is then obtained where each structure is randomly assigned a color.

In a spectral image, the baseline of each FTIR spectrum varies from one to the other. To correct this additive baseline effect and other unwanted variation effects, such as multiplicative scaling and interference effects, extended multiplicative signal correction (EMSC) was developed [10]. Thus, the baselines of recorded IR spectra are corrected independently for each image. EMSC has also been used for digital dewaxing process [11, 12]. During this step, the paraffin signal is kept constant for all the corrected spectra belonging to the same image. However, this signal differs for different images. Consequently, these non-uniform baselines and paraffin signals prevent a batched multi-image analysis with homogenized colors for the same structure between images. Further, the random color attribution of KM clusters, estimated independently on each spectral image, represents another source of heterogeneity. An identical biological structure will thus be represented by different colors on different images. In addition, KM is usually applied with a number of clusters empirically chosen by using i) a fixed number of clusters, regardless of the patient or the tissue zone analyzed or ii) a trial-and-error procedure which progressively increases the number of clusters until the estimated partition approximately matches the histological image as interpreted by the pathologist.

In order to solve these problems, in this study, we present an objective and automatic spectral histology based on the standardization of the FTIR image pre-processing steps and the application of crisp cluster validity indices (CCVI) [13–15]. This approach was applied on normal colon originating from wild-type mice and from areas identified from biopsies of patients with colon cancer.

2. Material and methods

2.1. Samples

Four formalin-fixed paraffin-embedded (FFPE) biopsies of normal colons were collected from 4 C57BL/6 mice. For validating our study on human tissues, 15 FFPE biopsies of normal parts from surgically removed colons were collected from 15 patients diagnosed with colon cancer. These normal parts were located in different zones of the colons: cecum, left

colon, right colon, sigmoid and rectosigmoid. The patient tissue set was composed of 9 female and 6 male patients, aged between 47 and 89 years-old (Table 1).

For each biopsy, two consecutive 6 μm thick tissue sections were used. The first was mounted on a calcium fluoride (CaF_2) window (Crystran, Dorset, UK) for FTIR imaging, without any particular preparation, especially no chemical dewaxing. The second was mounted on a glass window, stained by Hematoxylin-Eosin (HE) and used as reference to correlate with spectral histology by a pathologist.

2.2. FTIR image acquisition

FTIR imaging was performed with a Spectrum Spotlight 300 FTIR imaging system coupled to a Spectrum one FTIR spectrometer (Perkin Elmer, Courtabœuf, France). This system is equipped with a liquid nitrogen-cooled 16-element mercury cadmium telluride detector, calibrated for imaging in transmission mode. The imaging system and the sample compartment were continuously purged with dry air in order to reduce the impact of carbon dioxide and water vapor.

FTIR images were recorded on the spectral range 750 to 4000 cm^{-1}, with a spatial resolution of 6.25 μm, a spectral resolution of 4 cm^{-1} and averaged over 16 scans per pixel. A reference spectrum from the CaF_2 window was acquired prior to image acquisition with 240 accumulations and subtracted automatically from each recorded image by the Spectrum Image software (Perkin Elmer).

For each sample, up to 4 different tissue zones and 1 pure paraffin zone were imaged with these acquisition parameters. Totally, 72 FTIR images of paraffinized tissue and 15 of pure paraffin have been recorded. Each image contains between 5000 and 22000 pixels. Each pixel represents the absorbance recorded over 1626 wavenumbers uniformly distributed between 750 and 4000 cm^{-1}, defining a full spectrum.

2.3. Data processing

Data analysis were carried out using in-house scripts written in Matlab (The Mathworks, Natick, MA), unless otherwise specified.

2.3.1. Preprocessing

The IR spectral images have intrinsic undesirable signals that must be corrected by specific pre-processing steps in order to extract the informative signal coming from the biological material.

Atmospheric absorptions of water vapor and CO_2 were corrected on each FTIR image using the Spectrum Image software. The most informative spectral range for biological samples being the 900-1800 cm^{-1} [16], this fingerprint region was used for further numerical treatments.

After applying atmospheric correction and limiting the spectral range, the FTIR spectral images are mathematically defined as: $\mathbf{S^{t,n}} = \{\mathbf{s_l^{t,n}}, l = 1,\ldots,L_{t,n}\}, 1 \leq n \leq N_t, 1 \leq t \leq T$. T is the total number of analyzed tissue sections. N_t is the number of spectral images acquired on the t^{th} tissue section. $\mathbf{s_l^{t,n}} \in \mathbb{R}^{1\times D}$ is the l^{th} spectrum (containing D absorbance values) of the n^{th} FTIR image acquired on the t^{th} tissue section. $L_{t,n}$ is the number of spectra composing the n^{th} image acquired on the t^{th} tissue section. $N = \sum_{t=1}^{T} N_t$ is the total number of spectral images acquired on paraffinized tissue sections. The total number of spectra acquired on the N images is defined by $L = \sum_{t=1}^{T}\sum_{n=1}^{N_t} L_{t,n}$.

In order to correct the spectra from paraffin contribution and scattering effects, EMSC was used. EMSC is a linear pre-processing method developed to correct FTIR spectra from physical light scattering effects. The EMSC model has been extended by incorporating an orthogonal subspace in order to correct spectra from irrelevant effects [10], according to the following equation:

$$\mathbf{s_l^{t,n}} = a_l^{t,n}\hat{\mathbf{s}}^{\mathbf{t,n}} + \mathbf{b_l^{t,n}}\mathbf{I^t} + \mathbf{c_l^{t,n}}\mathbf{P} + \mathbf{e_l^{t,n}} \tag{1}$$

$\hat{\mathbf{s}}^{\mathbf{t,n}} \in \mathbb{R}^{1\times D}$ is the reference spectrum estimated as the mean spectrum of the n^{th} FTIR image acquired on the t^{th} tissue section. $\mathbf{I^t} \in \mathbb{R}^{M\times D}$ is the orthogonal subspace matrix composed of M components modeling the irrelevant effects of the t^{th} tissue section. $\mathbf{c_l^{t,n}P}$ is used to model the physical light scattering effects by a Q-order polynomial function, where $\mathbf{P} \in \mathbb{R}^{(Q+1)\times D}$ is the transpose of the Vandermonde matrix of the D wavenumbers. $\mathbf{e_l^{t,n}} \in \mathbb{R}^{1\times D}$ is the model error vector. $a_l^{t,n}$ is the regression coefficient of $\hat{\mathbf{s}}^{\mathbf{t,n}}$ to $\mathbf{s_l^{t,n}}$. $\mathbf{b_l^{t,n}} \in \mathbb{R}^{1\times M}$ and $\mathbf{c_l^{t,n}} \in \mathbb{R}^{1\times(Q+1)}$ are the regression coefficient vectors of $\mathbf{I^t}$ to $\mathbf{s_l^{t,n}}$ and $\mathbf{P}$ to $\mathbf{s_l^{t,n}}$, respectively. $a_l^{t,n}$, $\mathbf{b_l^{t,n}}$ and $\mathbf{c_l^{t,n}}$ are estimated by ordinary least squares in order to minimize the modeling error $\sum_{d=1}^{D}\left(e_{l,d}^{t,n}\right)^2$.

Once the model is estimated, the $\mathbf{s_l^{t,n}}$ spectrum is corrected by the following equation:

$$\tilde{\mathbf{s}}_\mathbf{l}^{\mathbf{t,n}} = a_l^{t,n}\hat{\mathbf{s}}^{\mathbf{t,n}} + \mathbf{e_l^{t,n}} \tag{2}$$

which can be seen as the fitting of the recorded spectrum $\mathbf{s_l^{t,n}}$ on the reference spectrum $\hat{\mathbf{s}}^{\mathbf{t,n}}$. $\mathbf{e_l^{t,n}}$ is thus the informative part of Eq. (2) since it models the biochemical differences between the mean spectrum and the l^{th} spectrum of the n^{th} FTIR image acquired on the t^{th} tissue section.

In order to correct the spectra from undesirable effects such as the effective optical path length, a normalization is applied on the corrected spectrum according to the following equation:

$$\bar{\tilde{\mathbf{s}}}_{\mathbf{l}}^{\mathbf{t,n}} = \hat{\mathbf{s}}^{\mathbf{t,n}} + \frac{\mathbf{e}_{\mathbf{l}}^{\mathbf{t,n}}}{a_l^{t,n}} = \frac{\tilde{\mathbf{s}}_{\mathbf{l}}^{\mathbf{t,n}}}{a_l^{t,n}} \tag{3}$$

The corrected FTIR spectral images are thus defined as $\bar{\tilde{\mathbf{S}}}^{\mathbf{t,n}} = \{\bar{\tilde{\mathbf{s}}}_{\mathbf{l}}^{\mathbf{t,n}}, l = 1, \ldots, L_{t,n}\}$.

Here, the FTIR images being recorded on FFPE samples without chemical dewaxing, the paraffin spectral contribution must be corrected. To this aim, a principal component analysis (PCA) was applied on a FTIR image acquired on a pure paraffin zone of the considered tissue section. Then the first 10 principal components (PCs) and the average spectrum of the FTIR paraffin image were retained to construct the orthogonal subspace matrix $\mathbf{I}^{\mathbf{t}}$ in Eq. (1) [12]. To model the physical light scattering effects, a 4-order polynomial function was used.

EMSC is very valuable since it accomplishes three tasks at the same time: i) it corrects spectra from scattering effects, ii) it neutralizes spectra from paraffin variability, iii) it normalizes spectra from optical path length.

2.3.2. KM clustering

After the EMSC pre-processing, a clustering method was applied on FTIR images for highlighting the colon tissue structures. In data mining, KM [17] is one of the most common unsupervised learning method. It aims to divide a given dataset of L observations in D dimensions into k given clusters, where each cluster is represented by the mean of its points, known as its centroid.

By definition, KM procedure starts by selecting randomly k points as initial centroids and executes iteratively two steps:

i) For all the points from the dataset, each point is assigned to the cluster whose centroid is the nearest using a chosen distance metric.

ii) The k centroids are updated according to the new partition computed in step i).

The above steps iteratively repeated until no point assignment changes. KM algorithm thus converges to a local minimum of the objective function, which is defined as the within-cluster sum of point-to-centroid distances.

The used distance metric is chosen as the Euclidean distance. To facilitate the visual analysis of KM clustering results, a color-coded image is generated where each spectrum belonging to a cluster can be exclusively represented by a single color. These images were then compared to the adjacent HE-stained tissue section by a pathologist, in order to assign each cluster to its corresponding histological structure.

2.3.3. Hierarchical double CCVI application

In data mining, to automatically estimate the optimal number of clusters k_{opt} for crisp clustering algorithm (such as KM), CCVIs have been developed. Theoretically, a validity index is a mathematical function that measures the quality of a clustering partition. By performing KM partitions with various values of k, CCVI calculates the ratio between the compactness and the separation of clusters for each partition. The k_{opt} is thus defined as the number of clusters giving the optimal CCVI value.

To perform an objective and automated spectral histology, we have recently proposed a hierarchical double application of CCVIs [18] for KM clustering. For each FTIR image, this method runs these following steps:

i) The CCVI is first applied on the KM results of spectral images, with $2 \leq k \leq 20$. The optimal number of clusters estimated by the CCVI is the number of main clusters k_{main} composing the dataset.

ii) Then, for each i^{th} main cluster, $1 \leq i \leq k_{main}$, KM is applied (with $2 \leq k \leq 20$) and a second application of the CCVI on these KM results estimates the optimal number of sub-clusters $k_{sub}(i)$.

iii) The final k_{opt} is thus the sum of the estimated optimal number of sub-clusters. The corresponding optimal double CCVI partition is obtained by assembling all the k_{main} estimated optimal sub-partitions together.

2.4. Classification accuracy rate

To compare two different clustering partitions estimated from the same dataset, a correspondence mapping m must be computed between their clusters. Each cluster i of the first partition is linked to the cluster $j = m(i)$ of the second partition with which it has the highest number of common pixels $H_{i,j} = H_{i,m(i)}$. Then, the similarity between the two partitions is measured by the classification accuracy rate τ [19]. This criterion is defined as the percentage of pixels which belong to the linked clusters: $\tau = 100 \times \frac{\sum_{i=1}^{k} H_{i,m(i)}}{L}$, where L is the number of pixels composing the dataset and k the number of clusters. τ is equal to 100% for identical partitions.

3. Results and discussion

IR spectral imaging experiments were performed on colon tissue sections either of C57BL/6 mice or of tumor-free mucosa of patients originating from surgically resected colorectal tumors.

3.1. Limitations of classical spectral histology of colon tissue

On spectral images recorded on FFPE mice colon tissue sections, IR spectral histology is usually performed by combining EMSC digital dewaxing with KM clustering [11, 20] (Figure 1). This approach is applied image by image, leading to random color-coded images and making their comparison difficult as shown in Figure 2. This figure presents 2 color-coded images (Figures 2B) acquired on 2 distinct areas of the same C57BL/6 murine colon tissue section (Figure 2A). Compared to their corresponding HE-stained image, data show a different cluster assignment for each color-coded image. For example, the muscular tunic structure is colored in brown in Figure 2B1 whereas it is in yellow in Figure 2B2.

In addition, with KM clustering, the number of clusters assigned to a specific tissue structure can vary from one image to another. For example, the crypts are represented by 3 and 6 clusters in Figure 2B1 and 2B2, respectively. The same problems occurred on images acquired from colon of several C57BL/6 (data not shown).

3.2. Development of a multi-image spectral histology of mice colon tissue

As shown in Figure 3A, to overcome these limitations, we developed a multi-image spectral histology consisting of analyzing simultaneously a set of spectral images by combining a joint EMSC procedure to a joint clustering step. Spectral images were taken from different areas either of a same colon tissue section or of different tissue sections from different mice.

3.2.1. Joint EMSC

Classically, for numerical dewaxing, the EMSC procedure is applied separately on each FTIR image of paraffinized tissue using its own reference spectrum $\hat{\mathbf{s}}^{\mathbf{t,n}}$ and paraffin model $\mathbf{I}^{\mathbf{t}}$ (see section 2.3). To correct the spectra of different images using a single model, joint EMSC was developed. This method simultaneously corrects the spectra of the set of N spectral images by using the same reference spectrum $\hat{\mathbf{s}}$ and the same paraffin PCA model $\mathbf{I}$. The joint correction is obtained by applying Eq. (1-3) using $\hat{\mathbf{s}}^{\mathbf{t,n}} = \hat{\mathbf{s}}$ and $\mathbf{I}^{\mathbf{t}} = \mathbf{I}$, whatever $1 \leq n \leq N_t$ and $1 \leq t \leq T$. Thus, the corrected spectra of the N images present the same baseline and paraffin contributions, while keeping their specific biochemical information.

As shown in Figure 3A, $\mathbf{I}$ was generated by a PCA applied on the paraffin images acquired from all tissue sections, and $\hat{\mathbf{s}}$ by averaging the spectra of all the tissue section images. Figure 4A presents corrected spectra obtained after joint EMSC performed on spectral images recorded on 2 different zones of the same paraffinized murine colon tissue section. The corrected spectra of each image, in blue and red, have the same baseline and paraffin contributions. These data demonstrate

that joint EMSC removes the baseline and paraffin bias inherent to conventional EMSC (Figure 4B). After applying this joint EMSC procedure, preprocessed data were then subjected to joint KM clustering.

3.2.2. Conventional joint KM clustering

In our study, KM was implemented by simultaneously processing the whole spectra of all images; this method is referred to here as joint KM clustering. This was performed using either conventional or automatic joint KM clustering. Interestingly, k centroids estimated by joint KM clustering are common to all the analyzed images, contrary to classical spectral histology which computes different k centroids for each image.

When applying conventional joint KM clustering (Figure 3B1), the set of spectral images was partitioned into k clusters as described in section 2.3, with k chosen empirically. At the intra-individual level, this was first applied to 2 distinct areas for the same murine colon tissue section using $k = 10$ in accordance to our previous results [4]. As shown in Figure 5, the 2 reconstructed color-coded images present the same color code for each histological structure. For example, the lamina propria of both images are represented in light green, or the muscular tunics in aqua.

Secondly, data obtained at the inter-individual level are presented in Figure 6 using $k = 8$. For a set of FTIR images recorded on 4 different mice, each tissue structure is found to be represented by the same colors, *e.g.* the crypts were coded in 3 colors, dark orange, dark blue and light blue.

3.2.3. Automated joint KM clustering

Besides the above conventional approach which is completely subjective, we developed an automated joint KM clustering (Figure 3B2). This allows to automatically estimate the number of clusters k necessary to access the common data structure of the whole set of spectral images. This approach is based on the recently described hierarchical double application of CCVIs [18]. Pakhira-Bandyopadhyay-Maulik (PBM) [21] was selected here as the most effective validity index that exactly matches the main colon tissue structures [18]. The efficiency of the automated joint KM clustering using PBM is demonstrated in Figure 7 in which the main colon histological structures are perfectly revealed both at the intra-individual and inter-individual levels.

3.3. Application of automated multi-image spectral histology to human colon tissue

Joint EMSC and automated joint KM clustering were applied on FTIR images recorded on normal human colon tissue sections. At the intra-individual level, Figure 8 shows 4 reconstructed color-coded images acquired on distinct areas of

the same colon tissue section of individual #1. These images present the same color code for each tissue structure with an estimated number of clusters $k_{opt} = 11$. Each cluster can be easily attributed to corresponding histological structures including crypts (3 clusters), lamina propria (4 clusters), muscularis mucosae (1 cluster) and submucosa (3 clusters).

At the inter-individual level, this approach was applied by simultaneously studying spectral images acquired on colon of 15 different individuals. Interestingly, the colon tissue organization presents the same color code ($k_{opt} = 9$) for all the individuals tested (Figure 9). This is consistent with the fact that the spectral information stemming from a given histological structure is unequivocally homogeneous between individuals.

In a previous FTIR spectral imaging study [20] performed on human colon tissue, we showed that a specific cluster can be assigned to the mucus present inside the crypt. This was possible as glycoprotein mucins, which are the main components of the mucus, display characteristic bands at 1044, 1076 and 1125 cm^{-1}. Here, we investigated which of the 5 crypt clusters C1-5 can be assigned to the mucus. Figure 10 presents the second derivatives of the centroids of these 5 crypt clusters estimated from the dataset of the 15 individuals. Three characteristic mucin bands can be mainly found in cluster C1 which, consequently, can be assigned to the secreted mucus.

In conclusion, our results clearly demonstrate the efficiency of combining the joint EMSC and the automated joint KM clustering, both at the intra- and inter-individual levels.

3.4. Sensitivity of the batched multi-image spectral histology to the reference spectrum $\hat{\mathbf{s}}$ *and the paraffin model* $\mathbf{I}$ *of the joint EMSC*

As previously reported [12, 22, 23], EMSC for digital dewaxing is usually applied independently on each spectral image. Consequently, the reference spectrum $\hat{\mathbf{s}}^{\mathbf{t,n}}$ is computed as the mean spectrum of the considered image, and the paraffin model $\mathbf{I}^{\mathbf{t}}$ is constructed using the paraffin spectral image acquired on a non tissular area of the same histological slice.

On the contrary, our proposed joint EMSC procedure corrects simultaneously all the spectra of all images by using the same reference spectrum $\hat{\mathbf{s}}$ and the same paraffin model $\mathbf{I}$ as described in section 3.2.1. Here, $\hat{\mathbf{s}}$ was computed as the mean spectrum of all the spectral images and $\mathbf{I}$ as the PCA model generated by using the paraffin images acquired from all the tissue sections (Figure 3A). However, due to the multi-image nature of our approach, multiple values can be assigned to $\mathbf{I}$ and $\hat{\mathbf{s}}$. Indeed, each of these parameters can be constructed from any of the 15 tissue sections or from the whole tissue section set.

To evaluate the influence of $\mathbf{I}$ on the spectral histology, the complete procedure described in Figure 3A was run for the different $\mathbf{I}$ settings while $\hat{\mathbf{s}}$ is kept constant at the mean spectrum of all the tissue sections. Table 2 (upper part outlined in

red) presents the different classification accuracy rates τ (see section 2.4) calculated for each couple of **I** and $\hat{\mathbf{s}}$ parameters. Data show that the mean of τ is equal to $86 \pm 7.6\%$, indicating a high variability of paraffin IR spectral signatures between the individual tissue sections. In order to take this variability into account, **I** has to be generated from the whole paraffin IR images of all patients.

Next, the influence of $\hat{\mathbf{s}}$ was analyzed while **I** was modeled from whole paraffin IR images as stated above. Interestingly in this case, the mean of τ values is equal to $97 \pm 1.9\%$ (Table 2, lower part outlined in green). These results provide a direct experimental evidence that the EMSC model is independent of the choice of $\hat{\mathbf{s}}$, as previously mentioned [24].

4. Conclusion

In this work, beyond the classic mono-image IR spectral histology, developing combined joint EMSC and joint KM allows to obtain simultaneous batched multi-image IR spectral histology. Combined to hierarchical double CCVI application, this procedure becomes fully unsupervised and objective since the number of clusters is automatically determined. The efficacy of this method is shown on IR spectral images acquired on normal colon tissue sections, at the intra- and inter-individual levels. The main normal colon histological structures are correctly identified, including the secreted mucus, for the whole spectral dataset. Finally, we show that the EMSC digital dewaxing is insensitive to the reference spectrum but highly sensitive to the paraffin model.

Acknowledgements. The authors are grateful for financial support from Cancérople Grand-Est, Ligue contre le Cancer, the URCA technological platform of cellular and tissular imaging PICT-IBiSA, Région Champagne-Ardenne, Région Alsace and Ministère de lEnseignement Suprieur et de la Recherche. We thank Sylvie Ricord for linguistic assistance.

References

[1] N. Bergner, B. F. M. Romeike, R. Reichart, R. Kalff, C. Krafft, and J. Popp, Analyst **138**(14), 3983–3990 (2013).

[2] E. Ly, O. Piot, A. Durlach, P. Bernard, and M. Manfait, Analyst **134**(6), 1208–1214 (2009).

[3] A. Kallenbach-Thieltges, F. Großerüschkamp, A. Mosig, M. Diem, A. Tannapfel, and K. Gerwert, Journal of Biophotonics **6**(1), 88–100 (2013).

[4] J. Nallala, M. D. Diebold, C. Gobinet, O. Bouché, G. D. Sockalingum, O. Piot, and M. Manfait, Analyst **139**(16), 4005–4015 (2014).

[5] B. Bird, M. Miljković, S. Remiszewski, A. Akalin, M. Kon, and M. Diem, Laboratory Investigation **92**(9), 1358–1373 (2012).

[6] F. Großerueschkamp, A. Kallenbach-Thieltges, T. Behrens, T. Brüning, M. Altmayer, G. Stamatis, D. Theegarten, and K. Gerwert, Analyst **140**(7), 2114–2120 (2015).

[7] A. Benard, C. Desmedt, M. Smolina, P. Szternfeld, M. Verdonck, G. Rouas, N. Kheddoumi, F. Rothé, D. Larsimont, C. Sotiriou, and E. Goormaghtigh, Analyst **139**(5), 1044–1056 (2014).

[8] J. T. Kwak, A. Kajdacsy-Balla, V. Macias, M. Walsh, S. Sinha, and R. Bhargava, Scientific Reports **5**, 8758 (2015).

[9] J. Einenkel, U. D. Braumann, W. Steller, H. Binder, and L. C. Horn, Histopathology **60**(7), 1084–1098 (2012).

[10] N. K. Afseth and A. Kohler, Chemometrics and Intelligent Laboratory Systems **117**, 92–99 (2012).

[11] J. Nallala, C. Gobinet, M. D. Diebold, V. Untereiner, O. Bouché, M. Manfait, G. D. Sockalingum, and O. Piot, Journal of Biomedical Optics **17**(11), 116013 (2012).

[12] E. Ly, O. Piot, R. Wolthuis, A. Durlach, P. Bernard, and M. Manfait, Analyst **133**(2), 197–205 (2008).

[13] M. Halkidi, Y. Batistakis, and M. Vazirgiannis, Journal of Intelligent Information Systems **17**(2/3), 107–145 (2001).

[14] L. Vendramin, R. J. G. B. Campello, and E. R. Hruschka, Statistical Analysis and Data Mining **3**(4), 209–235 (2010).

[15] O. Arbelaitz, I. Gurrutxaga, J. Muguerza, J. M. Pérez, and I. Perona, Pattern Recognition **46**(1), 243–256 (2013).

[16] M. Khanmohammadi, A. Bagheri Garmarudi, S. Samani, K. Ghasemi, and A. Ashuri, Pathology & Oncology Research **17**(2), 435–441 (2011).

[17] J. MacQueen, Some methods for classification and analysis of multivariate observations, in: Proceedings of the fifth Berkeley symposium on mathematical statistics and probability, (1967), pp. 281–297.

[18] T. N. Q. Nguyen, P. Jeannesson, A. Groh, D. Guenot, and C. Gobinet, Analyst **140**(7), 2439 – 2448 (2015).

[19] M. Meila, Journal of multivariate analysis **98**(5), 873–895 (2007).

[20] A. Travo, O. Piot, R. Wolthuis, C. Gobinet, M. Manfait, J. Bara, M. E. Forgue-Lafitte, and P. Jeannesson, Histopathology **56**(7), 921–931 (2010).

[21] M. K. Pakhira, S. Bandyopadhyay, and U. Maulik, Pattern Recognition **37**(3), 487–501 (2004).

[22] J. Nallala, G. R. Lloyd, and N. Stone, Analyst **140**(7), 2369–2375 (2015).

[23] R. Wolthuis, A. Travo, C. Nicolet, A. Neuville, M. P. Gaub, D. Guenot, E. Ly, M. Manfait, P. Jeannesson, and O. Piot, Analytical Chemistry **80**(22), 8461–8469 (2008).

[24] A. Kohler, N. K. Afseth, and H. Martens, Handbook of Vibrational Spectroscopy (2010).

Table 1 Characteristics of the human colon samples. The symbol # refers to the individual sample number. Abbreviations: M = Male, F = Female.

#	Age	Sex	Colon zone
1	62	F	Cecum
2	74	M	Recto-Sigmoid
3	73	F	Right colon
4	47	M	Cecum
5	72	F	Recto-Sigmoid
6	71	M	Sigmoid
7	71	M	Left colon
8	89	F	Cecum
9	74	F	Sigmoid
10	56	M	Recto-Sigmoid
11	82	F	Sigmoid
12	76	F	Recto-Sigmoid
13	74	M	Cecum
14	76	F	Sigmoid
15	75	F	Right colon

Table 2 Classification accuracy rates between partitions estimated using different paraffin models $\mathbf{I}$ and reference spectra $\hat{\mathbf{s}}$. The τ values of the upper part (outlined in red) have been calculated by varying $\mathbf{I}$ models and keeping $\hat{\mathbf{s}}$ constant, while those of the lower part (outlined in green) by varying $\hat{\mathbf{s}}$ and keeping $\mathbf{I}$ models constant. The symbol # refers to the individual sample number.

#	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1		94	92	78	75	78	79	79	79	73	78	78	77	75	76
2	97		89	78	75	78	79	79	79	73	78	78	77	75	77
3	98	95		76	75	77	76	77	76	71	77	77	76	76	74
4	98	99	96		88	95	94	93	95	88	94	95	92	88	92
5	95	98	93	97		88	86	86	85	84	89	88	89	95	89
6	95	98	93	97	99		96	95	95	88	97	97	94	89	92
7	98	97	96	99	96	96		95	96	88	95	95	93	87	92
8	97	95	99	96	93	93	97		95	85	94	94	91	87	90
9	100	97	97	99	95	96	98	97		87	94	95	92	87	91
10	99	98	97	99	96	96	98	97	100		87	89	88	86	89
11	100	97	98	99	95	95	98	97	100	99		95	94	88	94
12	97	94	99	96	92	92	95	98	97	96	97		93	90	91
13	100	98	97	99	96	96	98	97	100	100	100	97		90	94
14	100	97	98	99	95	95	98	97	100	99	100	97	100		89
15	96	99	94	97	99	99	96	93	96	97	96	93	96	96	

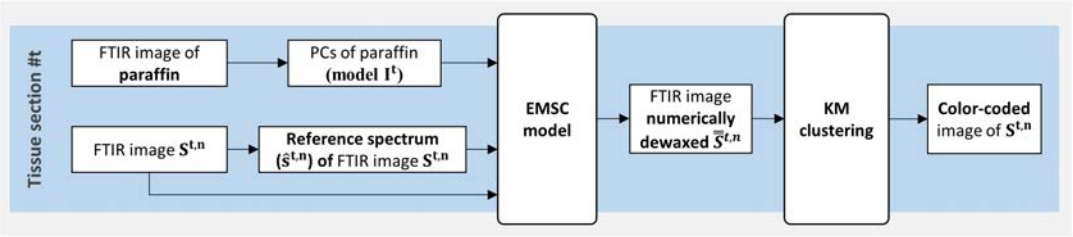


Figure 1 Flowchart of classical IR spectral histology on a paraffinized tissue section. EMSC preprocessing and KM clustering are independently applied on the tissue section by using two FTIR images acquired both on the biological tissue and on the pure paraffin.

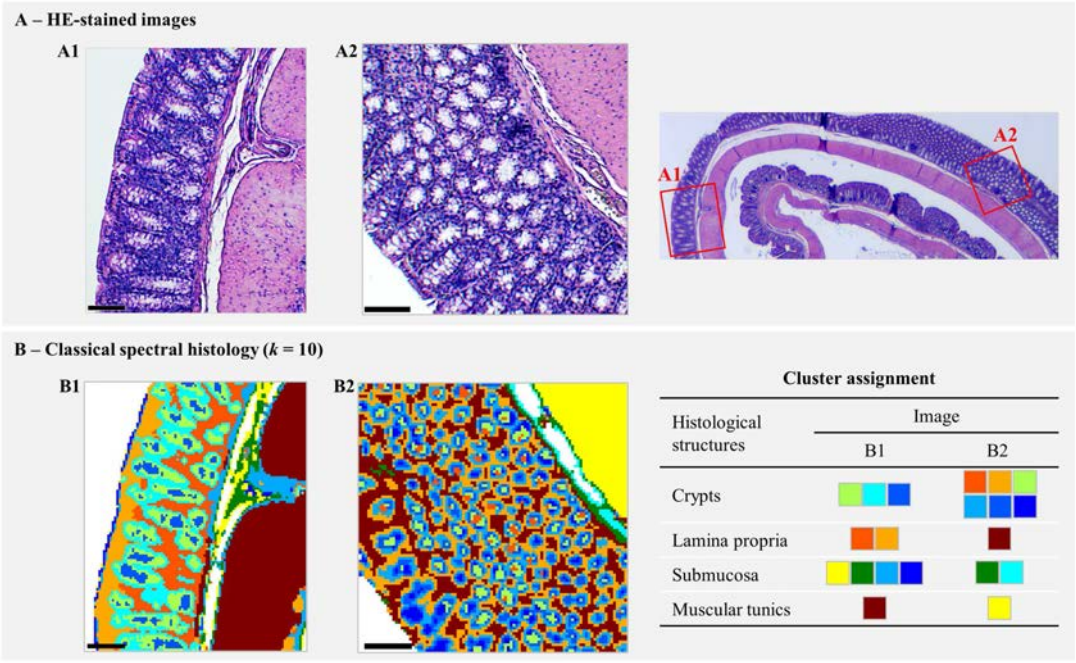


Figure 2 Comparison between conventional histology and color-coded images originating from the same murine colon tissue section. (A) HE-stained images; localizations of the examined tissue areas are presented in the right panel. (B) Color-coded images reconstructed from KM partitions ($k = 10$ clusters) of 2 FTIR images. Scale bars indicate 100 μm.

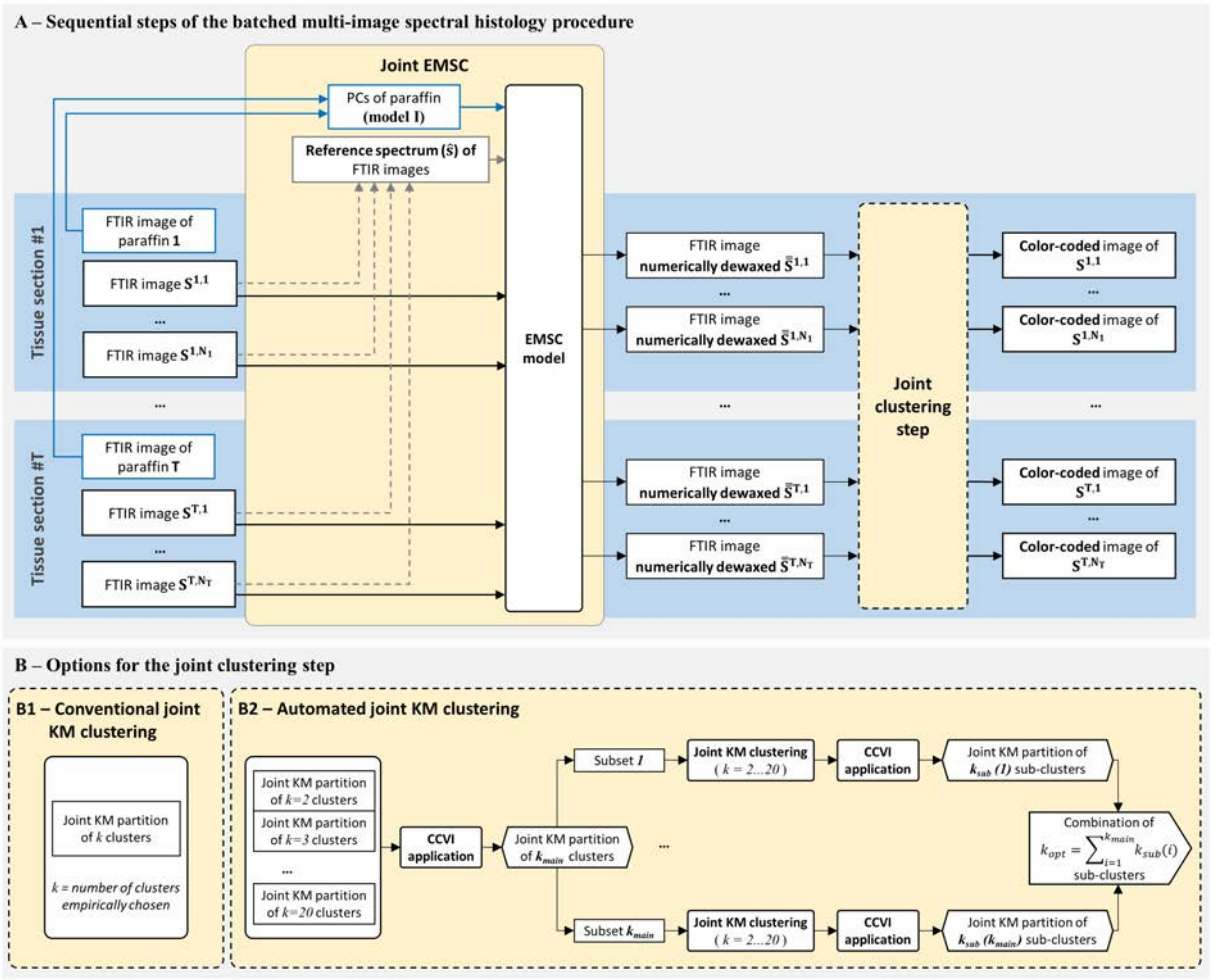


Figure 3 Flowchart of the batched multi-image spectral histology procedure applied to a set of N FTIR images. (A) PCs of paraffin and reference spectrum of paraffinized tissue were first generated, then followed by a joint EMSC preprocessing and a joint clustering step. (B) Details of the two possible options for the clustering step: conventional (B1) or automated (B2) joint KM.

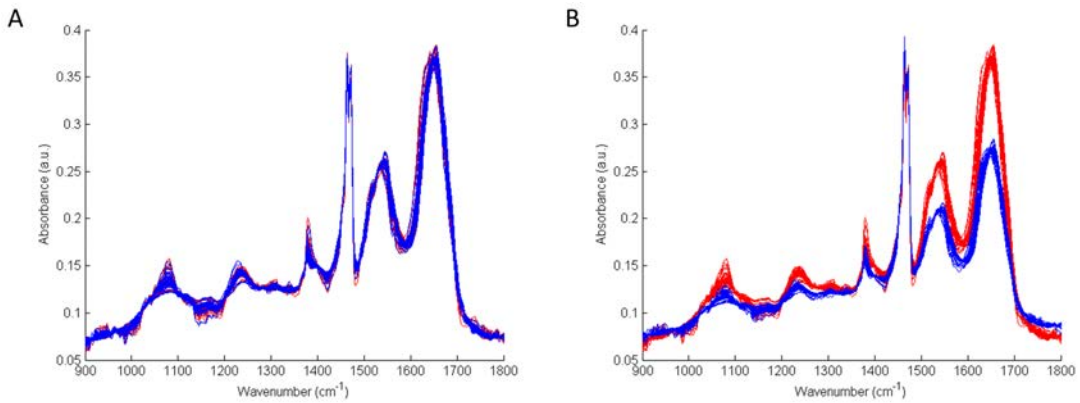


Figure 4 Comparison of corrected spectra after applying joint (A) and conventional (B) EMSC preprocessing. The red and blue spectra belong respectively to two different FTIR images acquired on the same murine tissue section.

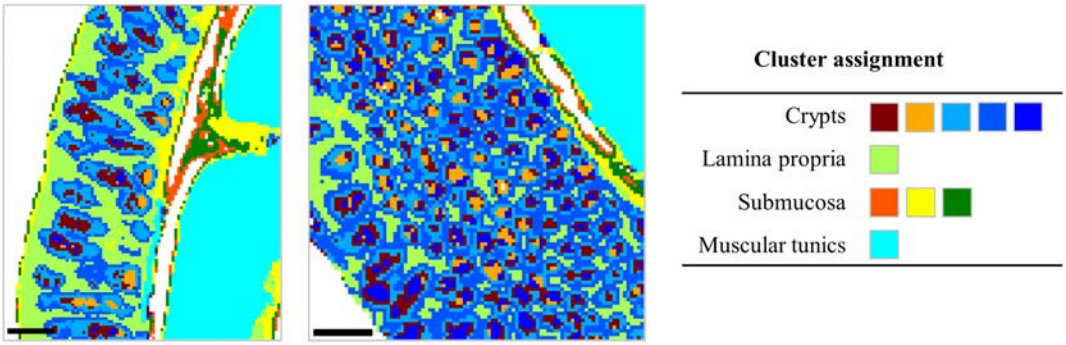


Figure 5 Color-coded images reconstructed from one single partition estimated by conventional joint KM ($k = 10$ clusters). Both images originate from the same murine colon tissue section. Corresponding HE images are shown in Figure 2. Scale bars indicate 100 μm.

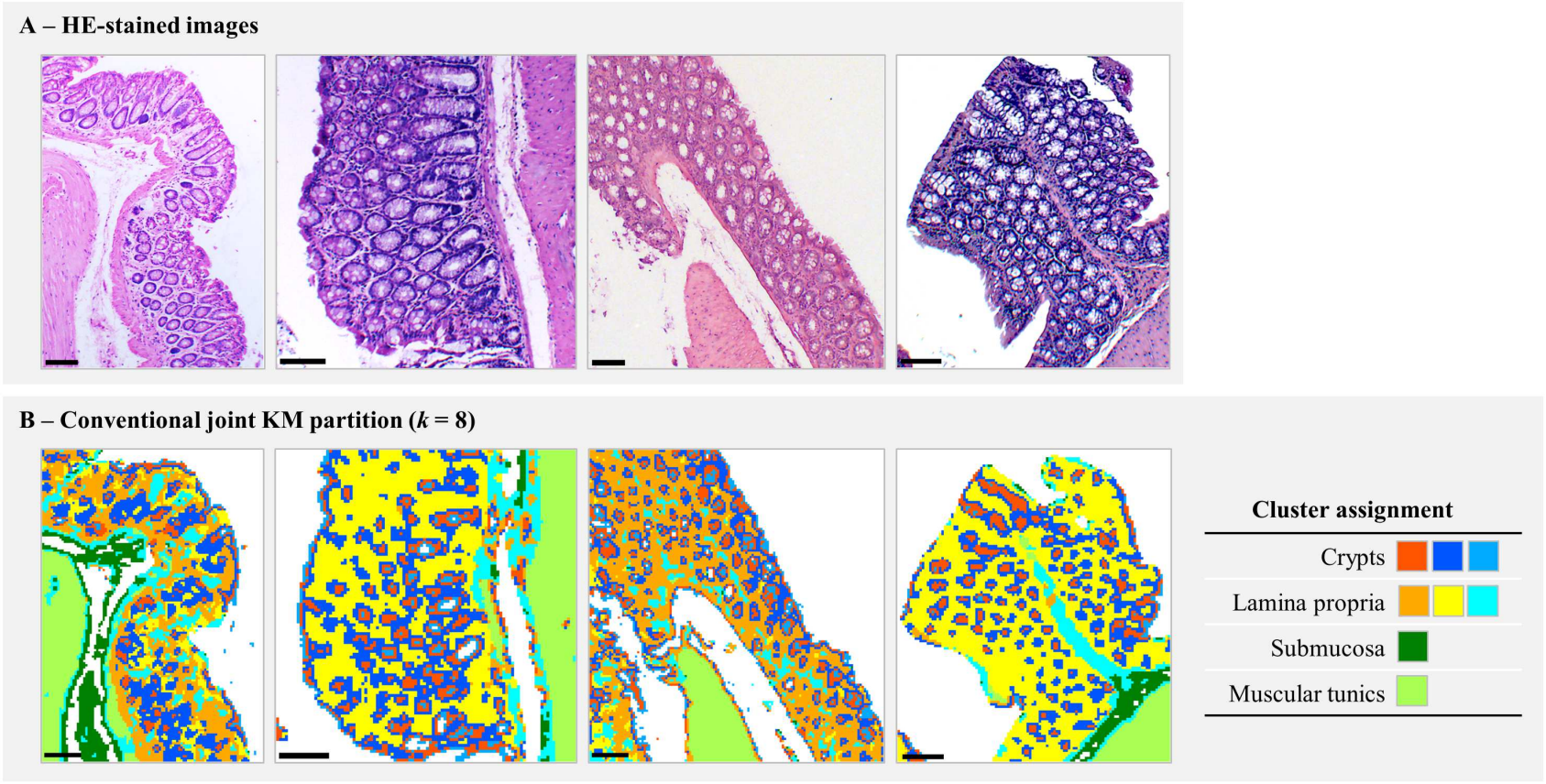


Figure 6 Comparison between conventional histology and color-coded images originating from murine colon tissue sections of four mice. (A) HE-stained images. (B) Color-coded images reconstructed from one single partition estimated by conventional joint KM ($k = 8$ clusters). Scale bars indicate 100 μm.

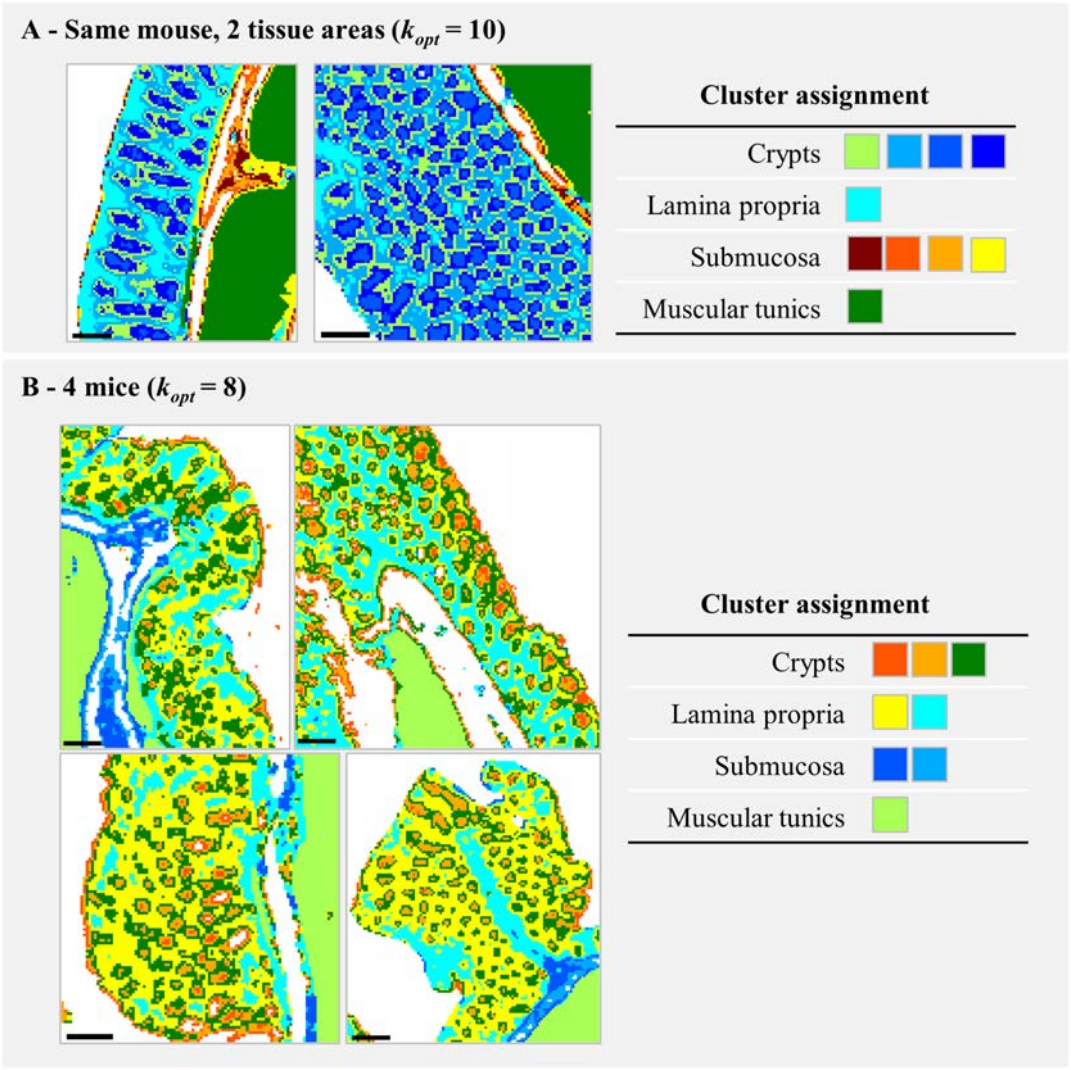


Figure 7 Color-coded images reconstructed from partitions estimated by automated joint KM using PBM index at (A) the intra-murine and (B) inter-murine level. Corresponding HE images of (A) and (B) are shown in Figures 2 and 6, respectively. Scale bars indicate 100 μm.

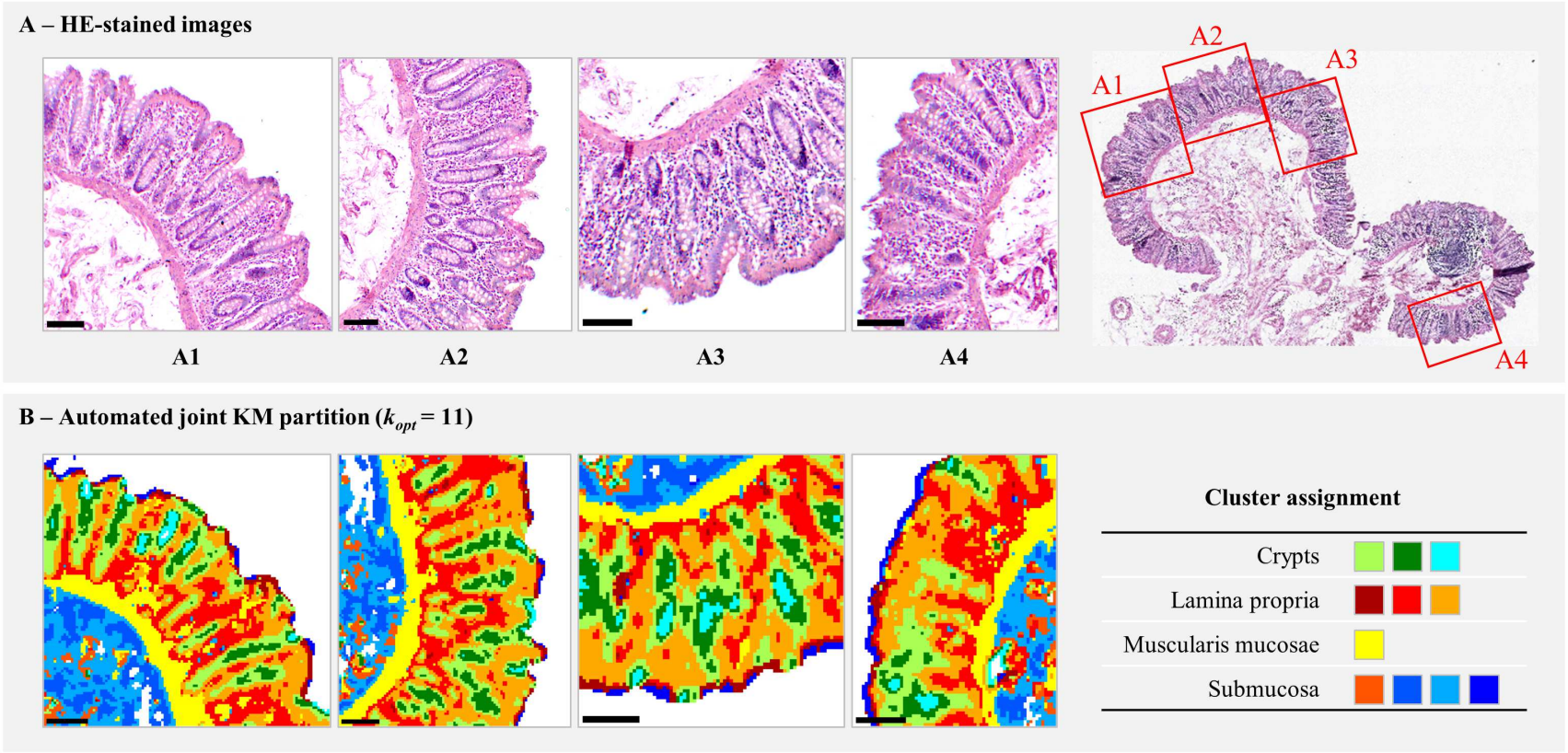


Figure 8 Comparison between conventional histology and color-coded images originating from patient #1. (A) HE-stained images; localizations of the examined tissue areas are presented in the right panel. (B) Color-coded images reconstructed from one single partition estimated by automated joint KM clustering using PBM ($k_{opt} = 11$ clusters); the images originate from the same colon tissue section. Scale bars indicate 100 μm.

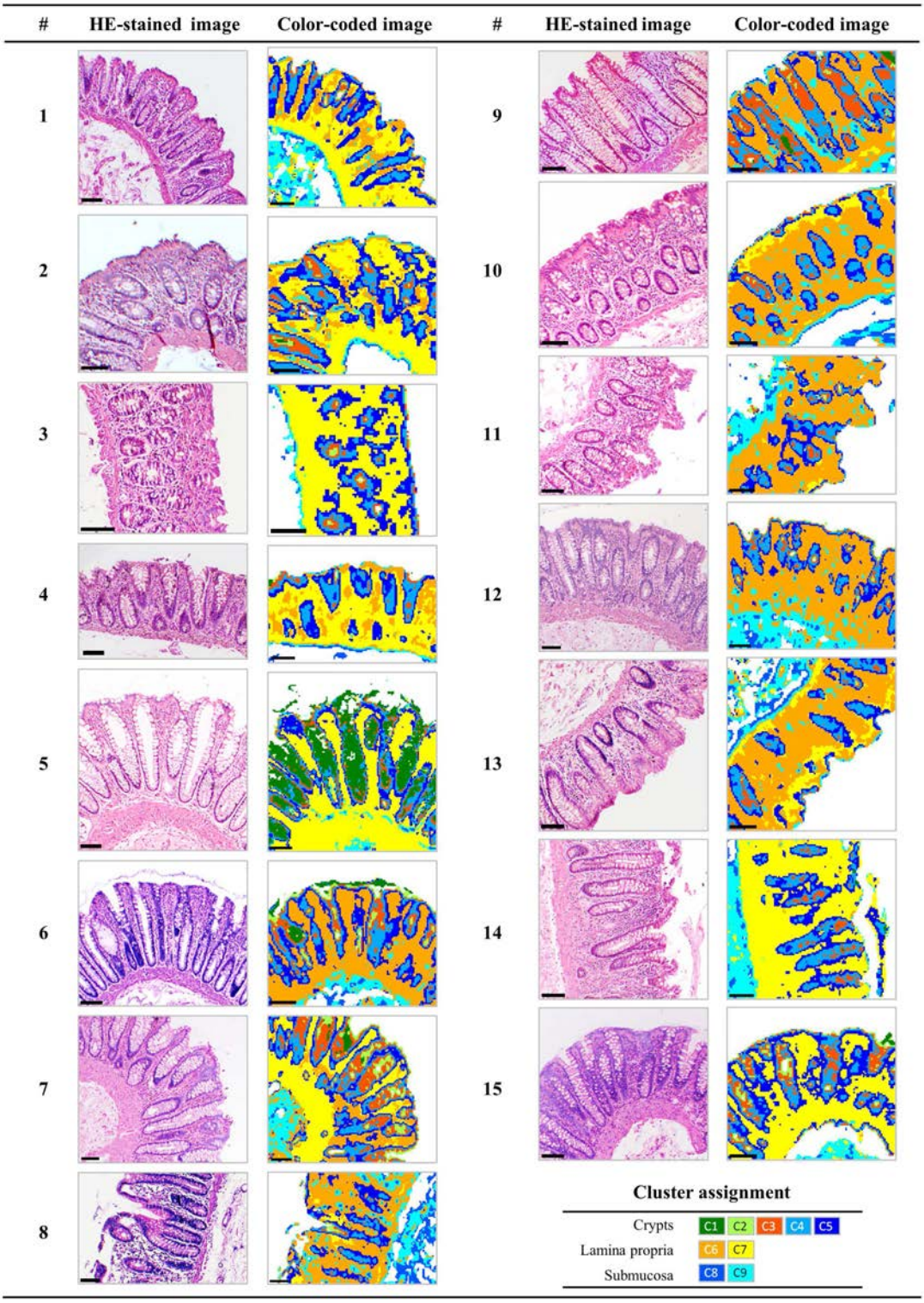


Figure 9 Comparison between conventional histology and color-coded images originating from colon tissue sections of 15 human individuals. Color-coded images were reconstructed from one single partition estimated by automated joint KM using PBM ($k_{opt} = 9$ clusters). The symbol # refers to the individual sample number. Scale bars indicate 100 μm.

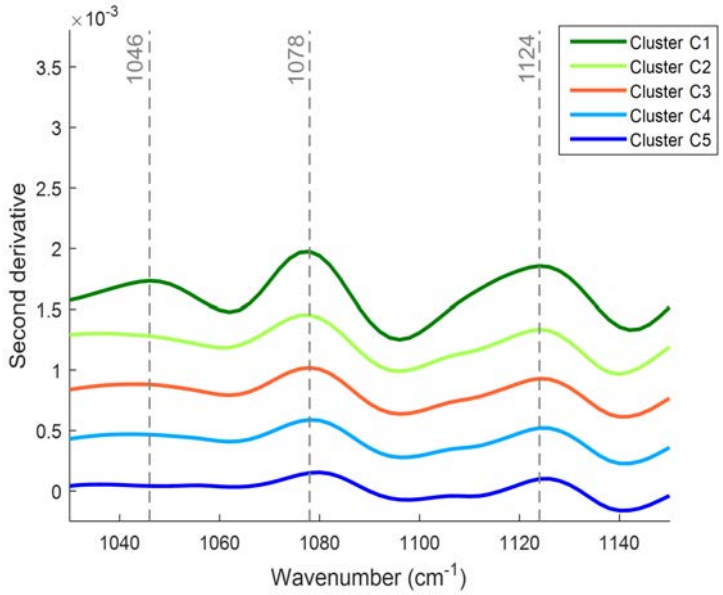


Figure 10 Second derivative of the five crypt cluster centroids from the partition presented on Figure 9. The characteristic bands of mucins are highlighted by vertical dashed lines.

IV.3 - Résultats supplémentaires

IV.3.1 - Confirmation de l'efficacité du protocole non-automatisé à l'échelle intra- et inter-individuelle chez l'Homme

Préalablement à son automatisation, nous avons testé le protocole non-automatisé d'histologie spectrale à l'échelle intra-individuelle (Figure IV.1) et inter-individuelle (Figure IV.2) chez l'Homme.

A l'échelle intra-individuelle, la Figure IV.1 montre qu'une même structure (voire une même sous-structure) tissulaire du côlon humain est affectée à une même couleur sur toutes les images. Des résultats similaires ont été également obtenus à l'échelle inter-individuelle (Figures IV.2), en utilisant un lot de 72 images provenant des 15 patients.

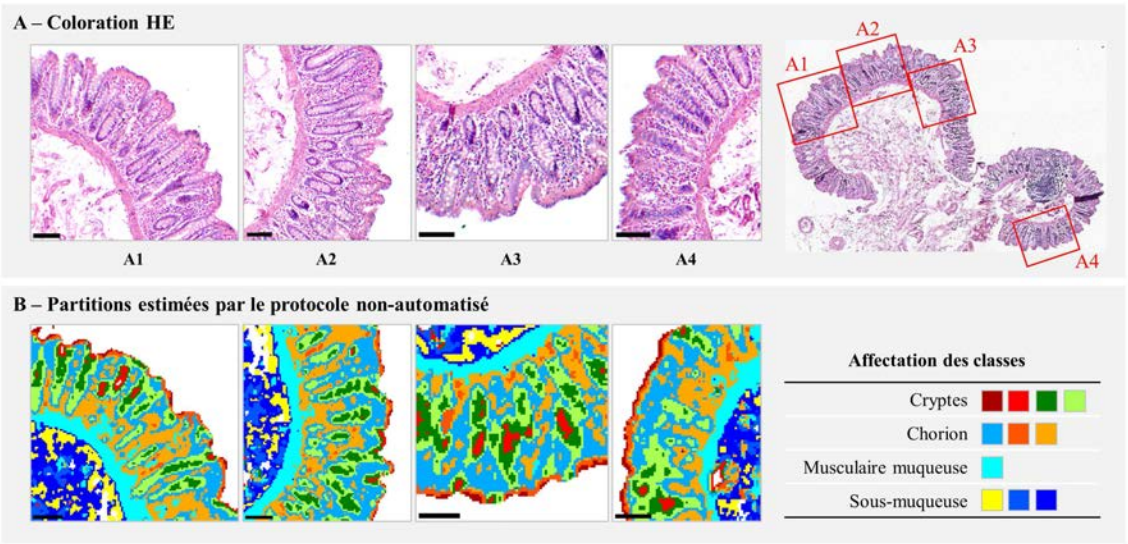


FIGURE IV.1 – Comparaison entre l'histologie conventionnelle et l'histologie spectrale multi-images à l'échelle intra-individuelle pour le patient #1. (A) Coloration HE. (B) Images en pseudo-couleurs reconstruites pour un nombre de classes $k = 11$ choisi empiriquement. Les images ont été acquises sur différentes zones d'un même échantillon. Échelles, 100 μm.

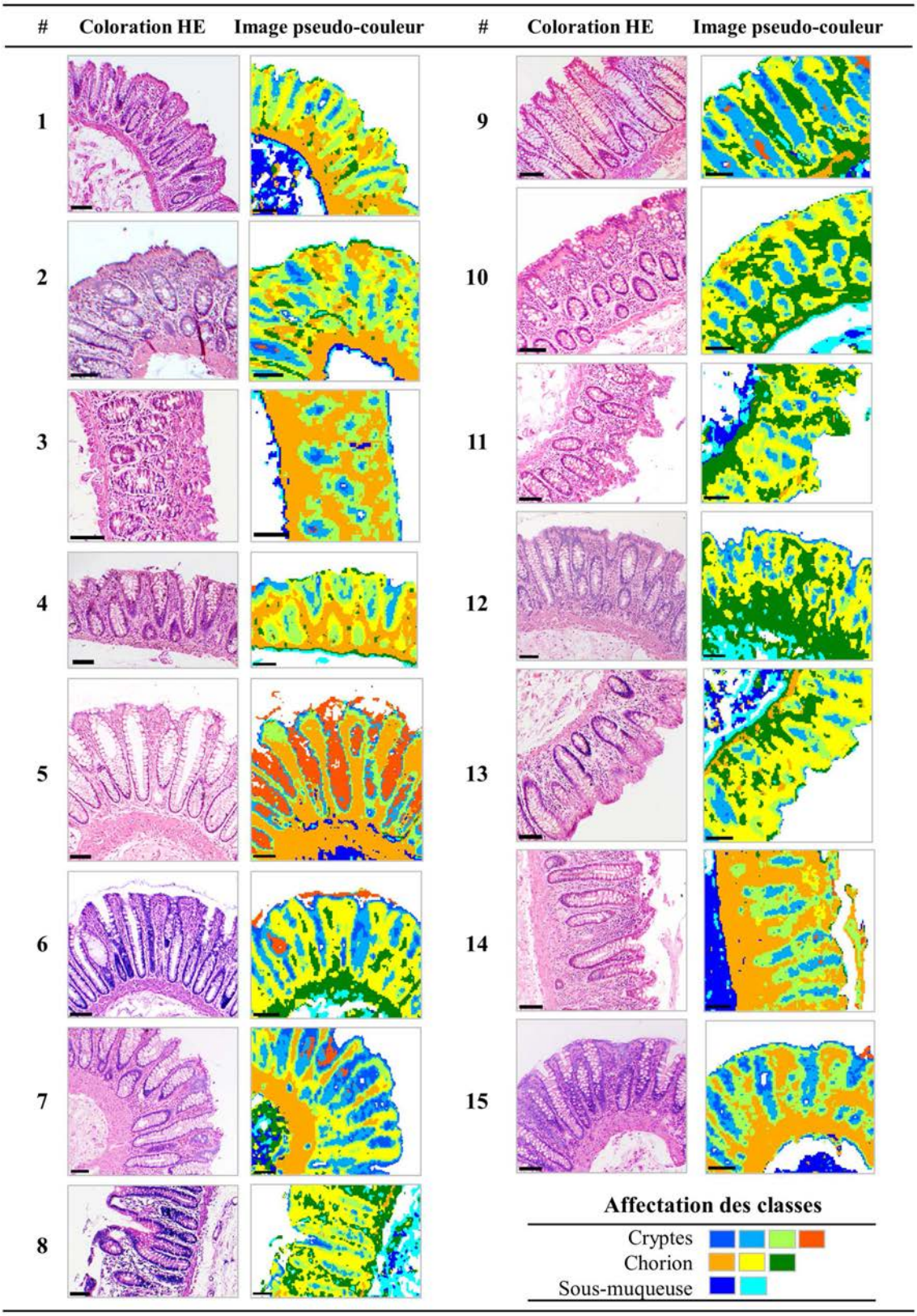


FIGURE IV.2 – Comparaison entre l'histologie conventionnelle et l'histologie spectrale multi-images à l'échelle inter-individuelle pour le lot de 72 images des 15 patients. Le nombre de classes $k = 9$ a été choisi empiriquement. Échelles, 100 μm.

IV.3.2 - Impact des paramètres de l'EMSC sur le nombre de classes estimé par le protocole automatisé

Dans l'article, nous avons montré que le taux de fidélité (classification accuracy rate τ) entre des partitions estimées par le protocole automatisé est très variable en fonction du modèle de paraffine mais indépendant du spectre de référence. Une explication possible serait que le modèle de paraffine et le spectre de référence impactent différemment l'estimation automatique du nombre de classes. Dans le but de répondre à cette question, nous avons étudié la variabilité du nombre de classes estimé en fonction du modèle de paraffine et du spectre de référence.

Dans un premier temps, comme le montre le schéma de la Figure IV.3, la double application de l'indice PBM a été répétée 15 fois (car 15 patients) sur les 72 images spectrales, en changeant de patient à chaque répétition pour construire le modèle de paraffine. Quant au spectre de référence, il est calculé comme spectre moyen des 72 images et reste donc constant. Les résultats du Tableau IV.1 (A) montrent que le nombre de classes varie entre 6 et 11.

Dans un second temps, la même procédure a été utilisée mais en changeant de patient à chaque répétition pour calculer le spectre de référence, tout en gardant constant le modèle de paraffine construit à partir de toutes les images spectrales de paraffine des 15 patients (Figure IV.4). Le Tableau IV.1 (B) montre que k_{opt} reste stable à 9 classes dans la majorité des cas (13/15).

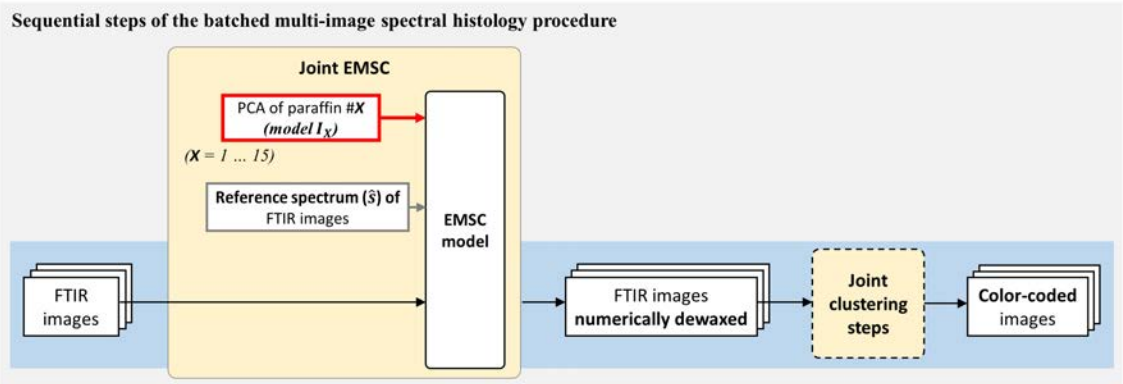


FIGURE IV.3 – Protocole automatisé avec application successive du modèle de paraffine I_X de chaque patient.

TABLE IV.1 – Nombre de classes k_{opt} estimé par le protocole automatisé pour (A) un modèle de paraffine variable et (B) un spectre de référence variable.

(A)

Modèle de paraffine du patient #	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
k_{opt}	6	6	8	9	9	10	9	10	9	9	10	10	11	9	9

(B)

Spectre de référence du patient #	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
k_{opt}	9	9	9	9	9	10	9	10	9	9	9	9	9	9	9

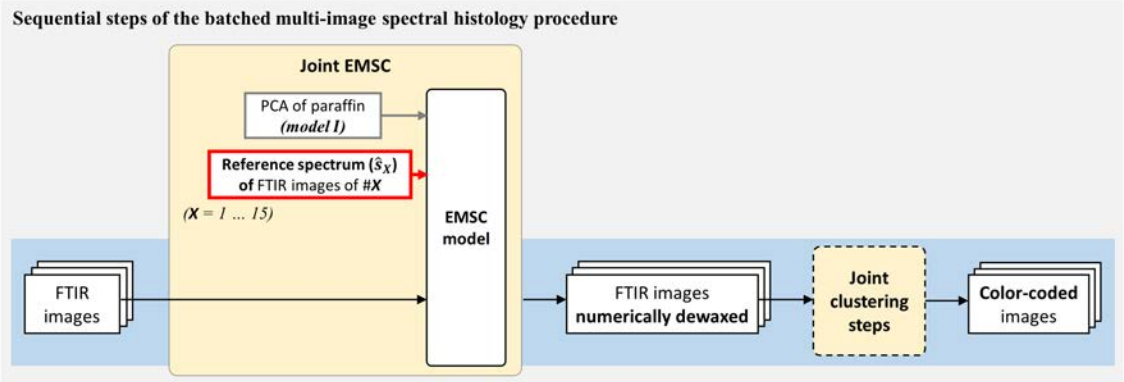


FIGURE IV.4 – Protocole automatisé avec application successive du spectre de référence $\hat{s}_X$ de chaque patient.

En conclusion, l'ensemble de ces résultats montre que k_{opt} estimé par le protocole automatisé est sensible au modèle de paraffine et indépendant du choix du spectre de référence. La variabilité du taux de fidélité τ aurait donc pour origine la variabilité de k_{opt}.

Chapitre V

Développement d'une histologie spectrale optimale par la métaheuristique

Sommaire

V.1 - Préambule 114

V.2 - Article #3 : "*Development of a memetic clustering algorithm for optimal spectral histology : application to FTIR images of normal human colon*" 117

V.3 - Résultats supplémentaires 152

V.3.1 - Confirmation de l'efficacité de CAM 152

V.3.2 - Modification du protocole automatisé par inclusion de CAM . . 152

V.1 - Préambule

Ce travail fait l'objet d'un manuscrit qui vient d'être soumis à *Analyst* :
I. Farah, T.N.Q. Nguyen, A. Groh, D. Guenot, P. Jeannesson, C. Gobinet, *"Development of a memetic clustering algorithm for optimal spectral histology : application to FTIR images of normal human colon"*.

Contexte et Objectif du projet

L'histologie spectrale IR est classiquement réalisée en appliquant une classification non-supervisée KM qui est un algorithme de recherche locale. A chaque exécution, KM converge vers un minimum local différent et dépendant de l'initialisation.

Afin de surmonter cet inconvénient, KM est généralement relancé de nombreuses fois en utilisant des initialisations aléatoires. Le résultat KM retenu est celui ayant les clusters les plus compactes. Cependant, aucune preuve théorique ne démontre que ce résultat est la solution optimale globale.

Des outils numériques, appelés métaheuristiques ont été proposés dans la littérature pour optimiser une fonction objectif de façon globale. Les algorithmes génétiques en sont les plus connus. Pour accélérer leur convergence, ils peuvent être hybridés avec une méthode de recherche locale tel que KM, donnant naissance aux algorithmes mémétiques.

Le but de ce travail a été de développer un algorithme mémétique dédié à la classification non-supervisée et optimale d'images spectrales afin de réaliser une histologie spectrale IR optimale.

Algorithme mémétique développé pour un clustering optimal

Le but de la méthode proposée, le CAM (Clustering par Algorithme Mémétique), est de minimiser de façon globale la fonction objectif de KM. Notre méthode étant basée sur un algorithme génétique, chaque chromosome représente une partition de clustering. Un chromosome est donc composé d'autant de gènes qu'il y a de spectres dans l'image spectrale analysée. Ainsi, chaque gène encode le numéro du cluster auquel chaque spectre a été

affecté. Une population est un ensemble composé de P chromosomes. Ici, la population initiale est générée aléatoirement.

Le cœur de l'algorithme consiste à itérer successivement les quatre étapes suivantes :

1) *Sélection :* Dans la population composée de P chromosomes, $P/2$ sont sélectionnés en utilisant la stratégie de la roulette couplée à une mise à l'échelle exponentielle. Cette règle permet de favoriser la sélection des chromosomes qui ont la valeur de la fonction objectif la plus petite, donc la partition ayant les clusters les plus compacts.
2) *Croisement :* Pour chaque couple de chromosomes parents choisis aléatoirement parmi les $P/2$ sélectionnés, deux chromosomes enfants sont construits en combinant des parties des deux chromosomes parents. Ainsi, la nouvelle population est composée des $P/2$ chromosomes parents et des $P/2$ chromosomes enfants.
3) *Mutation :* Pour chaque chromosome de cette nouvelle population, une perturbation génétique locale affecte 5% de ses gènes choisis aléatoirement. Cette opération permet d'éviter une convergence prématurée de l'algorithme.
4) *Recherche locale :* Dans le but d'explorer l'espace de recherche dans le voisinage d'un chromosome et ainsi d'accélérer la convergence de l'algorithme, N itérations de KM sont appliquées à chaque chromosome.

Ces quatre étapes sont répétées M fois. Dans la $M^{\text{ème}}$ population, le chromosome ayant les clusters les plus compacts correspond à la partition optimale. Un schéma récapitulatif de la méthode est proposé dans la Figure 1 de l'article.

Résultats

Les résultats présentés ont été obtenus sur des images spectrales IR acquises sur des coupes coliques saines provenant de cinq individus.

Comme pour les algorithmes génétiques, la convergence de la méthode CAM proposée ici dépend de la taille P de la population, du nombre d'itérations N de KM et du nombre d'itérations M de CAM. Dans ce travail, ces paramètres ont été choisis en utilisant une recherche par quadrillage répétée dix fois. La convergence, la stabilité et la rapidité de

l'algorithme appliqué sur une image spectrale sont assurées pour $P = 50$, $N = 15$ et $M = 20$.

Ces paramètres fixés, nous avons comparé notre méthode à trois autres :

i) KM puisque c'est la méthode de référence en histologie spectrale,

ii) GABC (Genetic Algorithm-Based Clustering) qui est l'un des algorithmes génétiques les plus cités dans la littérature pour réaliser un clustering optimal,

iii) GKA (Genetic KM Algorithm) qui est l'un des algorithmes mémétiques les plus cités dans la littérature pour réaliser un clustering optimal.

Les résultats obtenus en appliquant chaque méthode dix fois sur chaque image spectrale montrent que CAM estime à chaque fois les classes les plus compactes. En effet, les plus petites valeurs de la fonction objectif sont obtenues par CAM (Tableau 1 de l'article). De plus, CAM est la méthode la plus reproductible puisque, sur les dix réalisations, l'écart-type de sa fonction objectif est le plus petit (proche de zéro), quelque soit l'image spectrale analysée. Enfin, CAM est approximativement cinq fois plus rapide que GKA et GABC (Tableau 3 de l'article).

V.2 - Article #3 : "*Development of a memetic clustering algorithm for optimal spectral histology : application to FTIR images of normal human colon*"

I. Farah, T.N.Q. Nguyen, A. Groh, D. Guenot, P. Jeannesson, C. Gobinet.

Article soumis à *Analyst*, le 27-10-2015.

Analyst

Development of a memetic clustering algorithm for optimal spectral histology: application to FTIR images of normal human colon

Journal:	*Analyst*
Manuscript ID	AN-ART-10-2015-002227
Article Type:	Paper
Date Submitted by the Author:	27-Oct-2015
Complete List of Authors:	Farah, Ihsen; Université de Reims Champagne-Ardenne, Equipe M‼ éDIAN-Biophotonique et Technologies pour la Sant‼ é; CNRS UMR 7369, Matrice Extracellulaire et Dynamique Cellulaire (MEDyC) Nguyen, Thi Nguyet Que; Université de Reims Champagne-Ardenne, Equipe M‼ éDIAN-Biophotonique et Technologies pour la Sant‼ é; CNRS UMR 7369, Matrice Extracellulaire et Dynamique Cellulaire (MEDyC) Groh, Audrey; Université de Strasbourg, EA 3430 Progression tumorale et microenvironnement. Approches translationnelles et Epidémiologie. Fédération de Médecine Translationnelle de Strasbourg Guenot, Dominique; Université de Strasbourg, EA 3430 Progression tumorale et microenvironnement. Approches translationnelles et Epidémiologie. Fédération de Médecine Translationnelle de Strasbourg Jeannesson, Pierre; Universit‼ é de Reims Champagne-Ardenne, Equipe M‼ éDIAN-Biophotonique et Technologies pour la Sant‼ é; CNRS UMR 7369, Matrice Extracellulaire et Dynamique Cellulaire (MEDyC) Gobinet, Cyril; Université de Reims Champagne-Ardenne, Equipe M‼ éDIAN-Biophotonique et Technologies pour la Sant‼ é; CNRS UMR 7369, Matrice Extracellulaire et Dynamique Cellulaire (MEDyC)

Development of a memetic clustering algorithm for optimal spectral histology: application to FTIR images of normal human colon.†

Ihsen Farah[1,2], Thi Nguyet Que Nguyen [1,2], Audrey Groh[3], Dominique Guenot[3], Pierre Jeannesson[1,2],Cyril Gobinet[1,2*]

[1] *Université de Reims Champagne-Ardenne, Equipe MéDIAN-Biophotonique et Technologies pour la Santé, UFR de Pharmacie, 51 rue Cognacq-Jay, 51096 Reims Cedex, France.*

[2] *CNRS UMR 7369, Matrice Extracellulaire et Dynamique Cellulaire (MEDyC), Reims, France.*

[3] *Université de Strasbourg (Unistra), EA 3430 Progression tumorale et microenvironnement. Approches translationnelles et Epidémiologie. Fédération de Médecine Translationnelle de Strasbourg (FMTS), Bâtiment U1113, 3 Avenue Molière, 67200 Strasbourg, France.*

[*] *Corresponding author (email: cyril.gobinet@univ-reims.fr).*

†Electronic supplementary information (ESI) available

Abstract: The coupling between Fourier-transform infrared (FTIR) imaging and unsupervised classification is effective in revealing the different structures of human tissues based on their specific biomolecular IR signatures; thus the spectral histology of the studied samples is achieved.

However, the most widely applied clustering methods in spectral histology are local search algorithms, which converge to a local optimum, depending on initialization. Multiple runs of the techniques estimate multiple different solutions. Here, we propose a memetic algorithm, based on a genetic algorithm and a K-Means clustering refinement, to perform optimal clustering. In addition, this approach was applied to FTIR images acquired on normal human colon tissues originating from five patients. The results show the efficiency of the proposed memetic algorithm to achieve the optimal spectral histology of these samples, contrary to K-Means.

keywords: Memetic clustering, human colon, FTIR spectral imaging, optimal spectral histology

1 Introduction

Fourier transform infrared (FTIR) imaging is a non-destructive, non-invasive and label-free biophotonic technique based on the two-dimensional scan of the light absorbed by a sample. It has been successfully applied to elucidate the histological structure of biological tissues, such as brain [4], skin [21, 22], breast [3], colon [26, 20] and cervix [7].

Acquired images are mathematically defined as data cubes composed of two spatial and one spectral dimensions. Each pixel of a FTIR image represents an IR spectrum informative on the molecular composition of the tissue at this acquisition point.

The analysis of an FTIR image is complex for two reasons. Firstly, a high number (several thousands) of spectra composed of numerous wavenumbers (several hundreds) can be acquired. Secondly, the studied phenomenon can generate weak and subtle spectral responses. The interpretation of such a large and highly multi-dimensional data cube is possible by the development and the application of advanced

numerical data analysis tools. In particular, the application of partitional clustering methods, such as K-Means (KM) [18, 20] and Fuzzy C-Means (FCM) [29, 20], performs an IR spectral histology of the studied tissue, highly correlated to the conventionnal histology. However, these clustering methods are local search techniques, ensuring only the convergence of the algorithm to a local optimum, which is dependent on the initialization. Thus, applied several times to the same data cube, a partitional clustering algorithm can estimate highly variable solutions.

Metaheuristics are numerical methods designed to find the global optimal solution of any optimization problem. Genetic algorithms (GA) [11], ant colony optimization [5], particle swarm optimization [15] are popular examples of population-based metaheuristics. Numerous applications have been reported, such as the traveling salesman problem [6], vehicle routing problem [1], knapsack problem [2], bin-packing [9]. In addition, the hybridization between population-based metaheuritics and local refinement procedures have led to the development of the more efficient memetic algorithms [27].

In this study, we propose a memetic clustering method combining GA and KM to solve the problem of partitional clustering of IR spectral images acquired on normal human colon tissue. We show that our method outperformes KM by estimating the optimal histological partition.

2 Materials and methods

2.1 Sample preparation

Five formalin-fixed paraffin-embedded tissue blocks of normal zones were prepared from surgically excised colons from five patients with a colon cancer. For each tissue block, two consecutive $6\mu m$ thick slices were cut with a microtome. For FTIR image acquisition, the first tissue section was mounted onto a calcium fluoride (CaF_2) window (Crystran, Dorset, UK) which is transparent for IR light. For conventional histological analysis, the second tissue section was mounted on a glass window and stained by Harris' Hematoxylin and Eosin (HE). In this study, the stained tissue section is used as a reference

for the comparison with its corresponding FTIR images.

2.2 FTIR image acquisition

FTIR images were acquired using a Spectrum Spotlight 300 FTIR imaging system coupled to a Spectrum one FTIR spectrometer (Perkin Elmer, Courtabœuf, France), equipped with a nitrogen-cooled Mercury Cadmium Telluride 16-pixel-line detector.

FTIR images were collected with a 6.25 μm spatial resolution, a 4 cm^{-1} spectral resolution, and a 16 scan-averaged accumulation in a mid-IR range of 750 to 4000 cm^{-1}. A 240 scan-averaged reference spectrum was recorded from a blank area of the CaF_2 window in order to substract the background spectrum from the recorded FTIR images, using the Spectrum Image software (Perkin Elmer).

On each tissue section, two FTIR images were collected: (i) one on the tissue area for spectral histology, and (ii) one on a pure paraffin zone for numerical dewaxing.

2.3 Spectral data preprocessing

Before applying clustering methods, preprocessing steps must be applied in order to correct the spectra from parasitic signals.

Firstly, an atmospheric correction was performed on each FTIR image by Spectrum Image software to remove water vapor and CO_2 contributions.

Secondly, the spectral range was limited to the 900-1800 cm^{-1} fingerprint region of biological samples [16].

Thirdly, the spectra were numerically corrected for paraffin signal and baseline, and normalized using the Extended Multiplicative Signal Correction (EMSC) method [32, 17] employing the same parameters as previously described in [28].

After the preprocessing stage, two different clustering algorithms were applied independently on each

IR image in order to highlight the tissue structures of the studied samples: (i) the classical KM clustering which is a local search method, and (ii) a memetic clustering algorithm developed to globally optimize the clustering problem.

2.4 Partitional clustering

2.4.1 K-Means clustering

KM [23] is the most popular unsupervised classification method. Its aim is to partition into k clusters a set $X = \{x_i \mid 1 \leq i \leq n\}$ of n patterns where $x_i = \{x_{il} \mid 1 \leq l \leq d\}$ is the i^{th} pattern composed of d features. The clusters are estimated by minimizing the total within-cluster variation defined as:

$$f(X, W, C) = \sum_{j=1}^{k} \sum_{i=1}^{n} w_{ij} \|x_i - c_j\|^2. \tag{1}$$

$C = \{c_j \mid 1 \leq j \leq k\}$ is the set of barycenters (also called centroids) where $c_j = \{c_{jl} \mid 1 \leq l \leq d\}$ is the barycenter of the j^{th} cluster. $\|x_i - c_j\| = \sqrt{\sum_{l=1}^{d} (x_{il} - c_{jl})^2}$ is the Euclidean distance between the i^{th} pattern and the j^{th} barycenter. $W = \{w_{ij} \mid 1 \leq i \leq n, 1 \leq j \leq k\}$ is the membership matrix where:

$$w_{ij} = \begin{cases} 1 & \text{if the } i^{\text{th}} \text{ pattern belongs to the } j^{\text{th}} \text{ cluster} \\ 0 & \text{otherwise} \end{cases}. \tag{2}$$

W must respect the following constraints:

$$\sum_{i=1}^{n} w_{ij} \geq 1, \qquad 1 \leq j \leq k, \tag{3}$$

$$\sum_{j=1}^{k} w_{ij} = 1, \qquad 1 \leq i \leq n, \tag{4}$$

$$\sum_{j=1}^{k} \sum_{i=1}^{n} w_{ij} = n. \tag{5}$$

Constraint (3) means that each cluster must have at least one pattern. Constraints (4) and (5) imply that each pattern must be assigned to a unique cluster. A KM partition is defined as: $\psi = \{\psi_i | 1 \leq i \leq n\}$

with $\psi_i = j$ such that $w_{ij} = 1$ and $j \in \{1, 2, ..., k\}$.

This KM optimization problem can be addressed using the following algorithm:

i) Choose randomly k patterns as barycenters.

ii) Assign each pattern to the nearest barycenter in terms of Euclidean distance.

iii) Update barycenters using $c_j = (\sum_{i=1}^{n} w_{ij} x_i)/(\sum_{i=1}^{n} w_{ij}),\ 1 \leq j \leq k$.

iv) Repeat steps (ii) and (iii) until no reassignement of patterns occurs.

2.4.2 Memetic clustering

A memetic algorithm (MA)[27] is a global optimization method based on the hybridization between a population-based approach and a local refinement technique. In this paper, a MA coupling a GA and a KM local search, named memetic clustering (MC), is proposed for the clustering of infrared spectral images. This MC method globally minimizes the KM objective function $f(X, W, C)$ defined in equation (1).

Problem encoding A clustering partition is encoded as a chromosome ψ_p, such that $W^{(p)}$ and $C^{(p)}$ correspond to its membership matrix and centroid matrix, respectively. A chromosome is composed of n genes where each gene $\psi_{pi},\ 1 \leq i \leq n$, takes a value in $\{1, 2, ..., k\}$ according to the string-of-group-numbers encoding [14]. Then, the i^{th} gene represents the cluster number to which the i^{th} pattern belongs.

Initial population The initial population is composed by P chromosomes $\Psi = \{\psi_p | 1 \leq p \leq P\}$. For each chromosome ψ_p, k patterns are randomly selected as barycenters C_p. The i^{th} pattern x_i (i.e. the i^{th} gene) is assigned to the cluster with the nearest barycenter $c_j^{(p)}$, i.e. the barycenter which minimizes the squared Euclidean distance $\|x_i - c_j^{(p)}\|^2,\ 1 \leq i \leq n,\ 1 \leq j \leq k$. Thus, each chromosome corresponds to an initial partition of the n patterns into k clusters.

Selection operator GA is based on the generation of children chromosomes from the crossover of parent chromosomes selected from the current population. In this study, a dynamic parent selection was adopted using the roulette wheel strategy [11] coupled to the exponential scaling [25].

With exponential scaling, the fitness of chromosome ψ_p is measured by $g(\psi_p) = 1/f(X, W^{(p)}, C^{(p)})^{E(t)}$, where:

$$E(t) = \tan\Big[\Big(\frac{t}{T+1}\Big)\frac{\pi}{2}\Big]\rho. \tag{6}$$

t is the current iteration, T is the total number of iterations. $\rho \in]0;1[$ is a constant controlling the weight of the chromosomes during the selection process. Then, in the roulette wheel strategy, $P/2$ chromosomes are randomly selected with a probability proportional to their fitness values. During the first iterations of MC, the chromosomes have the same probability to be selected (because $E(t) \approx 0$, thus $g(\psi_p) \approx 1$, $\forall p$). At the end of the algorithm, chromosomes with high fitness have a high probability to be chosen (because $E(t) \gg 0$). The higher ρ, the earlier the chromosomes with the highest fitness will be fostered.

In addition, the elitism scheme is used in this work, *i.e.* the best chromosome is automatically selected.

Crossover operator The goal of the crossover operator is to exchange information between two selected parent chromosomes ψ_{p_1} and ψ_{p_2}, to generate two children chromosomes ψ_{c_1} and ψ_{c_2}. Let b_1 and b_2 be two integers randomly selected in $\{1, 2, ..., n-1\}$. The first child is composed by genes 1 to b_1 from the first parent, genes $b_1 + 1$ to b_2 from the second parent and genes $b_2 + 1$ to n from the first parent. Thus, $\psi_{c_1} = \{\psi_{p_1 1}, \ldots, \psi_{p_1 b_1}, \psi_{p_2(b_1+1)}, \ldots, \psi_{p_2 b_2}, \psi_{p_1(b_2+1)}, \ldots, \psi_{p_1 n}\}$. The second child is composed by genes 1 to b_1 from the second parent, genes $b_1 + 1$ to b_2 from the first parent and genes $b_2 + 1$ to n from the second parent. Thus, $\psi_{c_2} = \{\psi_{p_2 1}, \ldots, \psi_{p_2 b_1}, \psi_{p_1(b_1+1)}, \ldots, \psi_{p_1 b_2}, \psi_{p_2(b_2+1)}, \ldots, \psi_{p_2 n}\}$. This crossover operator can generate chromosomes containing empty clusters. In this case, a correction operator is applied on these chromosomes in order to randomly create new non-empty clusters.

The new population is composed by the $P/2$ selected parents and the $P/2$ generated children.

Mutation operator The mutation operator acts as a local genetic perturbation on each chromosome to prevent premature algorithm convergence. In our algorithm, we used the mutation operator defined in reference [19]. Each gene has a probability $p_m = 0.05$ to mutate, *i.e.* to change its value. For the p^{th} chromosome, the value of the i^{th} gene subjected to a mutation is selected in the range $\{1, 2, \ldots, k\}$ by the roulette wheel procedure described above, using the following fitness function:

$$h_j = d_{max} - \|x_i - c_j^{(p)}\|^2 \tag{7}$$

where $d_{max} = \max_{1 \leq j \leq k} \{\|x_i - c_j^{(p)}\|^2\}$, $c_j^{(p)}$ is the j^{th} updated barycenter of the p^{th} chromosome. The gene value has a high probability to be equal to the cluster number with the nearest barycenter.

Local search operator The local search operator consists to refine chromosomes computed by the application of the selection, crossover and mutation operators. Its role is to explore the search space in the chromosome neighbourhood in order to accelerate the convergence toward a global optimum. In this work, a limited number N of KM steps is applied on each chromosome as the local search operator.

Finally, MC repeats the selection, crossover, mutation and local search operators until the number of iterations M is reached. The output of the algorithm is the best chromosome of the last population. The different steps of our MC are summarised in the flowchart presented in Figure 1.

2.5 Clustering pseudo-color-coded images and cluster assignment

After the preprocessing stage, KM and MC clustering were applied separately on each IR image. For each estimated partition, a unique color is attributed to pixels belonging to the same cluster. Then, the corresponding reconstructed color-coded image is visually analysed by an expert pathologist who annotates each cluster to its corresponding histological class by comparaison to a reference HE-stained image.

2.6 Quality measure of a partition

The quality of a partition estimated by KM or MC is measured by the KM objective function f defined in equation (1). For a given number of clusters k, the smaller the objective function, the better the partition.

3 Results and discussion

3.1 Conventional histology of normal human colon

Here, convential histology using HE staining is considered as the reference method for the morphological recognition of the tissue structures. Figure 2 shows the HE-stained colon tissue section of patient #1. This image illustrates the five main histological tissue structures of normal human colon. The outer layer (mucosa) is composed of tubular glands (crypts of Lieberkühn (structure 1)) and of connective tissue (lamina propria (structure 2)). A thin layer of muscle (muscularis mucosae (structure 3)) links the mucosa to the submucosa (structure 4) which is rich in adipose tissue. Lymphoid aggregates (structure 5) can locally appear in the lamina propria and the submucosa or extend from the lamina propria to the submucosa [30].

The HE-stained colon tissue sections of the four remaining patients are shown in Figures S1-S4 of the ESI document.

In the following, the KM and MC pseudo-color-coded images are compared to these HE sections for the assignment of the estimated clusters to the main tissue structures.

3.2 Limitations of KM clustering

In our study, KM is repeated 100 times with $k = 15$ clusters on each tissue section in order to evaluate the variability of KM results due to its random initialization. As mentioned in section 2.6, each partition is evaluated by its estimated objective function value. The second column of Table 1 presents the mean

and the standard deviation of these 100 quality measures for the five patients. These results show a high variability of KM objective function value. Consequently, a high variability of estimated clusters is visible on KM partitions, as can be seen in Figure 3 for the tissue section of patient #1. For example, the lamina propria is represented by 2, 3, and 1 clusters on the best (Figure 3(a)), the most frequent (Figure 3(b)), and the worst (Figure 3(c)) partitions, respectively. In addition, only 1% of KM results corresponds to the best partition, justifying the routine application of KM replicates for spectral histology [21, 22]. Since KM is a local search method, it converges to a local minimum [12]. Thus, there is no certainty that the best KM partition is the optimal one.

The best, the most frequent, and the worst partitions estimated by KM on the four remaining patients are available in Figures S1-S4 of the ESI document.

To overcome these KM limitations, MC is proposed to partition data in an optimal way.

3.3 Setting of MC algorithm parameters

A critical phase of MC is the right choice of its parameters, presented in section 2.4.2, in order to insure the convergence of the algorithm to the optimal solution. In this paper, a grid search using $k = 15$ clusters was used to choose the parameters varying as follows: $P \in \{10, 20, 30, 40, 50, 60, 70\}$, $M \in \{20, 40, 60, 80, 100\}$, $\rho \in \{0.1, 0.3, 0.5, 0.7, 0.9\}$ and $N \in \{1, 5, 10, 15, 20\}$. For each setting of the parameters, the variability of MC is evaluated over 10 replicates, using the IR spectral image acquired on the patient #1 tissue section.

For this spectral dataset, we found that the smallest population size P required to ensure the convergence of MC is for $\rho = 0.1$ (data not shown). Hence, ρ was fixed to 0.1 in the rest of this article.

The remaining parameters were selected by analyzing the mean and standard deviation of the quality measure of the 10 estimated partitions. For each value of N, these results can be represented by a 3D-map. Two kinds of map are observed as shown in figure 4. Whatever the values of P and M, the first kind is characterized by a variable mean and a high standard deviation typical of the non-convergence of MC, as shown in Figure 4(a) for $N = 1$. In contrast, the second kind of map is composed of a flat

region defined by a constant mean and a tiny standard deviation, as shown in Figure 4(b) for $N = 15$. This region is characteristic of the convergence of MC to the global optimal solution. However, on this convergence region, the smaller P and M, the shorter the computational time. Thus, in the following, this region is summarized by its minimum values of P and M. These results are presented in Table 2. Contrary to the case of $N \in \{1, 5\}$, MC converges to the optimal partition for $N \in \{10, 15, 20\}$ in the tested parameter ranges. Thus, the setting of paramaters was driven by the computational time shown in the last column of Table 2. In the following, the parameters of MC were fixed to $\rho = 0.1$, $N = 15$, $P = 50$, and $M = 20$ since this setting minimizes the computational time. The efficiency of this setting has been confirmed by the optimal convergence of MC on the spectral images of the remaining patients.

3.4 Efficiency of MC algorithm

In our study, MC is repeated 10 times with $k = 15$ clusters on each tissue section. MC variability is evaluated by the mean and the standard deviation of the 10 quality measures. These results are summarized in the third column of Table 1. Compared to KM clustering, MC estimates a better solution, since its quality measure is smaller for all the patients. Furthermore, MC is reproducible since its standard deviation is close to zero. These results are consistent with expectations since MC is a global optimization method contrary to KM which is a local search algorithm.

An example of the optimal partition estimated by MC for patient #1 is given in Figure 5. The main difference with the best KM partition (Figure 3(a)) is visible at the level of lamina propria represented by two clusters by KM and one by MC. Other small differences can be seen for the other main histological structures.

The optimal partitions estimated by MC on the four remaining patients are available in Figures S1-S4 of the ESI document. In some cases, KM clustering can converge to the optimal partition as shown in Figure S4 of the ESI document for patient #5. This event is rare as it depends on the data structure, the chosen number of clusters, and the KM initialization. On the contrary, MC always converges to the

optimal solution.

Other clustering algorithms based on metaheuristics have been proposed such as Genetic K-Means Algorithm (GKA) [19] and Genetic Algorithm-Based Clustering (GABC) [24]. These two algorithms were applied 10 times on each FTIR image. The mean and standard deviation of their corresponding quality measures are given in the columns 4-5 of Table 1. These results show that GKA and GABC are less efficient and less reproducible than MC algorithm, since their mean and standard deviation of quality measures are higher than those of MC.

An important characteristic of metaheuristics is the computational time. Table 3 presents the mean and standard deviation of the computational time of MC, GAK, and GABC over 10 replicates. These data show that MC is four times faster than GAK and GABC.

Several studies have shown the efficacity of metaheuristics applied to IR spectral data. For example, genetic algorithms have been developed for the supervised classification of FTIR spectra acquired on different species of bacteria [10], for the optimal selection of discriminant subsets of wavelengths [13, 8, 22, 31], and for the selection of the best sequence of preprocessing steps applied to spectral data [13]. To the best of our knowledge, this is the first study presenting a metaheuristics-based algorithm specifically developed for the clustering of IR images.

The proposed MC has been proven effective to perform the spectral histology of human normal colon tissues. Obviously, it can be applied to infrared images acquired on other human tissue samples. Moreover, MC is based on a general framework suitable to all kinds of data.

4 Conclusion

In this study, an optimal memetic clustering (MC) combining a genetic algorithm and a refinement by KM was developed. Applied on FTIR images acquired on normal human colon tissue samples originating from five patients, this method outperformed standard KM and two popular genetic algorithm-based clustering

techniques named GKA and GABC. Compared to these three methods, our algorithm reproducibly converges to the optimal solution. In addition, MC is four times faster than GKA and GABC. Owing to its general framework, our algorithm may be applied for the spectral histology of any kind of tissue and for the clustering of any kind of data.

Acknowledgments

The authors thank Cancéropôle Grand-Est, Ligue contre le Cancer, the URCA technological platform of cellular and tissular imaging PICT-IBiSA, Région Champagne-Ardenne, Région Alsace, and Ministère de l'Enseignement Supérieur et de la Recherche for financial support, and Shawn Hussain for linguistic assistance.

References

[1] B. M. Baker and M. A. Ayechew. A genetic algorithm for the vehicle routing problem. *Computers & Operations Research*, 30(5):787–800, 2003.

[2] J. C. Bansal and K. Deep. A modified binary particle swarm optimization for knapsack problems. *Applied Mathematics and Computation*, 218(22):11042–11061, 2012.

[3] A. Benard, C. Desmedt, M. Smolina, P. Szternfeld, M. Verdonck, G. Rouas, N. Kheddoumi, F. Rothé, D. Larsimont, C. Sotiriou, and E. Goormaghtigh. Infrared imaging in breast cancer: automated tissue component recognition and spectral characterization of breast cancer cells as well as the tumor microenvironment. *Analyst*, 139(5):1044–1056, 2014.

[4] N. Bergner, B. F. Romeike, R. Reichart, R. Kalff, C. Krafft, and J. Popp. Tumor margin identification and prediction of the primary tumor from brain metastases using FTIR imaging and support vector machines. *Analyst*, 138(14):3983–3990, 2013.

[5] M. Dorigo and C. Blum. Ant colony optimization theory: A survey. *Theoretical Computer Science*, 344(2-3):243–278, 2005.

[6] M. Dorigo and L. M. Gambardella. Ant colonies for the travelling salesman problem. *BioSystems*, 43(2):73–81, 1997.

[7] J. Einenkel, U.-D. Braumann, W. Steller, H. Binder, and L.-C. Horn. Suitability of infrared microspectroscopic imaging for histopathology of the uterine cervix. *Histopathology*, 60(7):1084–1098, 2012.

[8] G. N. Elliott, H. Worgan, D. Broadhurst, J. Draper, and J. Scullion. Soil differentiation using fingerprint Fourier transform infrared spectroscopy, chemometrics and genetic algorithm-based feature selection. *Soil Biology and Biochemistry*, 39(11):2888–2896, 2007.

[9] E. Falkenauer. A hybrid grouping genetic algorithm for bin packing. *Journal of heuristics*, 2(1): 5–30, 1996.

[10] R. Goodacre, B. Shann, R. J. Gilbert, E. M. Timmins, A. C. McGovern, B. K. Alsberg, D. B. Kell, and N. A. Logan. Detection of the dipicolinic acid biomarker in Bacillus spores using Curie-point pyrolysis mass spectrometry and Fourier transform infrared spectroscopy. *Analytical Chemistry*, 72 (1):119–127, 2000.

[11] J. H. Holland. Adaptation in natural and artificial systems. *Ann Arbor, University of Michigan Press*, 1975.

[12] A. K. Jain, M. N. Murty, and P. J. Flynn. Data clustering: a review. *ACM Computing Surveys*, 31 (3):264–323, 1999.

[13] R. M. Jarvis and R. Goodacre. Genetic algorithm optimization for pre-processing and variable selection of spectroscopic data. *Bioinformatics*, 21(7):860–868, 2005.

[14] D. R. Jones and M. A. Beltramo. Solving partitioning problems with genetic algorithms. In *Proceedings of the 4th International Conference on Genetic Algorithms*, pages 442–449, 1991.

[15] J. Kennedy. Particle swarm optimization. In *Encyclopedia of Machine Learning*, pages 760–766. Springer, 2010.

[16] M. Khanmohammadi, A. B. Garmarudi, K. Ghasemi, H. K. Jaliseh, and A. Kaviani. Diagnosis of colon cancer by attenuated total reflectance-Fourier transform infrared microspectroscopy and soft independent modeling of class analogy. *Medical Oncology*, 26(3):292–297, 2009.

[17] A. Kohler, N. Kristian Afseth, and H. Martens. Chemometrics in biospectroscopy. In *Handbook of Vibrational Spectroscopy*. John Wiley & Sons, Ltd, 2006.

[18] C. Krafft, D. Codrich, G. Pelizzo, and V. Sergo. Raman mapping and FTIR imaging of lung tissue: congenital cystic adenomatoid malformation. *Analyst*, 133(3):361–371, 2008.

[19] K. Krishna and M. N. Murty. Genetic K-means algorithm. *IEEE Transactions on Systems, Man, and Cybernetics-Part B : Cybernetics*, 29(3):433–439, 1999.

[20] P. Lasch, W. Haensch, D. Naumann, and M. Diem. Imaging of colorectal adenocarcinoma using FT-IR microspectroscopy and cluster analysis. *Biochimica et Biophysica Acta (BBA)-Molecular Basis of Disease*, 1688(2):176–186, 2004.

[21] E. Ly, O. Piot, R. Wolthuis, A. Durlach, P. Bernard, and M. Manfait. Combination of FTIR spectral imaging and chemometrics for tumour detection from paraffin-embedded biopsies. *Analyst*, 133(2): 197–205, 2008.

[22] E. Ly, O. Piot, A. Durlach, P. Bernard, and M. Manfait. Differential diagnosis of cutaneous carcinomas by infrared spectral micro-imaging combined with pattern recognition. *Analyst*, 134(6): 1208–1214, 2009.

[23] J. MacQueen. Some methods for classification and analysis of multivariate observations. *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, 1(281-297):14, 1967.

[24] U. Maulik and S. Bandyopadhyay. Genetic algorithm-based clustering technique. *Pattern Recognition*, 33(9):1455–1465, 2000.

[25] Z. Michalewicz. *Genetic algorithms + data structures = evolution programs*. Springer-Verlag, 1991.

[26] J. Nallala, C. Gobinet, M.-D. Diebold, V. Untereiner, O. Bouché, M. Manfait, G. D. Sockalingum, and O. Piot. Infrared spectral imaging as a novel approach for histopathological recognition in colon cancer diagnosis. *Journal of Biomedical Optics*, 17(11):116013–116013, 2012.

[27] F. Neri, C. Cotta, and P. Moscato. *Handbook of memetic algorithms*, volume 379. Springer, 2012.

[28] T. N. Q. Nguyen, P. Jeannesson, A. Groh, D. Guenot, and C. Gobinet. Development of a hierarchical double application of crisp cluster validity indices: a proof-of-concept study for automated FTIR spectral histology. *Analyst*, 140(7):2439–2448, 2015.

[29] D. Sebiskveradze, V. Vrabie, C. Gobinet, A. Durlach, P. Bernard, E. Ly, M. Manfait, P. Jeannesson, and O. Piot. Automation of an algorithm based on fuzzy clustering for analyzing tumoral heterogeneity in human skin carcinoma tissue sections. *Laboratory Investigation*, 91(5):799–811, 2011.

[30] P. M. Treuting and S. M. Dintzis. *Comparative anatomy and histology: a mouse and human atlas*. Academic Press, 2011.

[31] J. Trevisan, P. P. Angelov, P. L. Carmichael, A. D. Scott, and F. L. Martin. Extracting biological information with computational analysis of Fourier-transform infrared (FTIR) biospectroscopy datasets: current practices to future perspectives. *Analyst*, 137(14):3202–3215, 2012.

[32] R. Wolthuis, A. Travo, C. Nicolet, A. Neuville, M. P. Gaub, D. Guenot, E. Ly, M. Manfait, P. Jean-

nesson, and O. Piot. IR spectral imaging for histopathological characterization of xenografted human colon carcinomas. *Analytical Chemistry*, 80(22):8461–8469, 2008.

List of Tables

1 Mean $\overline{f}$ and standard deviation σ of the partition quality measures computed over 100 replicates for KM, and 10 for MC, GAK and GABC. For each patient, bold values represent the smallest $\overline{f}$ among the four tested clustering methods. 19

2 MC convergence regions characterised by the number of KM iterations N, the minimum population size P, the minimum number of iterations M, the mean $\overline{f}$ and the standard deviation σ of the quality measures, and the mean computational time $\overline{t}$ in seconds, over 10 replicates. For a given N, "/" means that MC does not converge whatever the values of P and M. 20

3 Mean $\overline{t}$ and standard deviation σ_t of the computational time (in seconds) over 10 replicates for MC, GAK and GABC. For each patient, bold values represent the smallest $\overline{t}$ among the three tested clustering methods. 21

	$\overline{f} \pm \sigma$			
Patient	*KM*	*MC*	*GAK*	*GABC*
#1	21.0075 ± 0.1970	$\mathbf{20.6118} \pm 0.0008$	20.6317 ± 0.0164	22.4859 ± 0.1769
#2	4.1730 ± 0.0628	$\mathbf{3.9734} \pm 0.0004$	3.9918 ± 0.0208	4.4089 ± 0.0627
#3	11.4789 ± 0.2418	$\mathbf{10.3126} \pm 0.0027$	10.3202 ± 0.0001	11.2471 ± 0.1049
#4	15.5702 ± 0.1684	$\mathbf{15.2106} \pm 0.0050$	15.2366 ± 0.0230	16.8276 ± 0.1886
#5	5.4735 ± 0.0838	$\mathbf{5.3722} \pm 0.0004$	5.3723 ± 0.0004	6.0431 ± 0.0679

Table 1: Mean $\overline{f}$ and standard deviation σ of the partition quality measures computed over 100 replicates for KM, and 10 for MC, GAK and GABC. For each patient, bold values represent the smallest $\overline{f}$ among the four tested clustering methods.

N	P	M	$\overline{f}$	σ	$\bar{t}$
1	/	/	/	/	/
5	/	/	/	/	/
10	70	80	20.6115	0.0005	3906.3951
15	50	20	20.6118	0.0008	793.6144
20	40	40	20.6114	0.0004	1285.7603

Table 2: MC convergence regions characterised by the number of KM iterations N, the minimum population size P, the minimum number of iterations M, the mean $\overline{f}$ and the standard deviation σ of the quality measures, and the mean computational time $\bar{t}$ in seconds, over 10 replicates. For a given N, "/" means that MC does not converge whatever the values of P and M.

	$\bar{t} \pm \sigma_t$		
Patient	*MC*	*GAK*	*GABC*
#1	**793.6144** $\pm$ 19.7452	3713.1001 $\pm$ 35.3542	3349.3905 $\pm$ 6.1103
#2	**553.4608** $\pm$ 37.0554	2136.6738 $\pm$ 21.0275	2722.8798 $\pm$ 26.6593
#3	**632.9907** $\pm$ 46.0887	2561.5067 $\pm$ 29.0262	2090.8268 $\pm$ 27.4400
#4	**728.1194** $\pm$ 6.8380	3599.2612 $\pm$ 69.6974	2871.8835 $\pm$ 21.8029
#5	**412.0091** $\pm$ 42.4198	1775.1262 $\pm$ 29.5005	2221.7898 $\pm$ 18.7104

Table 3: Mean $\bar{t}$ and standard deviation σ_t of the computational time (in seconds) over 10 replicates for MC, GAK and GABC. For each patient, bold values represent the smallest $\bar{t}$ among the three tested clustering methods.

List of Figures

1 Flowchart of the developed memetic clustering (MC) algorithm. 23

2 HE-stained image of normal human colon tissue of patient #1. Its main histological tissue structures are annotated by numbers: (1) the crypts, (2) the lamina propria, (3) the muscularis mucosae, (4) the submucosa and (5) the lymphoid aggregate. Scale bar indicates 100 μm. 24

3 Examples among the 100 KM partitions estimated for patient #1: (a) the best (f = 20.638), (b) the most frequent (f = 20.931), and (c) the worst (f = 21.771) partitions. Scale bars indicate 100 μm. The cluster assignments are detailed below each pseudo-color-coded image. 25

4 Examples of grid search maps for (a) $N = 1$ and (b) $N = 15$, using $\rho = 0.1$. Each map represents the quality measure mean $\overline{f}$, over 10 MC replicates, in function of the population size P and the number of iterations M. Error bars represent the standard deviation σ. . . 26

5 Optimal partition estimated by MC for patient #1 (f = 20.611). Scale bar indicates 100 μm. The cluster assignments are detailed beside the pseudo-color-coded image. 27

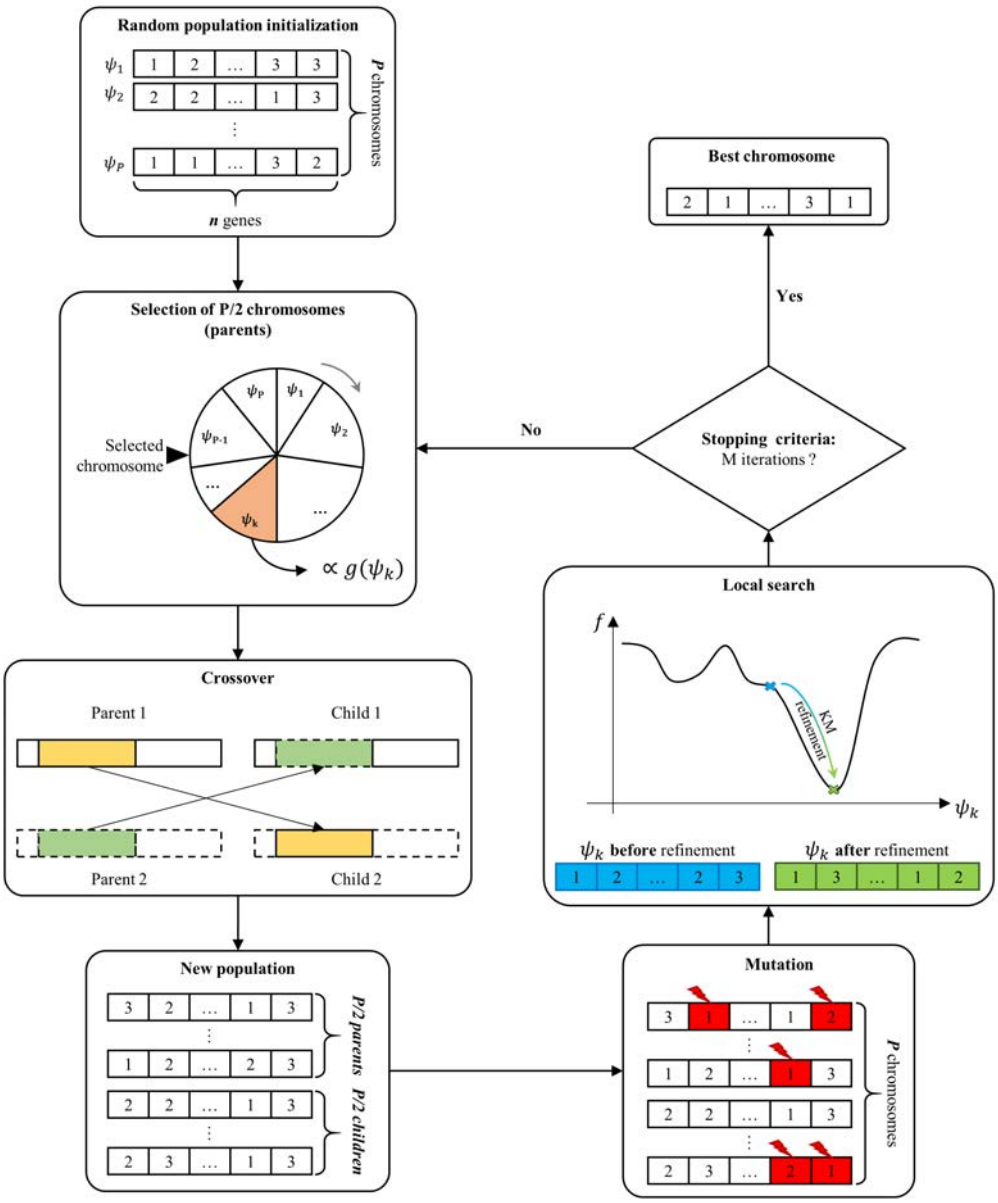


Figure 1: Flowchart of the developed memetic clustering (MC) algorithm.

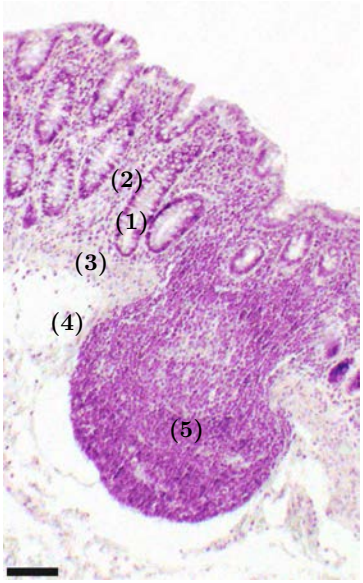


Figure 2: HE-stained image of normal human colon tissue of patient #1. Its main histological tissue structures are annotated by numbers: (1) the crypts, (2) the lamina propria, (3) the muscularis mucosae, (4) the submucosa and (5) the lymphoid aggregate. Scale bar indicates 100 μm.

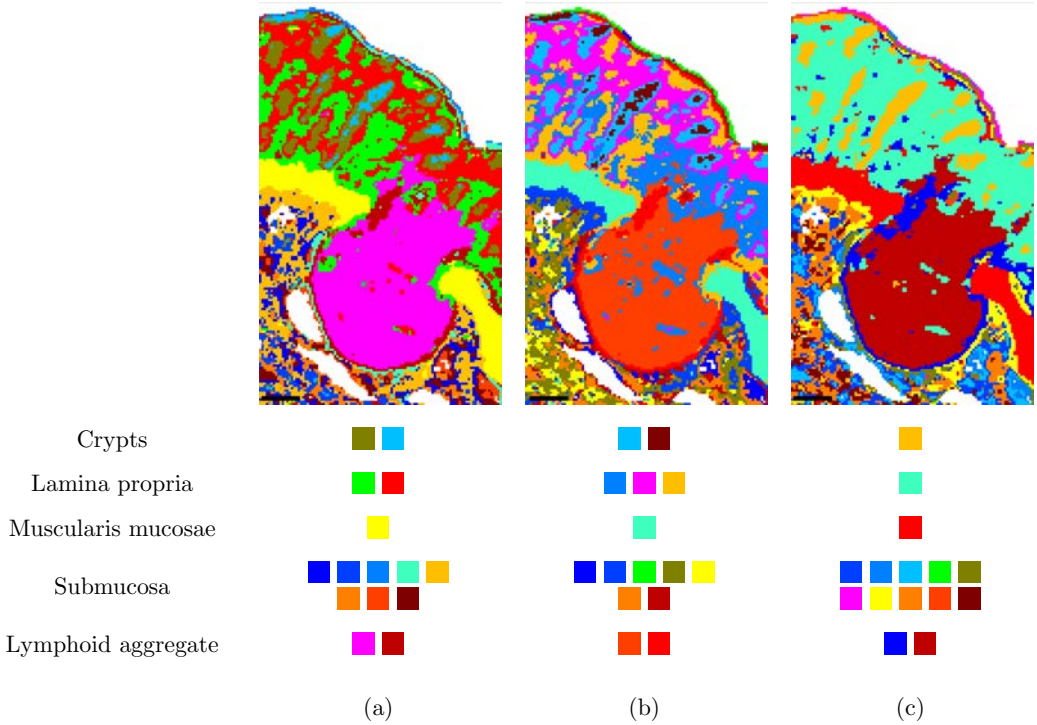
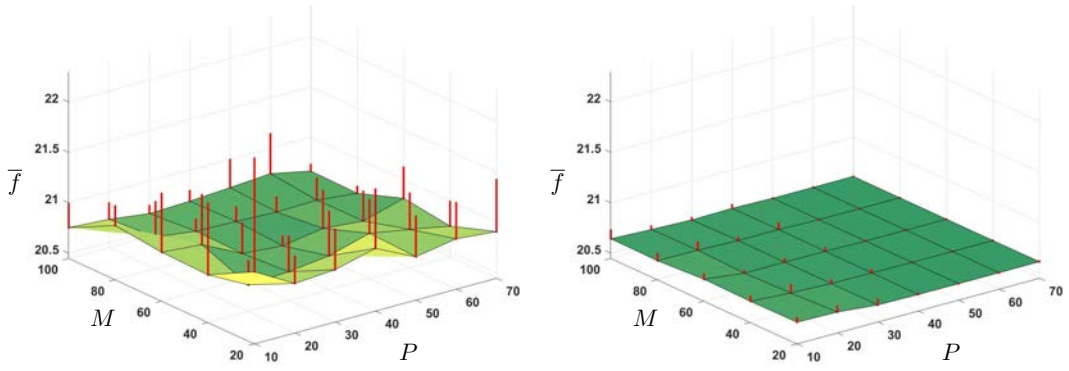


Figure 3: Examples among the 100 KM partitions estimated for patient #1: (a) the best ($f = 20.638$), (b) the most frequent ($f = 20.931$), and (c) the worst ($f = 21.771$) partitions. Scale bars indicate 100 μm. The cluster assignments are detailed below each pseudo-color-coded image.



(a) (b)

Figure 4: Examples of grid search maps for (a) $N = 1$ and (b) $N = 15$, using $\rho = 0.1$. Each map represents the quality measure mean $\overline{f}$, over 10 MC replicates, in function of the population size P and the number of iterations M. Error bars represent the standard deviation σ.

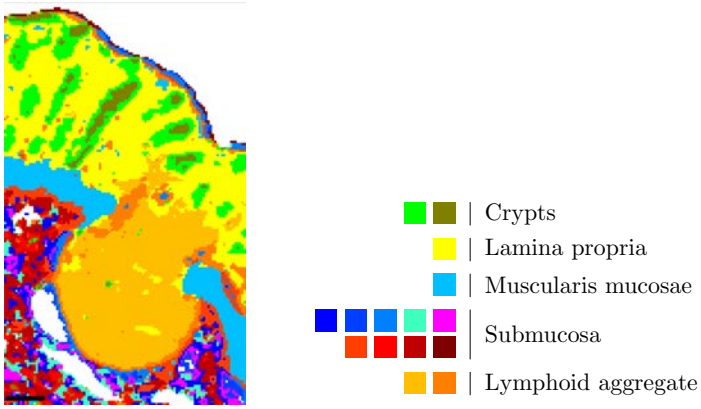


Figure 5: Optimal partition estimated by MC for patient #1 ($f = 20.611$). Scale bar indicates 100 μm.

The cluster assignments are detailed beside the pseudo-color-coded image.

Electronic supplementary information (ESI)

List of Figures

S1 Comparison between conventional histology and pseudo-color-coded images reconstructed from KM and MC partitions for patient #2. Scale bars indicate 100 μm. The quality measure f of each partition is provided above each pseudo-color-coded image. The cluster assignments are detailed below each pseudo-color-coded image. 2

S2 Comparison between conventional histology and pseudo-color-coded images reconstructed from KM and MC partitions for patient #3. Scale bars indicate 100 μm. The quality measure f of each partition is provided above each pseudo-color-coded image. The cluster assignments are detailed below each pseudo-color-coded image. 3

S3 Comparison between conventional histology and pseudo-color-coded images reconstructed from KM and MC partitions for patient #4. Scale bars indicate 100 μm. The quality measure f of each partition is provided above each pseudo-color-coded image. The cluster assignments are detailed below each pseudo-color-coded image. 4

S4 Comparison between conventional histology and pseudo-color-coded images reconstructed from KM and MC partitions for patient #5. Scale bars indicate 100 μm. The quality measure f of each partition is provided above each pseudo-color-coded image. The cluster assignments are detailed below each pseudo-color-coded image. 5

Figure S1: Comparison between conventional histology and pseudo-color-coded images reconstructed from KM and MC partitions for patient #2. Scale bars indicate 100 μm.

The quality measure f of each partition is provided above each pseudo-color-coded image. The cluster assignments are detailed below each pseudo-color-coded image.

Figure S2: Comparison between conventional histology and pseudo-color-coded images reconstructed from KM and MC partitions for patient #3. Scale bars indicate 100 μm. The quality measure f of each partition is provided above each pseudo-color-coded image. The cluster assignments are detailed below each pseudo-color-coded image.

Figure S3: Comparison between conventional histology and pseudo-color-coded images reconstructed from KM and MC partitions for patient #4. Scale bars indicate 100 μm. The quality measure f of each partition is provided above each pseudo-color-coded image. The cluster assignments are detailed below each pseudo-color-coded image.

Figure S4: Comparison between conventional histology and pseudo-color-coded images reconstructed from KM and MC partitions for patient #5. Scale bars indicate 100 μm. The quality measure f of each partition is provided above each pseudo-color-coded image. The cluster assignments are detailed below each pseudo-color-coded image.

V.3 - Résultats supplémentaires

V.3.1 - Confirmation de l'efficacité de CAM

Afin de vérifier les résultats présentés dans l'article, des images spectrales ont été acquises sur les échantillons coliques provenant de 10 patients supplémentaires. Sur ces données, 100 répétitions de KM et 10 répétitions de CAM, GKA et GABC ont été réalisées.

Les résultats du Tableau V.1 montrent que la valeur moyenne $\overline{f}$ de la fonction objectif estimée par CAM est plus petite que celle estimée par les autres méthodes pour les 10 patients. Sur ces nouveaux échantillons, notre algorithme est donc le plus efficace en terme d'optimisation de la fonction objectif de KM. Dans 70% des cas, l'écart-types σ de la fonction objectif estimée par CAM est inférieur à celui estimé par KM, GKA et GABC. La reproductibilité de notre méthode est donc confirmée.

TABLE V.1 – Moyenne $\overline{f}$ et écart-type σ de la fonction objectif de KM calculés sur 100 répétitions de KM et 10 répétitions de CAM, GAK et GABC. Pour chaque patient, la valeur en gras représente la plus petite valeur de $\overline{f}$ parmi les quatre méthodes de clustering.

	$\overline{f} \pm \sigma$			
Patient	KM	CAM	GABC	GAK
#6	$10{,}2551 \pm 0{,}4134$	**9,8608** $\pm 0{,}0198$	$11{,}2873 \pm 0{,}4531$	$9{,}886 \pm 0{,}0123$
#7	$4{,}3107 \pm 0{,}0521$	**4,226** $\pm 0{,}0001$	$4{,}8626 \pm 0{,}1131$	$4{,}2282 \pm 0{,}0027$
#8	$7{,}4039 \pm 0{,}2042$	**7,1419** $\pm 0{,}00003$	$8{,}1197 \pm 0{,}1988$	$7{,}1425 \pm 0{,}0006$
#9	$12{,}9793 \pm 0{,}2487$	**12,7084** $\pm 0{,}0001$	$14{,}5747 \pm 0{,}6804$	$12{,}7099 \pm 0{,}0013$
#10	$8{,}4715 \pm 0{,}0795$	**8,3419** $\pm 0{,}0047$	$9{,}3445 \pm 0{,}1793$	$8{,}3488 \pm 0{,}0189$
#11	$18{,}9947 \pm 0{,}489$	**18,2972** $\pm 0{,}0097$	$21{,}5243 \pm 0{,}5589$	$18{,}3225 \pm 0{,}0215$
#12	$6{,}2882 \pm 0{,}1504$	**6,0691** $\pm 0{,}005$	$7{,}3884 \pm 0{,}365$	$6{,}0723 \pm 0{,}0001$
#13	$7{,}6362 \pm 0{,}0824$	**7,558** $\pm 0{,}00004$	$8{,}7269 \pm 0{,}2127$	$7{,}558 \pm 0{,}00003$
#14	$21{,}8311 \pm 0{,}1574$	**21,6353** $\pm 0{,}0002$	$25{,}12 \pm 0{,}755$	$21{,}6363 \pm 0{,}0009$
#15	$6{,}1889 \pm 0{,}1068$	**6,0319** $\pm 0{,}0008$	$6{,}8544 \pm 0{,}2361$	$6{,}0329 \pm 0{,}0016$

V.3.2 - Modification du protocole automatisé par inclusion de CAM

Le protocole automatisé présenté dans le second article (Chapitre IV) a été développé en utilisant KM comme méthode de classification non-supervisée. Bien que dans ce protocole 20 réplications de KM soient appliquées, aucune garantie n'est donnée quant à sa

convergence vers le minimum global. C'est pourquoi nous avons remplacé KM par CAM pour définir un protocole automatisé et optimal.

Ce nouveau protocole a été appliqué au niveau multi-images à l'échelle intra-individuelle de chacun des 15 patients. En parallèle, le protocole automatisé (basé sur KM) présenté dans le second article (Chapitre IV) a également été appliqué. Les nombres de classes ainsi estimés par ces deux protocoles sont présentés dans le Tableau V.2. La comparaison entre les résultats de ces deux protocoles a été faite en calculant le taux de bonne classification (classification accuracy rate) défini dans la section II.4 du second article.

TABLE V.2 – Nombre de classes optimaux k_{opt} estimé par les deux protocoles automatisés basés sur CAM et KM. La similarité entre les partitions estimées par ces deux méthodes a été calculée par le taux de bonne classification τ.

Patient	CAM	KM	τ (%)
#1	14	14	100
#2	16	16	100
#3	11	11	100
#4	12	12	100
#5	18	17	97
#6	13	15	89
#7	8	8	100
#8	11	11	100
#9	14	14	100
#10	7	7	100
#11	22	22	90
#12	10	10	100
#13	14	14	100
#14	14	14	100
#15	10	10	100

Pour 13 patients, les nombres de classes estimés par les deux protocoles sont identiques. De plus, lorsque ce nombre de classes est différent, les partitions correspondantes présentent une très forte similarité ($\tau \geqslant 89\%$). Les deux protocoles automatisés sont donc optimaux. Cependant, étant basé sur les algorithmes génétiques, CAM présente un temps de calcul beaucoup plus grand que celui de KM. Appliqué sur des images spectrales IR volumineuses, le protocole automatisé basé sur KM sera donc préféré.

Chapitre VI

Conclusion et perspectives

Dans le but d'automatiser le choix du nombre de classes de KM, la double application hiérarchique d'indices de validité a été développée pour des images spectrales d'échantillons de côlon. Grâce à cette approche proposée dans le premier article (Chapitre III), les structures histologiques du côlon peuvent être différenciées de façon automatique. A notre connaissance, les deux seules tentatives d'automatisation du nombre de classes en classification non-supervisée ont porté sur une application unique des indices de validité en FCM. Ainsi, Wang *et al.*[81] en 2007 et plus récemment dans notre laboratoire Sebiskveradze *et al.*[72] en 2011 ont échoué dans la réalisation d'une histologie spectrale automatisée de ganglions lymphatiques et de tumeurs cutanées. De plus, aucune étude d'imagerie spectrale n'a été publiée sur l'utilisation des indices de validité en KM ou HCA.

Les résultats de notre travail de thèse ont montré de façon explicite l'intérêt d'une double application hiérarchique des indices de validité en imagerie spectrale IR. Cependant, cette méthodologie présente certaines limites. En effet, nous avons constaté que pour moins de 10% des images, certaines structures histologiques du côlon sont retrouvées mêlées à d'autres structures. Par exemple, concernant le patient #14, la partition estimée montre que le chorion et la musculaire muqueuse appartiennent à la même classe rouge (Figure VI.1). Afin de retrouver ces deux structures histologiques, nous avons appliqué une troisième fois KM et PBM sur les pixels de cette classe rouge, avec k allant de 2 à 20. Comme le montre la Figure VI.2A, cette classe a été partitionnée par PBM en 2 sous-classes majeures (magenta et marron) aisément attribuables aux chorion et à la musculaire muqueuse ; une troisième classe résiduelle (vert foncé) semble être artéfactuelle.

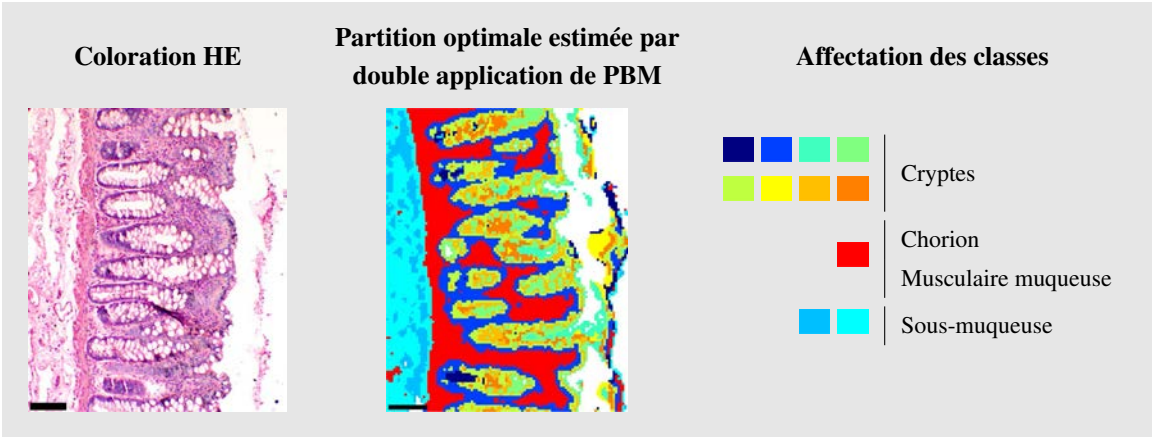


FIGURE VI.1 – Exemple d'une identification imparfaite de structures histologiques estimées par double application hiérarchique de PBM. Les structures du chorion et de la musculaire muqueuse sont représentées par une seule et même classe. Échelles, 100 μm.

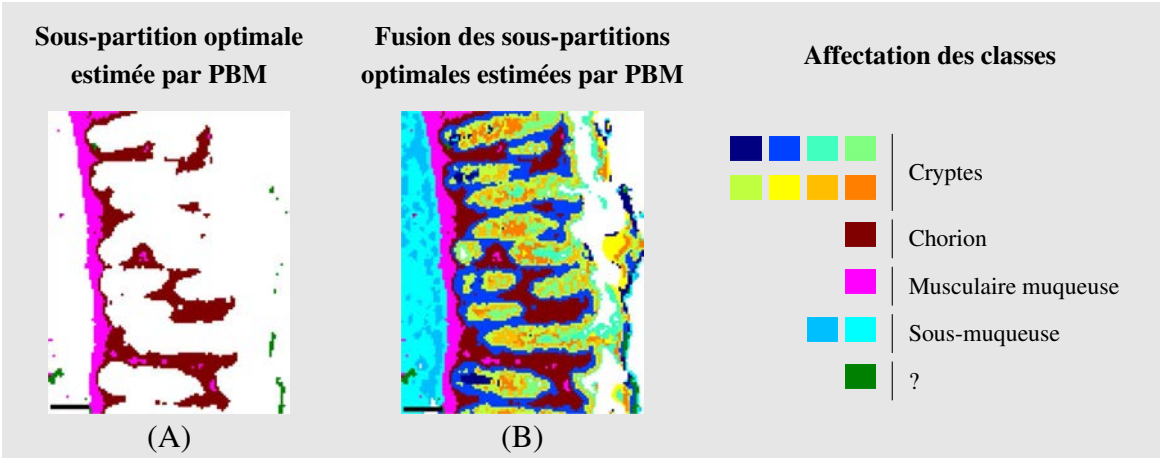


FIGURE VI.2 – Résultat d'une troisième application d'indice PBM sur une classe sélectionnée. (A) L'application de l'indice PBM sur la classe sélectionnée, représentant une mixture du chorion et de la musculaire muqueuse. (B) L'image en pseudo-couleur est construite en combinant les estimations des applications de l'indice PBM. Échelles, 100 μm.

Ces sous-partitions estimées par un troisième niveau hiérarchique de PBM ont été ensuite fusionnées aux résultats obtenus par la double application de PBM permettant ainsi de retrouver toutes les structures histologiques attendues (Figure VI.2B).

Par ailleurs, une troisième application de PBM est capable de révéler des informations supplémentaires comme le montre l'analyse des îlots lymphoïdes (Figure VI.3). En effet, comparativement aux données obtenues par la double application de PBM et par la coloration HE, on peut détecter la présence du tissu de soutien, le réseau réticulaire, et une grande hétérogénéité "structurée" du tissu lymphoïde. Cette hétérogénéité pourrait correspondre aux différentes sous-populations de cellules immunitaires présentes au sein

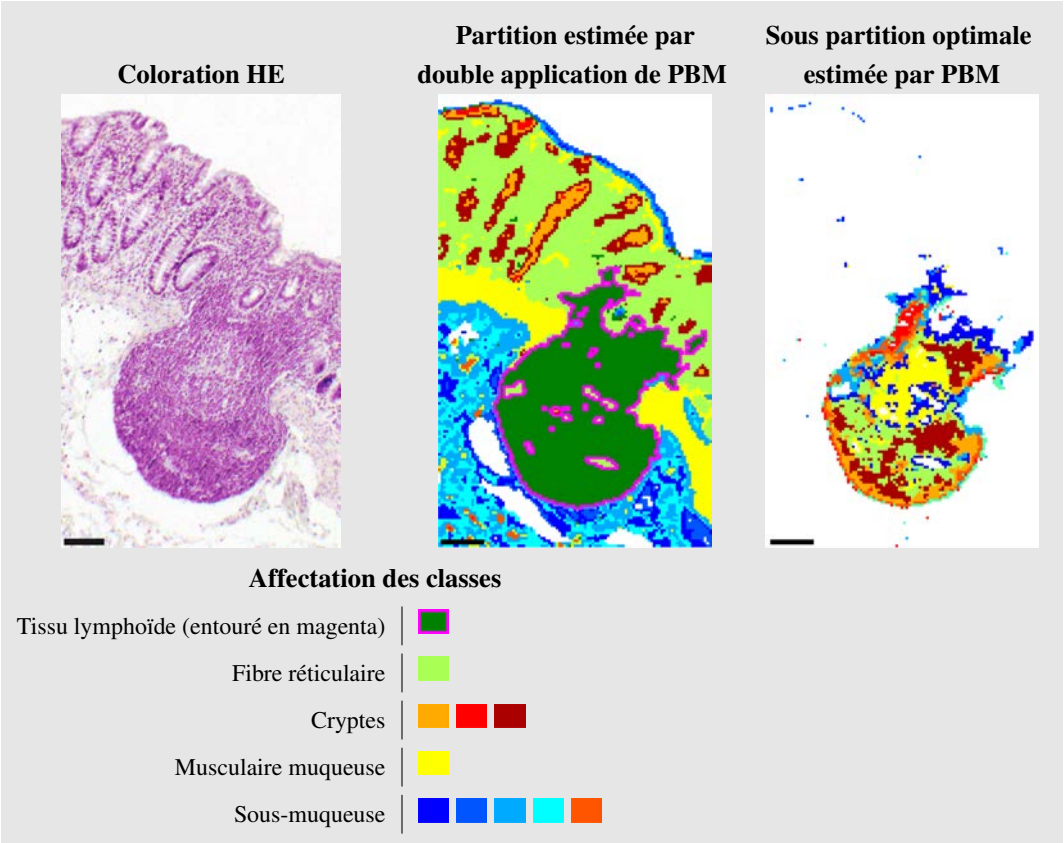


FIGURE VI.3 – Exemple de l'identification des sous-structures d'un îlot lymphoïde par double application hiérarchique de PBM. Échelles, 100 μm.

de l'îlot lymphoïde. Il faut souligner que sur le plan fonctionnel, l'îlot lymphoïde pourrait jouer un rôle important dans la relation entre inflammation et cancer colorectal[74].

Les deux situations précédentes illustrent parfaitement le fait que le nombre d'applications hiérarchiques des indices de validité est dépendant de la structure des données spectrales à analyser. Actuellement, nous tentons de moduler de façon automatique ce nombre en fonction du cluster considéré, donc de la structure histologique sous-jacente. Cette approche est basée sur une analyse de la distribution du nombre optimal de classes k_{opt} estimé par PBM, sur différents sous-ensembles de spectres tirés aléatoirement.

Les processus bayésiens hiérarchiques de Dirichlet récemment décrits[75] pourraient constituer une nouvelle approche d'histologie spectrale automatisée. Dans le cadre d'une classification non-supervisée, ces approches probabilistes modélisent parfaitement la problématique du clustering hiérarchique automatisé. En effet, chaque cluster est considéré comme un modèle de mélange où chaque composante est distribuée selon un processus de Dirichlet. De plus, le caractère non-paramétrique des processus de Dirichlet permet de s'af-

franchir de la connaissance *a priori* du nombre de classes. L'adaptation de ces méthodes à notre thématique nécessite toutefois la mise en place d'une collaboration étroite avec des statisticiens spécialisés.

* * *

Dans le second article (Chapitre IV), nous avons étendu l'étude précédente à une histologie spectrale multi-images. Nous sommes donc en mesure d'analyser simultanément un grand nombre d'images spectrales à l'échelle intra- et inter-individuelle, et d'en extraire les structures tissulaires communes. Dans des travaux précédents sur l'imagerie Raman et IRTF du côlon chez le nouveau-né, Krafft *et al.*[42] ont déclaré que la variabilité spectrale inter-individuelle pouvait être un obstacle majeur à l'application des techniques de classification non-supervisée ; excluant ainsi toute analyse simultanée d'un échantillon de plusieurs individus. Nos résultats contredisent cette étude puisque notre méthodologie permet de s'affranchir de cette variabilité inter-individuelle.

En conclusion, sur le plan anatomo-pathologique, notre approche simplifie considérablement la comparaison structurale des échantillons biologiques, et facilitera nécessairement l'interprétation de leur statut physiopathologique.

Étant donné qu'en histologie conventionnelle après coloration HE, les côlons murin et humain présentent de fortes ressemblances structurales (Figures II.3 et II.5), nous avons procédé à une histologie spectrale conjointe sur ces deux types de coupes. De façon surprenante, l'exploitation simultanée de données spectrales d'espèces différentes révèle que la quasi-totalité des structures histologiques sont communes aux deux espèces, indiquant une même "base" (socle) d'informations biomoléculaires. Par exemple, comme le montre la Figure VI.4, les cryptes murines et humaines sont représentées en bleu, marron, rouge et vert.

Chez le patient, ces résultats originaux pourraient être intéressants pour caractériser des phases importantes de la progression tumorale comme celles de la transition polype-cancer, adénome-carcinome et carcinome-métastase. L'analyse conjointe par histologie spectrale de modèles murins spécifiques de ces états pathologiques et d'échantillons de

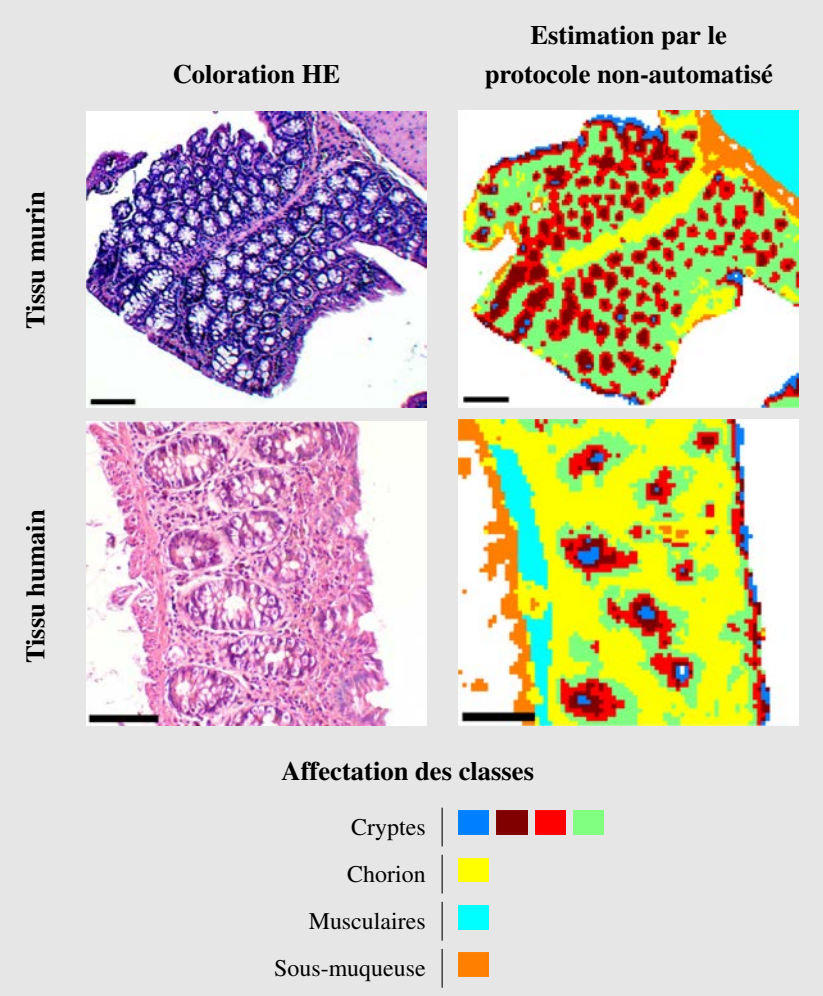


FIGURE VI.4 – Exemple d'histologie spectrale inter-espèce par le protocole non-automatisé, avec $k = 7$ classes. Échelles, 100 μm.

patients permettrait d'extraire des signatures spectrales communes. A terme, il est envisageable que cette imagerie spectrale inter-espèce puisse venir appuyer les méthodes conventionnelles de suivi du statut tumoral du patient.

* * *

Dans les travaux précédents, le clustering a été réalisé par KM qui est une méthode de recherche locale. Par conséquent, des initialisations multiples de KM appliqué sur des images spectrales IR induisent inévitablement, au niveau des classes, des répartitions différentes et donc des attributions différentes aux structures tissulaires. Afin de surmonter ces limitations, nous avons développé dans le troisième article (Chapitre V) un algorithme

mémétique réalisant une optimisation globale du problème de clustering. Appliqué aux images spectrales IR du côlon, notre méthode permet d'estimer une partition optimale, indépendante de l'initialisation de notre algorithme et surpassant KM et deux autres métaheuristiques en terme d'optimalité, de temps de calcul et de reproductibilité.

En histologie spectrale IR, la majorité des études exploite toutes les informations contenues dans une gamme spectrale allant de 900 à 1800 cm^{-1}, constituant la "bio-fingerprint region"[79]. Or, comme l'ont montré certains auteurs[63,79], il est possible d'extraire de manière supervisée, au sein de cette gamme spectrale, des bio-marqueurs plus spécifiques des structures histologiques étudiées. En se basant sur ce concept d'extraction de bio-marqueurs, nous avons développé un algorithme génétique de sélection non-supervisée de variables dédié à l'imagerie spectrale IR. Chaque chromosome est composé de N gènes codant les nombres d'onde sélectionnés ; N étant choisi actuellement par l'utilisateur. Brièvement, après initialisation des chromosomes, chaque itération de l'algorithme comprend les étapes suivantes :

- Pour chaque chromosome, application de KM sur les données réduites à ses nombres d'onde (gènes) ;
- Pour chaque chromosome, calcul de la "fitness" qui représente l'homogénéité spatiale de la partition KM estimée ;
- Sélection des chromosomes avec une probabilité proportionnelle à la "fitness" ;
- Croisement des chromosomes sélectionnés ;
- Mutation de la nouvelle génération de chromosomes.

La Figure VI.5 présente les résultats préliminaires de cette approche appliquée sur une image spectrale IR de côlon humain, et montre que les structures histologiques sont correctement retrouvées. L'objectif sous-jacent à long terme de l'application de métaheuristiques aux données spectrales IR est de mettre au point une méthode de classification non-supervisée optimale multi-objectifs, capable d'une estimation conjointe et automatique du nombre de classes et des variables d'intérêt.

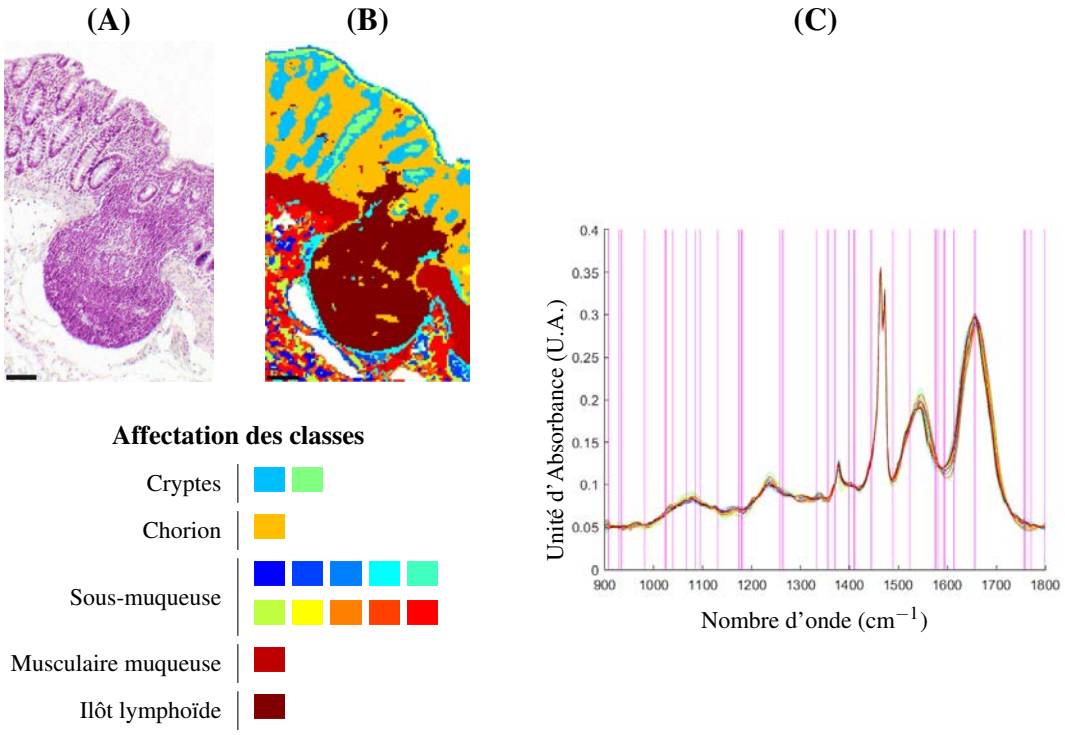


FIGURE VI.5 – Extraction non-supervisée de bio-marqueurs par algorithme génétique. (A) Coloration HE. (B) Image en pseudo-couleurs reconstruite après application de KM avec $k = 15$, sur le jeu de données réduit aux 30 variables sélectionnées par l'algorithme génétique. (C) Localisation sur les centroïdes, des variables sélectionnées (en magenta). Échelles, 100 μm.

* * *

L'ensemble de ce travail d'histologie spectrale IR constitue une base solide pour investiguer, dans le futur, certaines étapes du processus de cancérogénèse colique chez l'Homme. Le programme de recherche envisagé repose sur des résultats précédents de notre équipe[77,83] révélant par imagerie spectrale IR une hétérogénéité tumorale marquée au sein de xénogreffes de carcinome colique. Il visera à déterminer si cette hétérogénéité spectrale, non détectable par analyse histologique conventionnelle, est porteuse de valeurs pronostique et prédictive, quant au développement tumoral et à la réponse à la chimiothérapie. Notre objectif sera de construire une base de données de marqueurs spectroscopiques, à valeur pronostique et prédictive, spécifiques des cancers coliques.

Bibliographie

[1] N. K. Afseth and A. Kohler, *Extended multiplicative signal correction in vibrational spectroscopy, a tutorial*, Chemometrics and Intelligent Laboratory Systems, 117 (2012), pp. 92–99.

[2] S. M. Ali, F. Bonnier, H. Lambkin, K. Flynn, V. McDonagh, C. Healy, T. C. Lee, F. M. Lyng, and H. J. Byrne, *A comparison of Raman, FTIR and ATR-FTIR micro spectroscopy for imaging human skin tissue sections*, Analytical Methods, 5 (2013), pp. 2281–2291.

[3] M. J. Baker, J. Trevisan, P. Bassan, R. Bhargava, H. J. Butler, K. M. Dorling, P. R. Fielden, S. W. Fogarty, N. J. Fullwood, K. a. Heys, C. Hughes, P. Lasch, P. L. Martin-Hirsch, B. Obinaju, G. D. Sockalingum, J. Sulé-Suso, R. J. Strong, M. J. Walsh, B. R. Wood, P. Gardner, and F. L. Martin, *Using Fourier transform IR spectroscopy to analyze biological materials*, Nature Protocols, 9 (2014), pp. 1771–1791.

[4] K. R. Bambery, B. R. Wood, and D. McNaughton, *Resonant Mie scattering (RMieS) correction applied to FTIR images of biological tissue samples*, Analyst, 137 (2012), pp. 126–132.

[5] A. Banas, K. Banas, A. Furgal-Borzych, W. M. Kwiatek, B. Pawlicki, and M. B. H. Breese, *The pituitary gland under infrared light – in search of a representative spectrum for homogeneous regions*, Analyst, 140 (2015), pp. 2156–2163.

[6] P. BASSAN, J. LEE, A. SACHDEVA, J. PISSARDINI, K. M. DORLING, J. S. FLETCHER, A. HENDERSON, AND P. GARDNER, *The inherent problem of transflection-mode infrared spectroscopic microscopy and the ramifications for biomedical single point and imaging applications*, Analyst, 138 (2013), pp. 144–157.

[7] P. BASSAN, A. SACHDEVA, J. H. SHANKS, M. D. BROWN, N. W. CLARKE, AND P. GARDNER, *Automated high-throughput assessment of prostate biopsy tissue using infrared spectroscopic chemical imaging*, in SPIE Medical Imaging, 2014, p. 90410D.

[8] C. BELEITES, G. STEINER, M. SOWA, R. BAUMGARTNER, S. SOBOTTKA, G. SCHACKERT, AND R. SALZER, *Classification of human gliomas by infrared imaging spectroscopy and chemometric image processing*, Vibrational Spectroscopy, 38 (2005), pp. 143–149.

[9] A. BELJEBBAR, S. DUKIC, N. AMHARREF, S. BELLEFQIH, AND M. MANFAIT, *Monitoring of Biochemical Changes through the C6 gliomas progression and invasion by Fourier Transform Infrared (FTIR) imaging*, Analytical Chemistry, 81 (2009), pp. 9247–9256.

[10] A. BENARD, C. DESMEDT, M. SMOLINA, P. SZTERNFELD, M. VERDONCK, G. ROUAS, N. KHEDDOUMI, F. ROTHÉ, D. LARSIMONT, C. SOTIRIOU, AND E. GOORMAGHTIGH, *Infrared imaging in breast cancer : automated tissue component recognition and spectral characterization of breast cancer cells as well as the tumor microenvironment*, Analyst, 139 (2014), pp. 1044–1056.

[11] N. BERGNER, B. F. M. ROMEIKE, R. REICHART, R. KALFF, C. KRAFFT, AND J. POPP, *Tumor margin identification and prediction of the primary tumor from brain metastases using FTIR imaging and support vector machines*, Analyst, 138 (2013), pp. 3983–3990.

[12] J. C. BEZDEK AND N. R. PAL, *Some new indexes of cluster validity*, IEEE Transactions on Systems, Man, and Cybernetics, Part B : Cybernetics, 28 (1998), pp. 301–315.

[13] B. BIRD, M. MILJKOVIĆ, S. REMISZEWSKI, A. AKALIN, M. KON, AND M. DIEM, *Infrared spectral histopathology (SHP) : a novel diagnostic tool for*

the accurate classification of lung cancer, Laboratory Investigation, 92 (2012), pp. 1358–1373.

[14] E. M. BURKA AND R. CURBELO, *Imaging ATR spectrometer*, 2000.

[15] T. CALINSKI AND J. HARABASZ, *A Dendrite Method for Cluster Analysis*, Communications in Statistics - Simulation and Computation, 3 (1974), pp. 1–27.

[16] P. CAPPER AND C. ELLIOTT, *Infrared Detectors and Emitters : Materials and Devices*, vol. 8, Springer Science & Business Media, 2013.

[17] C.-H. CHOU, M.-C. SU, AND E. LAI, *A new cluster validity measure and its application to image compression*, Pattern Analysis and Applications, 7 (2004), pp. 205–220.

[18] W. W. COBLENTZ, *Optical Notes*, Physical Review (Series I), 19 (1904), pp. 89–97.

[19] C. S. COLLEY, S. G. KAZARIAN, P. D. WEINBERG, AND M. J. LEVER, *Spectroscopic imaging of arteries and atherosclerotic plaques*, Biopolymers, 74 (2004), pp. 328–335.

[20] L. D. DAVID AND W. B. DONALD, *A Cluster Separation Measure*, IEEE Transactions on Pattern Analysis and Machine Intelligence, 1 (1979), pp. 224–227.

[21] M. DIEM, L. CHIRIBOGA, AND H. YEE, *Infrared spectroscopy of human cells and tissue. VIII. Strategies for analysis of infrared tissue mapping data and applications to liver tissue*, Biopolymers, 57 (2000), pp. 282–290.

[22] M. DIEM, A. MAZUR, K. LENAU, J. SCHUBERT, B. BIRD, M. MILJKOVIĆ, C. KRAFFT, AND J. POPP, *Molecular pathology via IR and Raman spectral imaging*, Journal of Biophotonics, 6 (2013), pp. 855–886.

[23] J. C. DUNN, *Well-Separated Clusters and Optimal Fuzzy Partitions*, Journal of Cybernetics, 4 (1974), pp. 95–104.

[24] J. EINENKEL, U.-D. BRAUMANN, W. STELLER, H. BINDER, AND L.-C. HORN, *Suitability of infrared microspectroscopic imaging for histopathology of the uterine cervix*, Histopathology, 60 (2012), pp. 1084–1098.

[25] C. EKLOUH-MOLINIER, T. HAPPILLON, N. BOULAND, C. FICHEL, M.-D. DIÉBOLD, J.-F. ANGIBOUST, M. MANFAIT, S. BRASSART-PASCO, AND O. PIOT, *Investigating the relationship between changes in collagen fiber orientation during skin aging and collagen/water interactions by polarized-FTIR microimaging*, Analyst, 140 (2015), pp. 6260–6268.

[26] H. FABIAN, N. A. N. THI, M. EIDEN, P. LASCH, J. SCHMITT, AND D. NAUMANN, *Diagnosing benign and malignant lesions in breast tissue sections by using IR-microspectroscopy*, Biochimica et Biophysica Acta (BBA)-Biomembranes, 1758 (2006), pp. 874–882.

[27] J. FAHRENFORT, *Attenuated total reflection : A new principle for the production of useful infra-red reflection spectra of organic compounds*, Spectrochimica Acta, 17 (1961), pp. 698–709.

[28] F. GROSSERÜSCHKAMP, A. KALLENBACH-THIELTGES, T. BEHRENS, T. BRÜNING, M. ALTMAYER, G. STAMATIS, D. THEEGARTEN, AND K. GERWERT, *Marker-free automated histopathological annotation of lung tumour subtypes by FTIR imaging*, Analyst, 140 (2015), pp. 2114–2120.

[29] I. GURRUTXAGA, I. ALBISUA, O. ARBELAITZ, J. I. MARTÍN, J. MUGUERZA, J. M. PÉREZ, AND I. PERONA, *SEP/COP : An efficient method to find the best partition in hierarchical clustering based on a new cluster validity index*, Pattern Recognition, 43 (2010), pp. 3364–3373.

[30] A. S. HAKA, L. H. KIDDER, AND E. N. LEWIS, *Combined spectroscopic imaging and chemometric approach for automatically partitioning tissue types in human prostate tissue biopsies*, in BiOS 2001 The International Symposium on Biomedical Optics, 2001, pp. 47–55.

[31] M. HALKIDI AND M. VAZIRGIANNIS, *Clustering Validity Assessment : Finding the optimal partitioning of a data set*, in Data Mining, 2001. ICDM 2001, Proceedings IEEE International Conference on, 2001, pp. 187–194.

[32] R. J. HATHAWAY, J. W. DAVENPORT, AND J. C. BEZDEK, *Relational duals of the c-means clustering algorithms*, Pattern Recognition, 22 (1989), pp. 205–212.

[33] P. HERAUD, S. CAINE, N. CAMPANALE, T. KARNEZIS, D. MCNAUGHTON, B. R. WOOD, M. J. TOBIN, AND C. C. A. BERNARD, *Early detection of the chemical changes occurring during the induction and prevention of autoimmune-mediated demyelination detected by FT-IR imaging*, NeuroImage, 49 (2010), pp. 1180–1189.

[34] L. J. HUBERT AND J. R. LEVIN, *A general statistical framework for assessing categorical clustering in free recall*, Psychological Bulletin, 83 (1976), pp. 1072–1080.

[35] C. HUGHES, L. GAUNT, M. BROWN, N. W. CLARKE, AND P. GARDNER, *Assessment of paraffin removal from prostate FFPE sections using transmission mode FTIR-FPA imaging*, Analytical Methods, 6 (2014), pp. 1028–1035.

[36] T. HUSER AND J. CHAN, *Raman spectroscopy for physiological investigations of tissues and cells*, Advanced Drug Delivery Reviews, 89 (2015), pp. 57–70.

[37] A. K. JAIN, *Data clustering : 50 years beyond K-means*, Pattern Recognition Letters, 31 (2010), pp. 651–666.

[38] L. H. KIDDER, I. W. LEVIN, E. N. LEWIS, V. D. KLEIMAN, AND E. J. HEILWEIL, *Mercury cadmium telluride focal-plane array detection for mid-infrared Fourier-transform spectroscopic imaging*, Optics Letters, 22 (1997), pp. 742–744.

[39] M. KIM AND R. S. RAMAKRISHNA, *New indices for cluster validity assessment*, Pattern Recognition Letters, 26 (2005), pp. 2353–2363.

[40] K. KOCHAN, P. HERAUD, M. KIUPEL, V. YUZBASIYAN-GURKAN, D. MCNAUGHTON, M. BARANSKA, AND B. R. WOOD, *Comparison of FTIR transmission and transfection substrates for canine liver cancer detection*, Analyst, 140 (2015), pp. 2402–2411.

[41] M. R. KOLE, R. K. REDDY, M. V. SCHULMERICH, M. K. GELBER, AND R. BHARGAVA, *Discrete Frequency Infrared Microspectroscopy and Imaging with a Tunable Quantum Cascade Laser*, Analytical Chemistry, 84 (2012), pp. 10366–10372.

[42] C. KRAFFT, D. CODRICH, G. PELIZZO, AND V. SERGO, *Raman and FTIR microscopic imaging of colon tissue : A comparative study*, Journal of Biophotonics, 1 (2008), pp. 154–169.

[43] C. KRAFFT, L. SHAPOVAL, S. B. SOBOTTKA, K. D. GEIGER, G. SCHACKERT, AND R. SALZER, *Identification of primary tumors of brain metastases by SIMCA classification of IR spectroscopic images*, Biochimica et Biophysica Acta (BBA)-Biomembranes, 1758 (2006), pp. 883–891.

[44] C. KRAFFT, L. SHAPOVAL, S. B. SOBOTTKA, G. SCHACKERT, AND R. SALZER, *Identification of primary tumors of brain metastases by infrared spectroscopic imaging and linear discriminant analysis*, Technology in cancer research & treatment, 5 (2006), pp. 291–298.

[45] S. KUMAR, C. DESMEDT, D. LARSIMONT, C. SOTIRIOU, AND E. GOORMAGHTIGH, *Change in the microenvironment of breast cancer studied by FTIR imaging*, Analyst, 138 (2013), pp. 4058–4065.

[46] J. T. KWAK, A. KAJDACSY-BALLA, V. MACIAS, M. WALSH, S. SINHA, AND R. BHARGAVA, *Improving prediction of prostate cancer recurrence using chemical imaging*, Scientific reports, 5 (2015), p. 8758.

[47] J. M. KWIATKOSKI AND J. A. REFFNER, *FT-IR microspectrometry advances*, Nature, 328 (1987), pp. 837–838.

[48] P. LASCH, M. DIEM, W. HÄNSCH, AND D. NAUMANN, *Artificial neural networks as supervised techniques for FT-IR microspectroscopic imaging*, Journal of Chemometrics, 20 (2007), pp. 209–220.

[49] P. LASCH, W. HAENSCH, D. NAUMANN, AND M. DIEM, *Imaging of colorectal adenocarcinoma using FT-IR microspectroscopy and cluster analysis*, Biochimica et Biophysica Acta (BBA) - Molecular Basis of Disease, 1688 (2004), pp. 176–186.

[50] L. S. LESLIE, T. P. WROBEL, D. MAYERICH, S. BINDRA, R. EMMADI, AND R. BHARGAVA, *High definition infrared spectroscopic imaging for lymph node histopathology*, PloS ONE, 10 (2015), p. e0127238.

[51] I. W. LEVIN AND R. BHARGAVA, *FOURIER TRANSFORM INFRARED VIBRATIONAL SPECTROSCOPIC IMAGING : Integrating Microscopy and Molecular Recognition*, Annual Review of Physical Chemistry, 56 (2005), pp. 429–474.

[52] E. N. Lewis, P. J. Treado, R. C. Reeder, G. M. Story, A. E. Dowrey, C. Marcott, and I. W. Levin, *Fourier Transform Spectroscopic Imaging Using an Infrared Focal-Plane Array Detector*, Analytical Chemistry, 67 (1995), pp. 3377–3381.

[53] E. Ly, O. Piot, A. Durlach, P. Bernard, and M. Manfait, *Differential diagnosis of cutaneous carcinomas by infrared spectral micro-imaging combined with pattern recognition*, Analyst, 134 (2009), pp. 1208–1214.

[54] E. Ly, O. Piot, R. Wolthuis, A. Durlach, P. Bernard, and M. Manfait, *Combination of FTIR spectral imaging and chemometrics for tumour detection from paraffin-embedded biopsies*, Analyst, 133 (2008), pp. 197–205.

[55] J. B. MacQueen, *Some Methods for classification and Analysis of Multivariate Observations*, Proceedings of the fifth Berkeley symposium on mathematical statistics and probability, 1 (1967), pp. 281–297.

[56] H. H. Mantsch, *The road to medical vibrational spectroscopy-a history*, Analyst, 138 (2013), pp. 3863–3870.

[57] L. M. Miller and P. Dumas, *Chemical imaging of biological tissue with synchrotron infrared light*, Biochimica et Biophysica Acta (BBA) - Biomembranes, 1758 (2006), pp. 846–857.

[58] X. Mu, M. Kon, A. Ergin, S. Remiszewski, A. Akalin, C. M. Thompson, and M. Diem, *Statistical analysis of a lung cancer spectral histopathology (SHP) data set*, Analyst, 140 (2015), pp. 2449–2464.

[59] J. Nallala, M.-D. Diebold, C. Gobinet, O. Bouché, G. D. Sockalingum, O. Piot, and M. Manfait, *Infrared spectral histopathology for cancer diagnosis : a novel approach for automated pattern recognition of colon adenocarcinoma*, Analyst, 139 (2014), pp. 4005–4015.

[60] J. Nallala, C. Gobinet, M.-D. Diebold, V. Untereiner, O. Bouché, M. Manfait, G. D. Sockalingum, and O. Piot, *Infrared spectral imaging as a novel approach for histopathological recognition in colon cancer diagnosis*, Journal of Biomedical Optics, 17 (2012), p. 116013.

[61] J. NALLALA, G. R. LLOYD, N. SHEPHERD, AND N. STONE, *High-resolution FTIR imaging of colon tissues for elucidation of individual cellular and histopathological features*, Analyst, (2015).

[62] J. NALLALA, G. R. LLOYD, AND N. STONE, *Evaluation of different tissue de-paraffinization procedures for infrared spectral imaging*, Analyst, 140 (2015), pp. 2369–2375.

[63] J. NALLALA, O. PIOT, M.-D. DIEBOLD, C. GOBINET, O. BOUCHÉ, M. MANFAIT, AND G. D. SOCKALINGUM, *Infrared imaging as a cancer diagnostic tool : introducing a new concept of spectral barcodes for identifying molecular changes in colon tumors*, Cytometry. Part A : the journal of the International Society for Analytical Cytology, 83 (2013), pp. 294–300.

[64] M. J. NASSE, M. J. WALSH, E. C. MATTSON, R. REININGER, A. KAJDACSY-BALLA, V. MACIAS, R. BHARGAVA, AND C. J. HIRSCHMUGL, *High-resolution Fourier-transform infrared chemical imaging with multiple synchrotron beams*, Nature Methods, 8 (2011), pp. 413–416.

[65] M. K. PAKHIRA, S. BANDYOPADHYAY, AND U. MAULIK, *Validity index for crisp and fuzzy clusters*, Pattern Recognition, 37 (2004), pp. 487–501.

[66] D. PEREZ-GUAITA, P. HERAUD, K. M. MARZEC, M. DE LA GUARDIA, M. KIUPEL, AND B. R. WOOD, *Comparison of transflection and transmission FTIR imaging measurements performed on differentially fixed tissue sections*, Analyst, 140 (2015), pp. 2376–2382.

[67] R. K. REDDY AND R. BHARGAVA, *Accurate histopathology from low signal-to-noise ratio spectroscopic imaging data*, Analyst, 135 (2010), pp. 2818–2825.

[68] P. J. ROUSSEEUW, *Silhouettes : A graphical aid to the interpretation and validation of cluster analysis*, Journal of Computational and Applied Mathematics, 20 (1987), pp. 53–65.

[69] S. SAHA AND S. BANDYOPADHYAY, *A symmetry based multiobjective clustering technique for automatic evolution of clusters*, Pattern Recognition, 43 (2010), pp. 738–751.

[70] R. SALZER, H. H. MANTSCH, J. MANSFIELD, E. N. LEWIS, AND G. STEINER, *Infrared and Raman imaging of biological and biomimetic samples*, Fresenius' Journal of Analytical Chemistry, 366 (2000), pp. 712–726.

[71] M. SATTLECKER, N. STONE, AND C. BESSANT, *Current trends in machine-learning methods applied to spectroscopic cancer diagnosis*, TrAC - Trends in Analytical Chemistry, 59 (2014), pp. 17–25.

[72] D. SEBISKVERADZE, V. VRABIE, C. GOBINET, A. DURLACH, P. BERNARD, E. LY, M. MANFAIT, P. JEANNESSON, AND O. PIOT, *Automation of an algorithm based on fuzzy clustering for analyzing tumoral heterogeneity in human skin carcinoma tissue sections*, Laboratory Investigation, 91 (2011), pp. 799–811.

[73] C. H. SEQUIN AND M. F. TOMPSETT, *Charge transfer devices*, Academic Press, New York, (1975).

[74] F. SIPOS AND G. MUZES, *Isolated lymphoid follicles in colon : Switch points between inflammation and colorectal cancer ?*, World Journal of Gastroenterology, 17 (2011), pp. 1666–1673.

[75] Y. W. TEH, M. I. JORDAN, M. J. BEAL, AND D. M. BLEI, *Hierarchical Dirichlet Processes*, Journal of the American Statistical Association, 101 (2006), pp. 1566–1581.

[76] S. TIWARI AND R. BHARGAVA, *Extracting Knowledge from Chemical Imaging Data Using Computational Algorithms for Digital Cancer Diagnosis*, The Yale journal of biology and medicine, 88 (2015), pp. 131–143.

[77] A. TRAVO, O. PIOT, R. WOLTHUIS, C. GOBINET, M. MANFAIT, J. BARA, M.-E. FORGUE-LAFITTE, AND P. JEANNESSON, *IR spectral imaging of secreted mucus : a promising new tool for the histopathological recognition of human colonic adenocarcinomas*, Histopathology, 56 (2010), pp. 921–931.

[78] P. M. TREUTING AND S. M. DINTZIS, *Comparative Anatomy and Histology : A Mouse and Human Atlas*, vol. 4, Academic Press, 2012.

[79] J. TREVISAN, P. P. ANGELOV, P. L. CARMICHAEL, A. D. SCOTT, AND F. L. MARTIN, *Extracting biological information with computational analysis of Fourier-*

transform infrared (FTIR) biospectroscopy datasets : current practices to future perspectives, Analyst, 137 (2012), pp. 3202–3215.

[80] L. VENDRAMIN, R. CAMPELLO, AND E. R. HRUSCHKA, *On the Comparison of Relative Clustering Validity Criteria*, in SDM, vol. 10, 2009, pp. 733–744.

[81] WANG, J. M. GARIBALDI, B. BIRD, AND M. W. GEORGE, *A novel fuzzy clustering algorithm for the analysis of axillary lymph node tissue sections*, Applied Intelligence, 27 (2007), pp. 237–248.

[82] G. P. WILLIAMS, *Infrared synchrotron radiation instrumentation and applications*, Review of Scientific Instruments, 63 (1992), pp. 1535–1538.

[83] R. WOLTHUIS, A. TRAVO, C. NICOLET, A. NEUVILLE, M.-P. GAUB, D. GUENOT, E. LY, M. MANFAIT, P. JEANNESSON, AND O. PIOT, *IR spectral imaging for histopathological characterization of xenografted human colon carcinomas*, Analytical Chemistry, 80 (2008), pp. 8461–8469.

[84] B. R. WOOD, K. R. BAMBERY, C. J. EVANS, M. A. QUINN, AND D. MCNAUGHTON, *A three-dimensional multivariate image processing technique for the analysis of FTIR spectroscopic images of multiple tissue sections*, BMC Medical Imaging, 6 (2006), p. 12.

[85] X. L. XIE AND G. BENI, *A validity measure for fuzzy clustering*, IEEE Transactions on Pattern Analysis and Machine Intelligence, 13 (1991), pp. 841–847.

[86] K. YEH, S. KENKEL, J.-N. LIU, AND R. BHARGAVA, *Fast Infrared Chemical Imaging with a Quantum Cascade Laser*, Analytical Chemistry, 87 (2015), pp. 485–493.

[87] K. R. ŽALIK AND B. ŽALIK, *Validity index for clusters of different sizes and densities*, Pattern Recognition Letters, 32 (2011), pp. 221–234.

NOUVEAUX DEVELOPPEMENTS EN HISTOLOGIE SPECTRALE IR : APPLICATION AU TISSU COLIQUE

Les développements continus en micro-spectroscopie vibrationnelle IR et en analyse numérique de données multidimensionnelles ont permis récemment l'émergence de l'histologie spectrale. A l'échelle tissulaire et sur une base biomoléculaire, cette nouvelle approche représente un outil prometteur pour une meilleure analyse et caractérisation de différents états physiopathologiques, et potentiellement une aide au diagnostic clinique. Dans ce travail, en utilisant un modèle tissulaire de côlon normal chez la Souris et chez l'Homme, nous avons apporté des améliorations à la chaîne de traitements des données afin d'automatiser et d'optimiser cette histologie spectrale.
En effet, dans un premier temps, le développement d'une double application hiérarchique d'indices de validité a permis de déterminer le nombre optimal de classes nécessaire à une caractérisation complète des structures histologiques. Dans un second temps, cette méthode a été généralisée à l'échelle interindividuelle par couplage d'un prétraitement par EMSC (Extended Multiplicative
Signal Correction) et d'une classification non-supervisée k-Means ; ce couplage étant appliqué conjointement à toutes les images spectrales IR. Enfin, compte tenu de l'essor des métaheuristiques et de leur capacité à résoudre des problèmes complexes d'optimisation numérique, nous avons transposé un algorithme mémétique aux données spectrales IR. Ce nouvel algorithme se compose d'un algorithme génétique et d'un raffinement par classification non-supervisée k-Means. Comparé aux méthodes classiques de clustering, cet algorithme mémétique appliqué aux images spectrales IR, a permis de réaliser une classification non-supervisée optimale et indépendante de l'initialisation.

Histologie spectrale IR, classification non-supervisée, k-Means, métaheuristique, algorithme mémétique, spectroscopie IR, histologie du côlon.

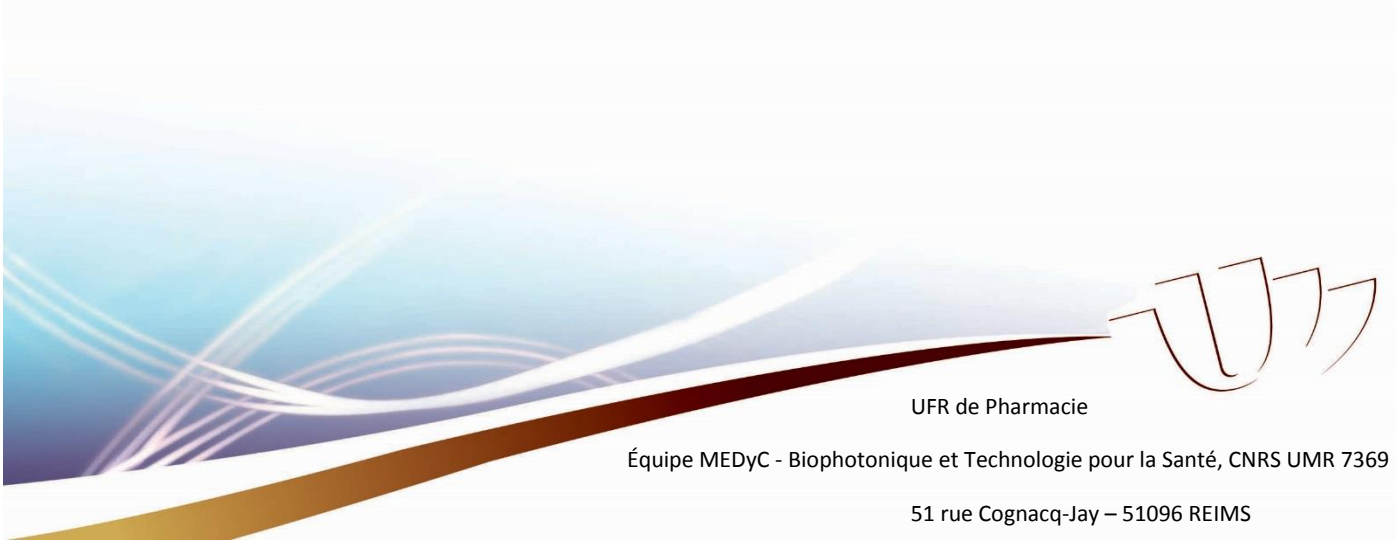
NEW DEVELOPMENTS IN IR SPECTRAL HISTOLOGY: APPLICATION TO COLON TISSUE

Recent developments in IR vibrational microspectroscopy and numerical multidimensional analysis have led to the emergence of spectral histology. At the tissue level, this new approach represents an attractive tool for a better analysis and characterization of pathophysiological states and for diagnostic challenges. Here, using normal murine and human colon tissues, data processing steps have been improved for automating and optimizing this spectral histology. First, the development of a hierarchical double application of validity indices permitted to determine the optimal number of clusters that correctly identified the different colon histological components.
Second, this method has been improved to perform spectral histology at the inter-individual level.
For this, EMSC (Extended Multiplicative Signal Correction) preprocessing has been successfully combined to k-Means clustering. Finally, given the ability of metaheuristics to solve complex optimization problems, a memetic algorithm has been developed for IR spectral data clustering. This algorithm is composed of a genetic algorithm and a k-Means clustering refinement. Compared with conventional clustering methods, our memetic algorithm allowed to generate an optimal and initialization-independent clustering.

IR spectral histology, clustering, k-Means, metaheuristic, memetic algorithm, IR spectroscopy, colon histology.

Discipline : PHYSIQUE

Spécialité : Bio-spectroscopie



UFR de Pharmacie

Équipe MEDyC - Biophotonique et Technologie pour la Santé, CNRS UMR 7369

51 rue Cognacq-Jay – 51096 REIMS