

UNIVERSITE DE REIMS CHAMPAGNE-ARDENNE

ECOLE DOCTORALE SCIENCES TECHNOLOGIE SANTE

THESE

Pour obtenir le grade de

DOCTEUR DE L'UNIVERSITE DE REIMS CHAMPAGNE-ARDENNE

Discipline : Informatique Biophysique

Présentée et soutenue publiquement le 12/12/2013

Par

Teddy HAPPILLON

Né le 25 mars 1984 à Soissons

Titre :

Aide au diagnostic de cancers cutanés et de la leucémie lymphoïde chronique par microspectroscopies vibrationnelles couplées à des analyses numériques multivariées

JURY

Pr. Michel MANFAIT	**Directeur de thèse**
Pr. Pierre CHAMINADE	**Rapporteur**
Pr. Robert SABATIER	**Rapporteur**
Pr. Michel HERBIN	**Président**
Pr. Xavier TROUSSARD	**Examinateur**
Dr. Jacques KLOSSA	**Examinateur**
Dr. Cyril GOBINET	**Examinateur**

Dans le monde il y a 10 types de personnes
Celles qui comprennent le binaire, et les autres

Remerciements

À Monsieur le Professeur Robert Sabatier

Je vous remercie très sincèrement d'avoir accepté de juger ce travail de thèse, et d'en être rapporteur scientifique. Veuillez trouver ici toute ma reconnaissance et soyez assuré de mon respect.

À Monsieur le Professeur Pierre Chaminade

Je vous suis très reconnaissant et vous remercie d'avoir accepté d'être rapporteur scientifique de ces travaux de thèse et de faire partie du jury de soutenance. Soyez assuré de toute ma reconnaissance et de mon respect.

À Monsieur le Professeur Xavier Troussard

Je vous exprime toute ma reconnaissance et vous remercie de faire partie de mon jury de thèse, mais également de l'intérêt que vous portez à mon travail et plus généralement pour votre contribution à l'élaboration du projet IHMO.

À Monsieur le Professeur Michel Herbin

Michel, je tiens très sincèrement à te remercier pour avoir tout de suite accepté de faire partie de mon jury de thèse, et sache que je suis également très honoré de présenter les résultats de mes travaux à un spécialiste du traitement d'image.

À Monsieur le Docteur Jacques Klossa

Jacques, je vous suis très reconnaissant de faire partie de mon jury de thèse, mais au-delà de ça, je tenais à vous remercier pour toute l'énergie que vous avez mise dans le projet IHMO et qui m'a permis de mener à bien des travaux qui m'ont passionné. Je vous remercie également de conjuguer professionnalisme, sympathie et bonne humeur à chaque fois que nous nous rencontrons.

À Monsieur le Professeur Michel Manfait

Je vous remercie de m'avoir accueilli au sein de votre laboratoire pour réaliser mon stage de master 2, et de m'avoir offert l'opportunité de réaliser une thèse à la fin de celui-ci. Je vous exprime également toute ma gratitude pour m'avoir encadré pendant ces 4 années de doctorat, pour ces discussions constructives et captivantes, et pour m'avoir permis, et même encouragé, à présenter mes travaux de recherche dans de nombreux congrès nationaux et internationaux.

À Monsieur le Docteur Cyril Gobinet

Cyril, je ne saurais combien te remercier pour tout ce que tu as pu m'apprendre pendant près de 5 ans. Tu m'as apporté tes connaissances de programmeur qui ont fait de moi un vrai Matlabiste, tu as toujours été disponible pour répondre à mes nombreuses questions, tu m'as laissé une précieuse autonomie et m'as réorienté dans les bons axes de recherche lorsque je faisais fausse route. Enfin, ta gentillesse, ta sympathie et ta bonne humeur quotidienne ont fait que toutes ces années sont passées en un éclair. Pour toutes ces raisons et bien d'autres, je te remercie très sincèrement.

Je tiens également à remercier

***Valérie Untereiner**, pour toutes ces connaissances que tu m'as apportées pendant ma thèse, pour ton travail toujours très rigoureux, pour ces semaines interminables que tu as passées dans le noir pour réaliser les acquisitions IHMO, mais également pour toutes ces discussions et nombreux éclats de rires que nous avons partagés dans le bureau, la cuisine et les couloirs, et pour ta gentillesse et ton aide précieuse, merci Valérie.*

***Abdelilah Beljebbar**, pour m'avoir apporté ton savoir en spectroscopie, que ce soit d'un point de vue matériel ou chimiométrique, mais également pour avoir pris beaucoup de temps à la correction d'article, je te remercie. Et comment ne pas te remercier pour ta sympathie, pour ton humour et pour ton rire plus que communicatif qui ont su rendre nos échanges très agréables, merci Abdel.*

***Sylvie Daliphard et Pascale Cornillet-Lefebvre**, je vous remercie très sincèrement pour votre investissement si fort dans le projet IHMO, pour m'avoir appris tant de choses en hématologie et ce, en restant pédagogues pour le néophyte que je suis. Je vous remercie également pour votre sympathie, votre gentillesse et votre disponibilité qui ont fait que j'ai pris un réel plaisir à travailler avec vous. Merci pour tout Sylvie et Pascale.*

***Irène Dupuis**, merci pour ton aide durant ces 4 années de thèse, mais surtout merci pour ta bonne humeur, pour les gâteaux dans la cuisine et les boutures de fleurs que je retrouvais sur mon bureau le matin.*

***Aurélie Trussardi-Régnier, Christine Terryn, Dominique Toubas, Frédéric Velard, Ganesh D. Sockalingum, Hamid Morjani, Jean-François Angiboust, Jennifer Geeraert, Laurence Van-Gulick, Olivier Piot, Pierre Jeannesson, Roselyne Garnotel, Valeriu Vrabie**, merci à tous pour votre aide, pour vos apports scientifiques et pour votre sympathie au cours de ces 4 années de doctorat.*

*Un immense merci à mes 3 drôles de dames, **Aurélie**, **Marie** et **Mathilde**, les furieuses de la biologie, qui m'ont été d'une grande aide pour la rédaction de ce manuscrit mais qui sont avant tout de grandes amies. Merci à vous les filles.*

*Mon super copain **Hadrien**, comment te remercier pour tous ces bons moments passés ensemble, au labo, au L-Bee, chez toi, chez moi, chez tout le monde (et dans un fameux salon parc des expos) ? Tu m'as soutenu moralement, tu m'as (re)mis au sport, tu as été plus que patient pour me supporter durant ma période de rédaction, bref tu as été un des moteurs de ma réussite. Et à ce titre, tu resteras un de mes plus grands amis. Je ne te remercierai jamais assez.*

*Je remercie également tous mes amis thésards, nouveaux docteurs et leur moitié, **Adrian**, **Caroline**, **Christophe**, **David**, **Elodie**, **Georges**, **Goutam**, **Imane**, **Jaya**, **Joan**, **Julien**, **Laura**, **Lucie**, **Mai Ahn**, **Mathieu**, **Michaël**, **Mohammed**, **Nathalie**, **Qué**, **The Thuong** et **Vincent**, avec qui j'ai passé des moments inoubliables au sein du labo mais aussi et surtout en dehors. Ces années n'auraient pas été les mêmes sans vous, merci à tous.*

*Je remercie profondément **mon frère**, **ma belle-sœur**, **ma moitié**, toute **ma famille** et **mes proches**, pour leur soutien et leurs encouragements, qui m'ont aidé à surmonter les moments difficiles.*

*Enfin, je remercie tout particulièrement et du fond du cœur, **mes parents**, qui toute leur vie se sont démenés pour que mon frère et moi, puissions aller le plus loin possible.*
Sans vous, je n'aurai jamais pu arriver jusqu'ici ! Merci.

Je dédie cette thèse à Andrée Happillon

Sommaire

LISTE DES ABRÉVIATIONS **7**

LISTE DES FIGURES **9**

PRÉAMBULE **10**

CHAPITRE 1 : INTRODUCTION ET GÉNÉRALITÉS **12**

I) PEAU ET CANCERS CUTANÉS 13

1) La peau *13*

a) L'épiderme 14

b) Le derme 16

c) L'hypoderme 17

2) Les cancers cutanés *17*

a) Description 17

b) Diagnostic clinique 19

c) Diagnostic histologique 21

d) Traitement des carcinomes cutanés 25

e) Méthodes alternatives de diagnostic 26

II) SANG ET LEUCÉMIE LYMPHOÏDE CHRONIQUE 28

1) Le sang *28*

a) Techniques de numération et d'identification des cellules sanguines en pratique quotidienne 34

b) Les maladies du sang ou hémopathies 34

c) Les leucémies aiguës 35

d) Les leucémies chroniques 35

2) La Leucémie Lymphoïde Chronique *36*

a) Diagnostic 37

b) Classification de Binet 38

c) Les principaux facteurs pronostiques 39

d) Traitements de la LLC 40

III) SPECTROSCOPIES VIBRATIONNELLES 42

1) Généralités *42*

2) Principe des spectroscopies vibrationnelles *43*

a) Spectroscopie d'absorption infrarouge 43

b) Spectroscopie de diffusion Raman 45

c) Règles de sélection 46

3) Instrumentation *48*

a) Le spectromètre infrarouge à transformée de Fourier 48

b) Le spectromètre Raman 50

4) Applications des spectroscopies vibrationnelles en oncologie *52*

IV) ANALYSE NUMÉRIQUE DES DONNÉES 53
1) Prétraitements *53*
a) Suppression des outliers 53
b) Suppression des spikes 54
c) Lissage 54
d) Correction de ligne de base 55
e) Normalisation 55
f) Correction de signaux contaminants 56
g) Extended Multiplicative Signal Correction 56
h) Sélection de variables 57
i) Extraction de variables 59
2) Classifications *61*
a) Généralités 61
b) Algorithmes de classification non supervisée 61
c) Algorithmes de classification supervisée 62
V) OBJECTIFS DE THÈSE 64
1) Aide au diagnostic de cancers cutanés par spectroscopie infrarouge *64*
2) Aide au diagnostic de la LLC par spectroscopie Raman *65*

CHAPITRE 2 : MATÉRIELS ET MÉTHODES 66

I) PRÉPARATION DES ÉCHANTILLONS 67
1) Coupes tissulaires *67*
2) Frottis sanguins *67*
II) INSTRUMENTATION 68
1) Microspectroscopie infrarouge à transformée de Fourier *68*
2) Microspectroscopie Raman *69*
III) PRÉTRAITEMENTS DES SPECTRES 72
1) Prétraitements des spectres infrarouge *72*
2) Prétraitements des spectres Raman *73*
3) Sélection de variables sur les spectres Raman *75*
IV) CLASSIFICATION DES SPECTRES 77
1) Classification non supervisée des spectres infrarouge : Fuzzy C-Means *77*
a) Principe 77
b) Avantages et limites 79
c) Indices de validité et m-rules 80
2) Classification des spectres Raman : Support Vector Machine *82*
a) Principe 82
b) Les SVMs linéaires 82
c) Les SVMs non linéaires 88
d) Avantages et limites 90

CHAPITRE 3 : RÉSULTATS ET DISCUSSIONS 92

PRÉAMBULE 93

ARTICLE #I 95

Présentation de la publication n°1 96

ARTICLE #II 104

Présentation de la publication n°2 105

ARTICLE #III 150

Présentation de la publication n°3 151

ARTICLE #IV 182

Présentation de la publication n°4 183

ARTICLE #V ET #VI 190

Présentation des publications n°5 et 6 191

ARTICLE #VII 217

Présentation de la publication n°7 218

CHAPITRE 4 : CONCLUSIONS ET PERSPECTIVES 224

I) AIDE AU DIAGNOSTIC DE CANCERS CUTANÉS PAR SPECTROSCOPIE INFRAROUGE 225

II) AIDE AU DIAGNOSTIC DE LA LLC PAR SPECTROSCOPIE RAMAN 226

BIBLIOGRAPHIE 229

PUBLICATIONS ET COMMUNICATIONS 245

ANNEXES 249

Liste des abréviations

ADN : Acide Désoxyribonucléique

ANN : Artificial Neural Network

ANR : Agence Nationale de la Recherche

ARC : Association pour la Recherche sur le Cancer

BCC : Basal Cell Carcinoma

BCR : B Cell Receptor

CaF_2 : Fluorure de Calcium

CD : Cluster Differenciation

CCD : Charged-Couple Device

CFS : Correlation-based Feature Selection

CP : Composante Principale

CO_2 : dioxyde de carbone

EMSC : Extended Multiplicative Signal Correction

FCBF : Fast Correlation-Based Feature Selection

FCM : Fuzzy C-Means

FCM-RBA : Fuzzy C-Means Redundant Based Algorithm

FCM-URBA : Fuzzy C-Means Upgraded Redundant Based Algorithm

FISH : Fluoresence In Situ Hybridization

FSI : Fuzzy Silhouette Width

FTIR : Fourier Transform Infrared

GIISI : Generalized Intra-Inter Silhouette Index

HAS : Haute Autorité de Santé

HCA : Hierarchical Clustering Analysis

H&E : Hématoxyline et Eosine

IHMO : Imagerie Hybride en Microscopie Oncologique

InVS : Institut de Veille Sanitaire

IPQ : Imagerie de Phase Quantitative

IR : Infrarouge

IRM : Imagerie par Résonance Magnétique

IRTF : Infrarouge à Transformée de Fourier

IV : Indice de Validité

KNN : K Nearest Neighbor

LDA : Linear Discriminant Analysis

LLC : Leucémie Lymphoïde Chronique

MASS : Ministère des Affaires Sociales et de la Santé

MCT : Mercury Cadmium Telluride

MGG : May-Grünwald Giemsa

MRDF : Maximum Representation and Discrimination Feature

NCLS : Nonnegatively Constrained Least Squares

NK : Natural Killer

OCT : Optical Coherence Tomography

OSI : Overlap and Separation Index

PC : Principal Component

PCA : Principal Component Analysis

PLS : Partial Least Square

PLS-DA : Partial Least Square – Discriminant Analysis

RBA : Redundant Based Algorithm

SBE : Sequential Backward Elimination

SCC : Squamous Cell Carcinoma

SCI : Separation and Compactness Index

SFD : Société Française de Dermatologie

SFS : Sequential Forward Selection

SJM : Schwämmle-Jensen Method

SMLR : Sparse Multinomial Logistic Regression

SNV : Standard Normal Variate

STCI : Separation and Total Compactness Index

SVM : Support Vector Machine

TCR : T Cell Receptor

TECSAN : Technologies pour la Santé et l'autonomie

TF : Transformée de Fourier

UV : Ultraviolet

WBC : Wisconsin Breast Cancer

ZAP-70 : Zeta-chain Associated Protein kinase 70

Liste des figures

FIGURE 1 : SCHÉMA REPRÉSENTATIF DES PRINCIPALES STRUCTURES DE LA PEAU 13
FIGURE 2 : REPRÉSENTATION DES QUATRE PRINCIPALES COUCHES DE L'ÉPIDERME ET SCHÉMA DU PROCESSUS DE KÉRATINISATION 14
FIGURE 3 : COUPE HISTOLOGIQUE D'UN CARCINOME CUTANÉ BASOCELLULAIRE COLORÉE À L'HÉMATOXYLINE ET ÉOSINE. 18
FIGURE 4 : COUPE HISTOLOGIQUE D'UN CARCINOME CUTANÉ SPINOCELLULAIRE COLORÉE À L'HÉMATOXYLINE ET ÉOSINE. 18
FIGURE 5 : COUPE HISTOLOGIQUE D'UN MÉLANOME COLORÉE À L'HÉMATOXYLINE ET ÉOSINE. 19
FIGURE 6 : SCHÉMA DU CALCUL DES INDICES DE CLARK ET DE BRESLOW 24
FIGURE 7 : SCHÉMA EXPLICATIF DE L'HÉMATOPOÏÈSE 29
FIGURE 8 : SCHÉMA DES DIFFÉRENTS LEUCOCYTES PRÉSENTS DANS LE SANG 30
FIGURE 9 : VUES AU MICROSCOPE DE FROTTIS SANGUINS COLORÉS 36
FIGURE 10 : SCORE DE MATUTES 37
FIGURE 11 : SCHÉMA DES 5 AIRES LYMPHOÏDES PALPABLES (GAUCHE) ET TABLEAU DE CLASSIFICATION DE BINET 38
FIGURE 12 : EXEMPLES DE MODES FONDAMENTAUX DE VIBRATION MOLÉCULAIRE 42
FIGURE 13 : DIAGRAMME DE JABLONSKI 46
FIGURE 14 : SCHÉMA INSTRUMENTAL D'UN SPECTROMÈTRE IRTF 49
FIGURE 15 : SCHÉMA INSTRUMENTAL D'UN SPECTROMÈTRE RAMAN 51
FIGURE 16 : IMAGEUR SPOTLIGHT IRTF 68
FIGURE 17 : CUBE DE DONNÉES SPECTRALES 69
FIGURE 18 : MODULE RAMAN 70
FIGURE 20 : HYPERPLAN ET POPULATIONS LINÉAIREMENT SÉPARABLES 83
FIGURE 21 : HYPERPLANS ET POPULATIONS NON LINÉAIREMENT SÉPARABLES 87
FIGURE 22 : POPULATIONS D'OBJETS IMBRIQUÉES DE FAÇON CONCENTRIQUE 87
FIGURE 23 : PROJECTION NON LINÉAIRE DE DONNÉES 88

Préambule

« Cancer », provenant d'un mot latin signifiant « crabe », est un terme général regroupant un ensemble de maladies très différentes les unes des autres, capables d'affecter n'importe quelle partie de l'organisme. L'une de ses caractéristiques principales est la prolifération incontrôlée de cellules dites « malignes », pouvant (selon les différents types de cancer) atteindre des parties de l'organisme adjacentes à celles dont ces cellules sont issues, voire envahir d'autres organes (métastases). Le cancer est une cause majeure de décès dans le monde. Cependant, un dépistage et un traitement précoces avant que la maladie n'atteigne un stade trop avancé, augmentent les chances de survie des patients. Le diagnostic est évoqué à partir de symptômes évocateurs sur la base d'un examen clinique ou lorsqu'un test de dépistage est positif. Il est ensuite confirmé ou infirmé par des explorations diverses telles un bilan biologique ou d'imagerie médicale. Finalement, le prélèvement et l'analyse au microscope des cellules et tissus concernés (examens histologiques et/ou cytologiques) permettront d'établir de façon définitive le diagnostic et d'en préciser le type. D'autres outils diagnostiques de confirmation (immunomarquage, biologie moléculaire, cytogénétique, etc.) basés sur des technologies modernes sont couramment utilisés par les professionnels impliqués au quotidien dans le diagnostic ou dans la recherche scientifique, dans le but de les améliorer ou d'en proposer de nouveaux plus performants.

C'est dans cette optique que les travaux réalisés au cours de cette thèse ont été menés. Ainsi, différents outils ont été développés pour apporter une aide au diagnostic de cancers, en couplant à la fois des technologies à fort potentiel discriminatoire et des outils de classification de données très performants. Ces études se sont articulées autour de deux axes principaux, tous deux basés sur l'utilisation de technologies de pointe que sont les microspectroscopies infrarouge et Raman, et par lesquelles ont pu être enregistrées des informations moléculaires sur les échantillons analysés. Ces données, riches en information et véritables signatures moléculaires, ont ensuite nécessité des prétraitements spécifiques à chaque étude, avant d'être considérées dans différents processus de classification.

La première étude concerne la discrimination de cancers cutanés *via* le développement d'un outil de classification autonome, basé sur un algorithme de classification non supervisée, et

appliqué à des données acquises par microspectroscopie infrarouge à l'échelle tissulaire. L'objectif est, non seulement de discriminer les différentes structures composant les biopsies étudiées, mais également d'apporter des informations complémentaires à celles obtenues de façon usuelle en anatomo-pathologie, comme la mise en avant de zones de progression tumorales, difficilement délimitables par les techniques conventionnelles.

La seconde étude, réalisée dans le cadre du projet ANR TecSan IHMO (Imagerie Hybride en Microscopie Oncologique), s'articule autour d'une approche diagnostique de la leucémie lymphoïde chronique, par l'enregistrement de spectres Raman à l'échelle cellulaire, et par l'usage de classifications supervisées. L'objectif est de montrer que les informations moléculaires (potentiellement complémentaires aux outils diagnostiques usuels) apportées par la spectroscopie Raman permettent une distinction entre cellule saine et tumorale.

Avant de présenter les résultats obtenus au cours de ces études, nous détaillerons dans un premier chapitre 1) des généralités quant à la peau, au sang ainsi qu'aux cancers leur étant associés et étudiés dans ces travaux de thèse (cancers de la peau et leucémie lymphoïde chronique respectivement), 2) les principes des spectroscopies vibrationnelles infrarouge et Raman et 3) une revue des méthodes d'analyses numériques classiquement appliquées en spectroscopie vibrationnelle.

Dans le chapitre 2, le matériel utilisé et la méthodologie suivie pour chaque étude sont présentés, telle la préparation des différents échantillons, les caractéristiques des spectromètres utilisés ainsi que les différents prétraitements et classifications appliqués aux spectres enregistrés.

Le chapitre 3 introduit et présente l'ensemble des résultats obtenus sous la forme de trois articles concernant le diagnostic de cancers cutanés, et trois autres articles pour l'approche diagnostique de la leucémie lymphoïde chronique. Dans ce chapitre, une publication supplémentaire est également présentée, comportant les résultats préliminaires d'une étude réalisée *via* le module d'imagerie hybride développé dans le cadre du projet IHMO et visant à diagnostiquer la malaria.

Le chapitre 4 expose les principales conclusions de ces travaux ainsi que leurs perspectives.

Chapitre 1 : Introduction et Généralités

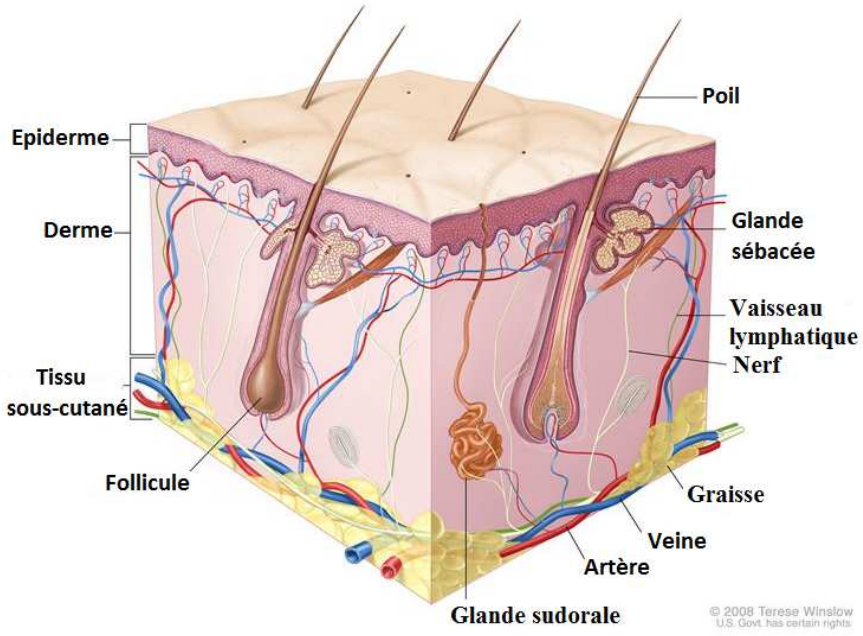
Le but de cette thèse étant de proposer de nouvelles approches d'aide au diagnostic des cancers cutanés et du sang, cette partie a pour objectif de décrire succinctement ces cancers, et également de présenter les technologies et analyses numériques employées dans ce but.

I) Peau et cancers cutanés

1) La peau (1-7)

La peau est l'organe le plus important du corps humain, que ce soit par sa taille ou son poids. Chez un homme adulte moyen, elle mesure environ 2 m^2 et représente 15% du poids du corps. Ses fonctions principales sont de protéger le corps contre les agressions mécaniques ou infectieuses, les radiations lumineuses et réguler la température et les fluides corporels.

L'épaisseur de la peau n'est pas homogène. Elle varie de 1 mm (paupières) à 4 mm (paume des mains et plante des pieds), avec une moyenne de 2 mm pour l'ensemble du corps. La peau, de composition identique quelle que soit la zone concernée, se structure en plusieurs couches, dont les principales sont l'épiderme, le derme et la couche sous-cutanée ou hypoderme (Figure 1).



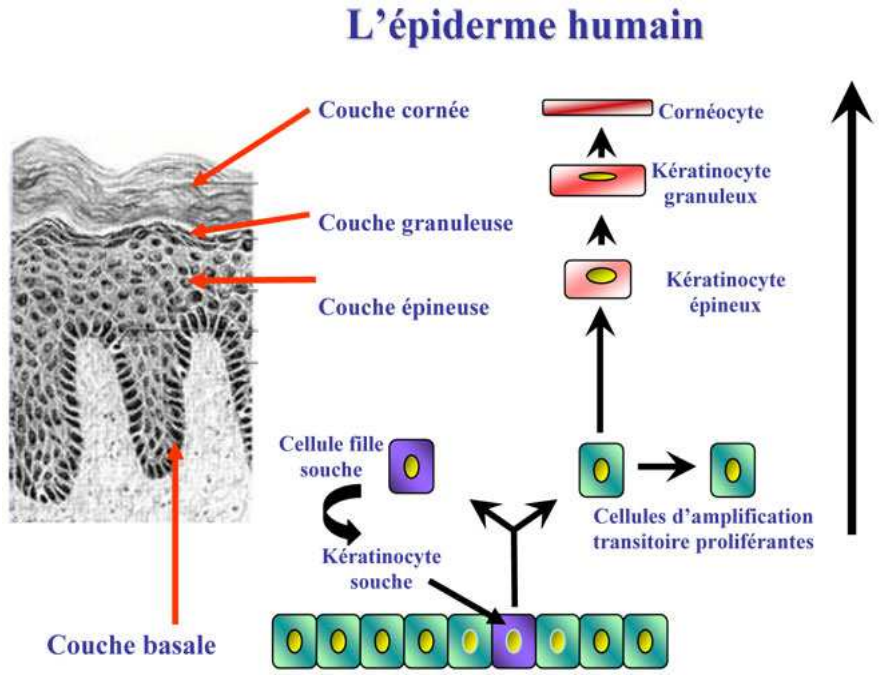
(Terese Winslow for the National Cancer Institute 2008, U.S. Govt.)

Figure 1 : Schéma représentatif des principales structures de la peau

La peau est constituée de différentes couches dont les principales sont l'épiderme, le derme et la couche sous cutanée. Ces couches se composent de différents éléments tels les glandes sébacées, les glandes sudorales, les follicules pileux, les vaisseaux sanguins ou lymphatiques.

a) L'épiderme

L'épiderme est un épithélium constituant la couche la plus externe de la peau (entre 0,1 et 0,2 mm d'épaisseur). Il est subdivisé en plusieurs sous-couches qui sont de la plus profonde à la plus superficielle, **la couche basale** (*stratum germinativum*), composée de kératinocytes assurant le renouvellement de l'épiderme, **la couche épineuse** (*stratum spinosum*), apportant des propriétés d'imperméabilité de la peau, **la couche granuleuse** (*stratum granulosum*), permettant la formation des futures cellules de la couche cornée, et **la couche cornée** (*stratum corneum*), véritable barrière de protection aux facteurs externes (Figure 2).



(www.inserm.fr)

Figure 2 : Représentation des quatre principales couches de l'épiderme et schéma du processus de kératinisation
L'épiderme est composé à 90% de kératinocytes migrant depuis la couche basale vers la couche cornée en passant par différents stades d'évolution. Les kératinocytes souches de la couche basale ont pour propriétés de pouvoir à la fois se régénérer (cellule fille souche) et se différencier (cellules d'amplification transitoire proliférantes). Au cours de ce processus migratoire, les kératinocytes accumulent de la kératine, entrent en apoptose (mort cellulaire programmée) et finissent par perdre leur noyau pour former les cornéocytes de la couche cornée.

Au sein de ces couches, on retrouve quatre types de cellules :

- **Les kératinocytes :**

  Ils représentent près de 85% des cellules de l'épiderme et sont présents à différentes étapes de maturation, selon une progression de 28 jours en moyenne de la couche basale jusqu'à la couche cornée. La Figure 2 présente ce processus de maturation appelé kératinisation. Chaque kératinocyte souche de la couche basale donne naissance à deux cellules filles similaires, l'une deviendra un kératinocyte souche identique à la cellule mère, la seconde entrera dans le processus de maturation durant lequel elle accumulera de la kératine. Au niveau de la couche granuleuse, ces cellules s'aplatissent, perdent leur noyau et entrent en apoptose (mort cellulaire programmée) pour former les cornéocytes de la couche cornée. La kératine présente en quantité dans leur cytoplasme couplée à un film lipidique présent à la surface de la peau donne à la couche cornée ses propriétés hydrophobes.

- **Les mélanocytes :**

  Ils représentent autour de 2,5% de la population totale épidermique. Les mélanocytes se situent dans la couche basale de l'épiderme et sont responsables de la pigmentation de la peau. Lorsqu'ils sont stimulés par les rayons ultra-violets, les mélanocytes sécrètent une protéine, la mélanine, dont la fonction est de protéger les kératinocytes souches de ces rayons. Cependant, chez les populations à peau noire, on peut retrouver ces cellules dans toutes les couches de l'épiderme et la production de mélanine y est spontanée. Les grains de mélanine sont produits dans le cytoplasme des mélanocytes et circulent dans leurs bras, pour être déposés dans les kératinocytes de la couche épineuse avec lesquels ils sont en contact. Plus les grains de mélanine sont présents dans les kératinocytes, plus la peau apparaîtra foncée.

- **Les cellules de Langerhans :**

  Elles constituent 3 à 8% des cellules épidermiques, et sont produites dans la moelle osseuse avant de migrer vers la couche épineuse de l'épiderme, où elles constituent la première ligne de défense immunitaire. Les cellules de Langerhans sont capables de digérer des corps étrangers (virus, bactéries ou allergènes) et présentent les produits de digestion aux lymphocytes, cellules du système immunitaire.

- **Les cellules de Merkel :**

  Situées au niveau de la couche basale, ces cellules sont des mécanorécepteurs qui jouent un rôle au niveau des sensations tactiles. Elles sont les moins nombreuses de l'épiderme, entre 0,5 et 5%, mais sont très présentes dans certaines régions de la peau telles la pulpe des doigts, la paume des mains ou encore la plante des pieds.

b) Le derme

C'est un tissu conjonctif fibreux et élastique d'une épaisseur variant de 1 à 4 mm en fonction de la zone du corps concernée. Le derme est situé sous l'épiderme, et lui est relié *via* la couche basale. Il est constitué à 80% d'eau, et présente des fibres d'élastine et de collagène en grand nombre lui assurant solidité, souplesse et élasticité. Peu de cellules sont présentes dans le derme. On y trouve essentiellement des fibroblastes (productrices de collagène et d'élastine), des lymphocytes et des macrophages, mais également des éléments de la peau dits « annexes » tels que les poils, les glandes sudoripares (productrices de sueur), les glandes sébacées (productrices de sébum) et les vaisseaux sanguins présents en nombre (apportant nutriments et cellules immunitaires). C'est également dans le derme que l'on retrouve les éléments nerveux dont les terminaisons libres progressent jusque dans l'épiderme et qui permettent de ressentir les sensations de toucher, chaleur et pression. Le derme peut être divisé en deux sous structures, le derme papillaire et le derme réticulaire :

- **Le derme papillaire :**

  C'est la couche supérieure du derme en contact avec l'épiderme. Elle mesure autour de 0,1 mm d'épaisseur et est principalement composée d'élastine et de collagène I. C'est dans cette couche que l'on trouve les récepteurs sensoriels (à l'exception des récepteurs de pression), des macrophages et des leucocytes. À sa surface, le derme papillaire apparaît irrégulier, ceci est dû à de nombreuses excroissances lui donnant un relief accidenté.

- **Le derme réticulaire :**

  C'est la couche la plus vaste du derme présentant beaucoup moins de fibroblastes et de cellules du système immunitaire mais possédant un réseau de fibres de collagène et d'élastine plus fourni. C'est ce réseau fibreux qui confère à la peau résistance, extensibilité, élasticité et au sein duquel on retrouve les glandes sébacées et sudoripares ainsi que les nerfs et follicules pileux.

c) L'hypoderme

L'hypoderme ou couche sous-cutanée, est la structure la plus profonde de la peau, mesurant entre 1 mm et 1 cm d'épaisseur. C'est un tissu souple, déformable, riche en fibroblastes et adipocytes (cellules stockant les graisses), qui sert de transition entre la peau et les organes qu'il recouvre. Les graisses présentes dans les adipocytes jouent un rôle de réserve énergétique sollicitée par exemple lors d'un effort intense, mais également un rôle d'isolant thermique. La fonction de protection de la peau peut être mise à mal par les nombreuses agressions répétées qu'elle subit au contact du monde extérieur. Certains facteurs de risque connus, comme par exemple l'exposition solaire, les cicatrices de brûlures, les pesticides, les hydrocarbures ou encore le tabac, peuvent entrainer l'apparition de cancers cutanés.

2) Les cancers cutanés (8-10)

a) Description

Ce sont les plus fréquents des cancers chez l'homme avec près de 80 000 cas nouvellement diagnostiqués chaque année en France. Les cancers cutanés prennent naissance lors d'une multiplication non contrôlée de cellules anormales au sein d'une couche de la peau. Les formes de cancers cutanés les plus courantes se développent à partir de l'épiderme et sont différentiables selon leur origine cellulaire :

- **Les carcinomes cutanés basocellulaires** (Basal Cell Carcinoma, BCC, Figure 3) sont les plus courants des carcinomes cutanés avec une incidence de 40 à 50 000 nouveaux cas par an en France. Ils se développent à partir de kératinocytes dans la couche la plus profonde de l'épiderme, la couche basale. Leur évolution reste lente et leur potentiel métastatique quasi nul (0,002% des cas). Les BCC non pris en charge pendant une longue période peuvent, dans de très rares cas, se développer en profondeur et atteindre des tissus sous-jacents, muscles, os et organes.

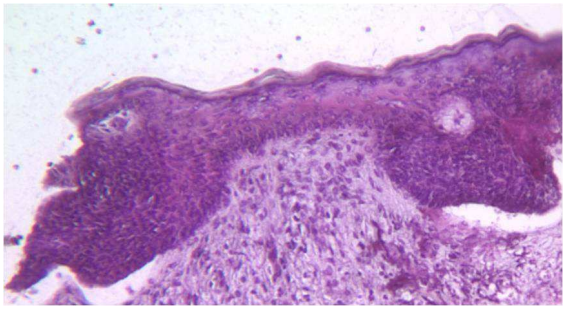


Figure 3 : Coupe histologique d'un carcinome cutané basocellulaire (Basal Cell Carcinoma, BCC) colorée à l'hématoxyline et éosine.

Les carcinomes cutanés basocellulaires sont non invasifs, se développent dans la lame basale, ont une évolution lente et ne métastasent pas.

- **Les carcinomes cutanés spinocellulaires** (Squamous Cell Carcinoma, SCC, Figure 4) sont moins fréquents que les BCC, avec une incidence de 10 à 12 500 nouveaux cas détectés en France chaque année, mais sont plus agressifs. Une lésion préexistante n'est pas nécessaire à leur apparition mais des kératoses actiniques (lésions précancéreuses dues à une forte exposition solaire) sont un facteur de risque à l'apparition de SCC. Leur développement se situe sur des zones de la peau exposées au soleil et plus précisément dans la couche épineuse de l'épiderme à partir de kératinocytes. Ils peuvent exister sous forme intra-épidermique, maladie de Bowen (11), et sont capables de métastaser, pouvant ainsi engager le pronostic vital des patients s'ils ne sont pas traités suffisamment tôt (5).

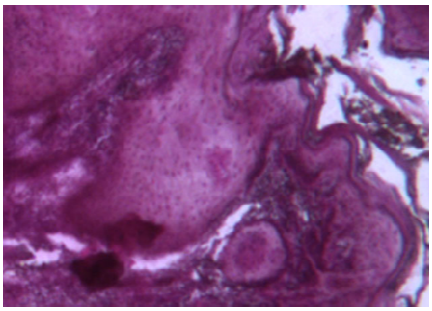


Figure 4 : Coupe histologique d'un carcinome cutané spinocellulaire (Squamous Cell Carcinoma, SCC) colorée à l'hématoxyline et éosine.

Les SCC peuvent se développer à partir de lésions cutanées dues à une forte exposition solaire. Ils se développent dans la couche épineuse de l'épiderme et ont un fort potentiel infiltrant, leur permettant de métastaser.

- **Les mélanomes** (Figure 5) sont issus des mélanocytes et sont des cancers métastatiques agressifs ne représentant que 5% de la totalité des cancers cutanés mais possédant le plus fort taux d'augmentation en France. Leur incidence a plus que triplé de 1980 à 2005 et plus de 9700 nouveaux cas ont été répertoriés en 2011 en France (9). Cette augmentation est essentiellement due à une surexposition aux rayons UV naturels ou artificiels, malgré les nombreuses campagnes de sensibilisation à l'exposition solaire menées depuis plusieurs années.

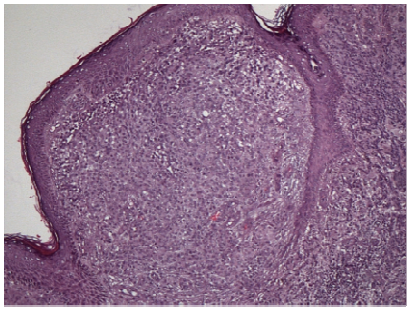


Figure 5 : Coupe histologique d'un mélanome colorée à l'hématoxyline et éosine.
Les mélanomes sont des cancers agressifs ayant un fort pouvoir métastatique. Ils peuvent apparaître à partir d'une peau saine ou résulter de la transformation d'un grain de beauté. Issus des mélanocytes, leur apparition relève généralement d'une surexposition solaire.

b) Diagnostic clinique

Le diagnostic clinique des cancers cutanés repose sur l'évaluation morphologique de la lésion par un dermatologue. Chaque catégorie présente des caractéristiques propres.
Les rapports de la Haute Autorité de Santé (HAS) et de la Société Française de Dermatologie (SFD) distinguent 3 sous-catégories de carcinomes cutanés basocellulaires (8, 12) :

- **Le BCC nodulaire :** nodule lisse, translucide, grisâtre et d'aspect perlé, formant une lésion élémentaire. Le BCC nodulaire croît progressivement et atteint une taille variable.
- **Le BCC superficiel :** plaque rouge bien délimitée, parfois recouverte de croûtes. Les BCC superficiels sont préférentiellement situés au niveau des parties couvertes de la peau.
- **Le BCC sclérodermiforme :** plaque dure et brillante difficile à délimiter et ressemblant à une cicatrice blanche. Sa progression peut être étendue de par la difficulté à le diagnostiquer en l'absence d'ulcération.

Le carcinome cutané spinocellulaire se caractérise par une lésion croûteuse, jaunâtre, durcie et présentant une ulcération centrale. Il peut également se caractériser par une lésion végétante ou bourgeonnante.

Selon un rapport coordonné par Martin et Bonerandi (11) et validé par l'Institut National du Cancer et la Haute Autorité de Santé, il existe beaucoup de sous-types de carcinomes cutanés spinocellulaires, variant de par leur forme et leur évolution.

Cependant, beaucoup de cas sont rares et leur morphologie n'est pas clairement définie dans la littérature.

Quelques exemples de sous-types de SCC :

- Une forme commune du SCC apparaissant sur une cicatrice, une ulcération chronique, une radiolésion ou *de novo* (sans lésion préexistante). Elle se caractérise par une tumeur saillante, infiltrante, de surface irrégulière et légèrement saignante.

- Les carcinomes cutanés spinocellulaires verruqueux, localisés sur les jambes, les pieds, la muqueuse buccale et pharyngée, la région génitale ou péri-anale. Ils ont une évolution lentement progressive, cliniquement végétante, mais d'extension infiltrante lente.

- Les formes à cellules fusiformes (sarcomatoïde) sont rares, en continuité avec l'épiderme ou présentent un déficit dans la synthèse de kératine.

- Les carcinomes cutanés spinocellulaire acantholythiques se localisent au niveau de la tête et du cou. Des cavités se forment au niveau des lobules tumoraux et peuvent contenir du matériel amorphe sans présence de sécrétion mucineuse.

Concernant **le mélanome** (13), les signes d'alerte les plus fréquents sont l'apparition d'une lésion (pigmentée ou non) ou la modification d'un grain de beauté (naevus) jusqu'alors stable. Le diagnostic est posé après une première analyse selon la règle ABCDE (14) :

- Lésion de forme **A**symétrique
- Ayant des **B**ords irréguliers, encochés, polycycliques
- De **C**ouleur hétérogène
- D'un **D**iamètre supérieur à 6 mm
- Dont l'**E**volution présente des changements (aspect, taille, couleur et épaisseur)

Ce diagnostic du mélanome est toujours confirmé par l'examen anatomopathologique d'une exérèse complète et non d'une biopsie qui peut être source d'erreurs.

c) Diagnostic histologique

Lorsqu'un spécialiste soupçonne la présence d'un cancer cutané, un prélèvement (biopsie) ou une exérèse complète peuvent être réalisés. Les échantillons ainsi obtenus sont soit congelés soit inclus en paraffine. La congélation permet de mettre le tissu dans un état de conservation rapidement après extraction. L'inclusion en paraffine nécessite quant à elle plus de temps mais présente l'avantage de pouvoir conserver l'échantillon plus longtemps sans qu'il ne se dégrade, certains échantillons sont d'ailleurs conservés plusieurs années.

De fines coupes (autour de 5 µm d'épaisseur) sont ensuite réalisées *via* un cryotome (pour les échantillons congelés) ou un microtome (pour les échantillons inclus en paraffine). Déposées sur lames de verre, les coupes sont ensuite colorées, classiquement à l'hématoxyline et éosine (H&E), avant d'être observées au microscope par un spécialiste en anatomopathologie.

Il est à noter que les coupes paraffinées subissent un déparaffinage chimique avant coloration.

Ce diagnostic histologique est donc basé sur des critères morphologiques, et fait office de méthode de référence pour confirmer ou infirmer les hypothèses faites en amont.

Le carcinome cutané basocellulaire (12) :

La Société Française de Dermatologie distingue 6 sous-types histologiques de carcinomes cutanés basocellulaires :

- **Le BCC nodulaire :** présente dans le derme plusieurs massifs bien délimités et constitués de cellules basaloïdes ayant des noyaux agencés en palissade en périphérie.

- **Le BCC superficiel :** présente des foyers tumoraux reliés à l'épiderme ou à des follicules pileux. Comme pour le BCC nodulaire, ils sont composés de cellules basaloïdes ayant des noyaux agencés en palissade en périphérie. En général, les cellules tumorales sont séparées du stroma (tissu induit par les cellules cancéreuses) par des artefacts de rétraction.

- **Le BCC infiltrant :** peut être de forme trabéculaire (cellules agencées en travées), avec de petits foyers tumoraux qui s'infiltrent dans le derme avec des délimitations floues. Il peut également être de forme micronodulaire, présentant une multitude de petits foyers tumoraux bien délimités.

- **Le BCC sclérodermiforme :** présente des foyers tumoraux en forme de cordons effilés où les cellules tumorales sont peu différenciées. En général, ces tumeurs s'étendent sur toute la hauteur du derme et peuvent parfois atteindre l'hypoderme.

- **Le BCC métatypique :** en comparaison des autres BCC, il présente une différenciation malpighienne (relative aux couches épithéliales profondes de l'épiderme) carcinomateuse. Notons toutefois que cette sous-catégorie de BCC reste controversée.

- **Le carcinome mixte ou composite :** est une association entre un carcinome cutané basocellulaire et spinocellulaire. Cependant chacun reste identifiable.

Le carcinome cutané spinocellulaire (11) **:**

Le carcinome cutané spinocellulaire est caractérisé par une prolifération de grandes cellules organisées en lobules (petits lobes) ou en travées, généralement mal délimités et où des globes de kératine sont fréquemment présents.

Cette tumeur apparaît dans l'épiderme, peut franchir la lame basale et s'infiltrer plus ou moins profondément dans le derme (si elle ne la franchit pas on parlera alors de maladie de Bowen). Dans certains cas elle peut atteindre l'hypoderme.

Les sous-types de SCC présentent également des caractéristiques propres d'un point de vue histologique comme par exemple :

- Les carcinomes cutanés spinocellulaires de forme commune présentant des bourgeons irréguliers composés de kératinocytes atypiques qui infiltrent le derme.

- La prolifération des formes verruqueuses reste longtemps bien différenciée avant que le carcinome ne s'infiltre plus en profondeur sans pour autant être fortement métastatique.

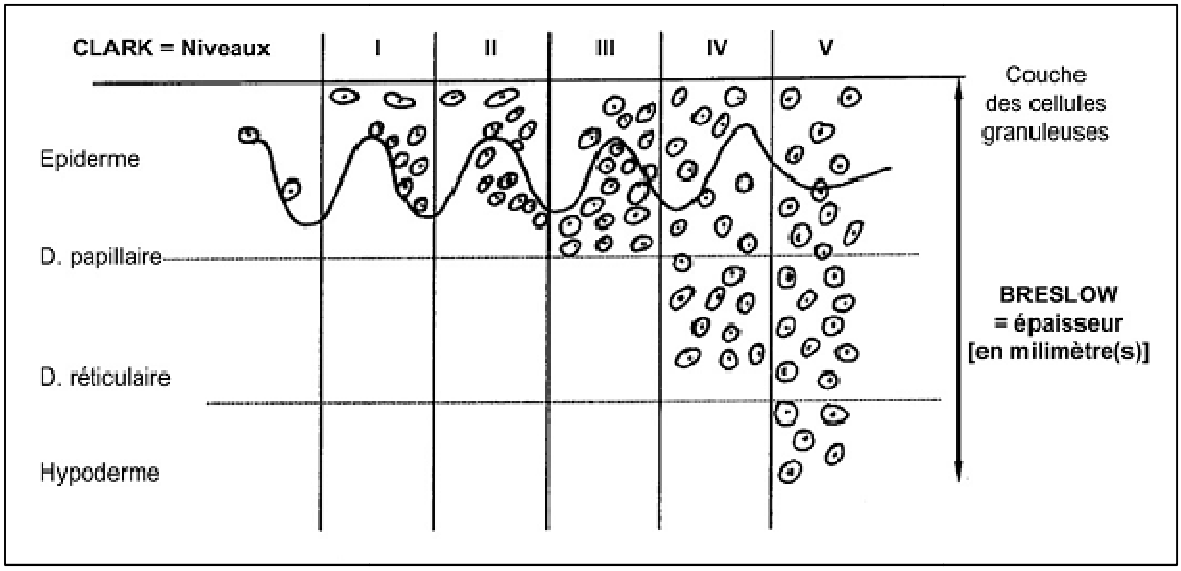
- En l'absence de signes histologiques, les carcinomes à cellules fusiformes sont caractérisés par immunohistochimie, en montrant qu'ils expriment de la cytokératine (constituants de kératine).

- Les formes acantholytiques présentent l'apparition de cellules kératinisées de petite taille et de forme anormale.

Le mélanome (14, 15) :

Le mélanome est caractérisé par une évolution en deux phases de croissance. La première est « horizontale » tout d'abord au dessus de la membrane basale (phase intra-épidermique), puis dans le derme (phase micro-invasive). La seconde phase est « verticale », le mélanome progressant plus profondément dans le derme devient de ce fait invasif.

Deux calculs d'indices déterminent l'évolution des mélanomes (Figure 6). **L'indice de Clark** caractérise le niveau d'infiltration (en mm) et **l'indice de Breslow** son épaisseur (en mm) depuis la surface de la couche granuleuse de l'épiderme jusqu'à la cellule tumorale la plus profonde.



(Lavie *et al*, 2007)

Figure 6 : Schéma du calcul des indices de Clark et de Breslow

Les valeurs de ces deux indices renseignent sur l'invasion (Clark) et l'épaisseur du mélanome (Breslow), deux informations à valeurs diagnostique et pronostique.

Il existe 4 sous-catégories histologiques de mélanomes (10) :

- **Le mélanome superficiel extensif** (aux alentours de 70% des cas) : se caractérise par une prolifération de mélanocytes isolés ou groupés en thèques (enveloppes). Ces mélanocytes sont présents sur toute la profondeur de l'épiderme. Lors de sa croissance verticale il pénètre le derme puis l'hypoderme.

- **Le mélanome nodulaire** (entre 5 et 20% des cas) : sa progression n'est que verticale, il est donc très vite invasif (quelques semaines à quelques mois). Généralement, il ne devient identifiable que lorsqu'il devient palpable (indice de Breslow élevé).

- **Le mélanome acral lentigineux** (entre 2 et 10% des cas) : est localisé sur la paume des mains, la plante des pieds et sous les ongles. Sa progression varie de quelques mois à plusieurs années.

- **Le mélanome de Dubreuilh** (10% des cas) : présente des mélanocytes atypiques sur les couches inférieures de l'épiderme et est généralement localisé au niveau des zones exposées au soleil comme le visage. Sa progression horizontale est lente (plusieurs années) avant qu'il ne progresse verticalement et devienne invasif.

d) Traitement des carcinomes cutanés

Les carcinomes cutanés basocellulaires, spinocellulaires et les mélanomes (16) sont traités dans la quasi-totalité des cas par chirurgie (selon les recommandations de la HAS) bien qu'il existe des alternatives à celle-ci (6). Les exérèses réalisées peuvent être d'importance variable, allant de l'exérèse suivie d'une suturation à l'exérèse en deux temps nécessitant une reconstruction plastique au vu de la localisation ou de la masse de tumeur à retirer. L'avantage principal de cette intervention chirurgicale est que la pièce retirée peut être analysée pour confirmer le diagnostic clinique mais également vérifier le retrait total de la tumeur.

Dans le cas des mélanomes primitifs (formes non métastasées), l'exérèse complète est le seul traitement curatif. En ce qui concerne les mélanomes de stade avancé, ils ne sont pas curables (6, 17). Pour le cas des BCC et des SCC, une étude basée sur 844 patients ayant subi une exérèse complète des tumeurs, a montré que 95% d'entre eux n'a pas récidivé après 3 ans (18).

Les exérèses sont donc réalisées (dans la quasi-totalité des cas) dans le but de supprimer les tumeurs dans leur totalité. Mais celles-ci sont plus ou moins étendues en surface ou en profondeur et sont plus ou moins invasives. Se pose alors la question de distance à la tumeur afin d'empêcher une récidive. On introduit ici la notion de marge d'exérèse (18).

Ces marges, fixées selon des recommandations standards « Standards, Options et Recommandations » (11, 19, 20), doivent être déterminées en fonction du type et de l'épaisseur de la tumeur mais ne doivent jamais dépasser 3 cm :

- Elles varient de 3 à 4 mm pour les BCC selon que leur pronostic soit bon, intermédiaire ou mauvais.
- Les SCC pouvant potentiellement métastaser, leur marge est supérieure à celles des BCC et relève du centimètre.
- Quant aux mélanomes, c'est leur épaisseur, donnée par l'indice de Breslow, qui détermine la marge à appliquer. Celle-ci varie de 0,5 mm à 3 cm.

Le mélanome de Dubreuilh étant non invasif pendant plusieurs années, une marge fixée à 1 cm est recommandée si elle est réalisable. Dans le cas contraire, une marge de 0,5 cm est tolérée avec un suivi histologique strict des berges.

e) Méthodes alternatives de diagnostic (21-23)

Beaucoup de méthodes non invasives et appliquées à l'étude de la peau ou au diagnostic de cancers cutanés ont été étudiées. Parmi ces différentes techniques, la plus communément utilisée est la dermoscopie (ou dermatoscopie).

Cette technique *in vivo* et non invasive apporte des informations morphologiques visibles et non visibles à l'œil nu. Elle permet notamment une observation détaillée de la distribution et architecture de la mélanine dans l'épiderme, dans la jonction dermo-épidermique ainsi que dans le derme. À ce titre, la dermoscopie présente surtout un intérêt pour l'examen des lésions pigmentaires (24).

On notera également que depuis l'apparition des téléphones nouvelle génération, des industriels se sont penchés sur le développement d'accessoires et applications pour smartphone permettant de transformer celui-ci en dermoscope numérique (25).

D'autres méthodes, basées sur la biophysique, ont également montré un potentiel discriminatoire dans la différenciation de tissus cancéreux (22) :

- **l'Imagerie par Résonance Magnétique** (IRM) (26) qui permet une distinction des tissus, et détermine le degré d'hydratation de ceux-ci mais à une résolution faible d'un point de vu dermatologique (de l'ordre de plusieurs centaines de microns).

- **la tomographie par cohérence optique** (Optical Coherence Tomography, OCT) (27, 28) basée sur la détection par interférométrie de photons rétrodiffusés cohérents, cette technique permet de localiser les interfaces tissulaires présentant des indices de réfraction différents.

- **la spectroscopie de fluorescence** (29) permet de mesurer l'autofluorescence de l'échantillon. Cette autofluorescence est due au renvoi de photons par un tissu ou molécules après excitation *via* une source lumineuse. Cette technique permet de renseigner sur la prolifération cellulaire ou encore sur la dégradation du collagène.

- **la microscopie confocale à balayage laser** (28, 30) est une technique de très haute résolution (de l'ordre du micron) basée sur la rétrodiffusion de la lumière par l'échantillon observé. Cette technique peut être appliquée *in vivo* et permet de différencier les cellules de l'épiderme en fonction de leurs couches respectives puisque seules les informations du point de focalisation sont enregistrées.

De nombreuses autres méthodes biophysiques existent telles que **l'imagerie polarisée** (23), **l'échographie** (31, 32), **la microscopie à cohérence optique** (Optical Coherence Microscopy, OCM) (33), **la profilométrie** (21), **la spectroscopie par réflectance diffuse** (34), **la spectroscopie térahertz** (35, 36), **la microscopie de fluorescence à deux photons** (37), ou encore **la spectroscopie de génération de seconde harmonique** (37).

En clinique seule la **dermoscopie** est utilisée, bien qu'uniquement basée sur des critères morphologiques. Aucune information à l'échelle moléculaire de la tumeur ne peut être obtenue *via* cette technique.

II) Sang et leucémie lymphoïde chronique

1) Le sang

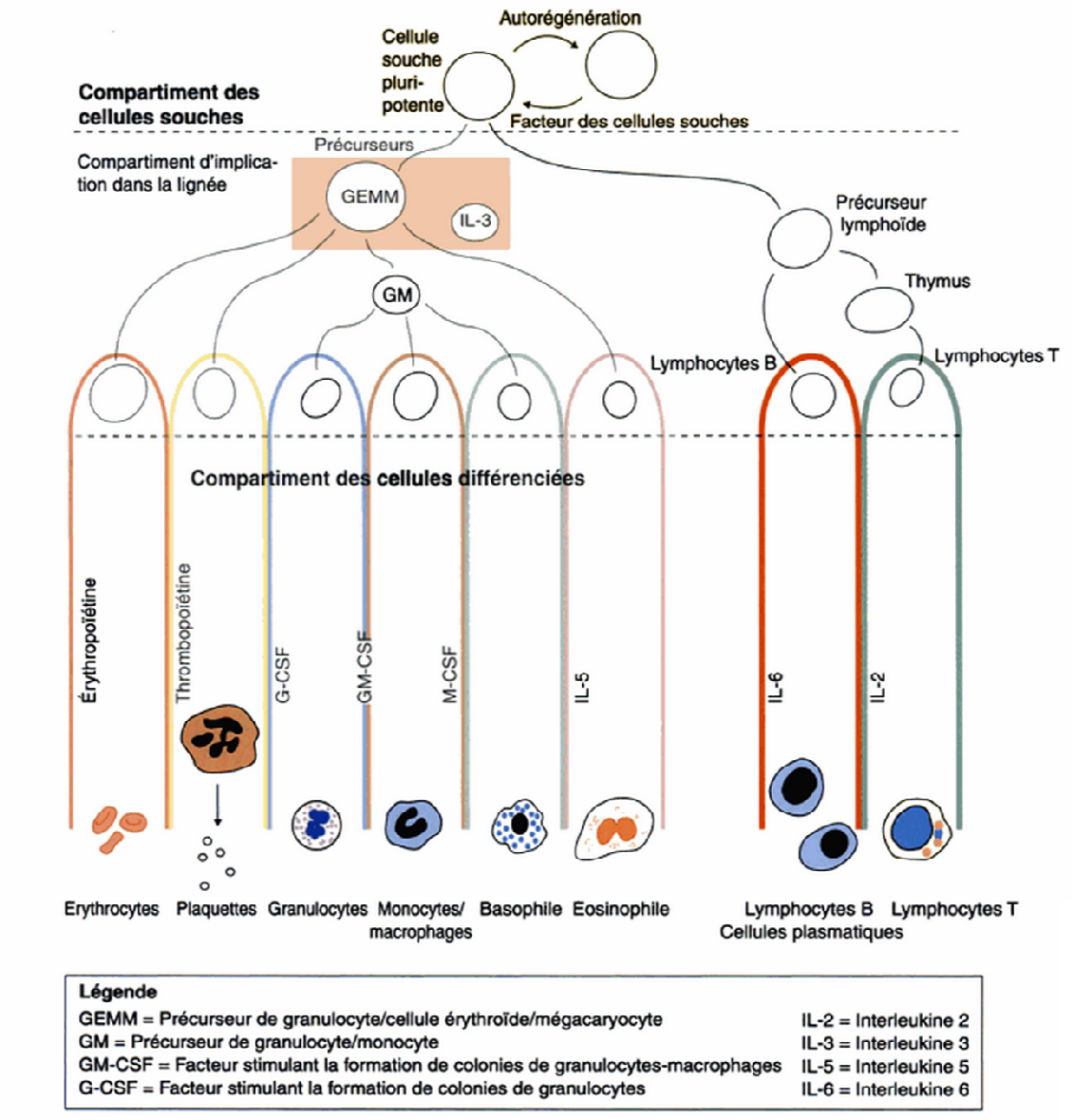
Le sang est un liquide présent à hauteur de 5 litres chez un adulte moyen mais dont le volume varie en fonction du sexe, de la taille et du poids de chaque personne. Il irrigue tous les tissus de l'organisme, y apportant les éléments nutritifs et l'oxygène, et y recueillant les déchets, *via* les artères, les veines et les capillaires. Il contient des cellules (45%) en suspension dans un liquide, le plasma (55%). La production de ces cellules est réalisée à partir de la cellule souche hématopoïétique, lors de **l'hématopoïèse**, ensemble des mécanismes incluant notamment l'auto-renouvellement des cellules souches de l'hématopoïèse et la différenciation en lignées, vers l'une ou l'autre des cellules du sang (Figure 7). L'hématopoïèse se situe à l'âge adulte dans la moelle osseuse.

Parmi les cellules sanguines on distingue trois catégories (38, 39) :

- **Les plaquettes :** aussi appelées thrombocytes, sont des cellules anucléées, formées lors de la thrombopoïèse par fragmentation du cytoplasme des mégacaryocytes contenus dans la moelle osseuse. Leur taille est de 2 à 5 µm et leur fonction biologique est hémostatique. Elles forment le clou plaquettaire qui obture les brèches se produisant sur les parois vasculaires (hémostase primaire) et de ce fait, interviennent en prévention ou pour stopper les hémorragies. Elles possèdent à leur surface des protéines leur assurant une adhésion aux parois des vaisseaux sanguins lésés. En cas de brèche vasculaire, les plaquettes adhèrent à l'endothélium lésé (couche la plus interne des vaisseaux sanguins) et sécrètent des facteurs permettant leur agrégation en un thrombus. Leur nombre varie de 150 000 à 450 000/mm^3 de sang chez un individu sain et leur durée de vie est d'environ 7 jours.

- **Les globules rouges :** ou hématies ou érythrocytes, sont les cellules les plus nombreuses du sang (environ 5 millions/mm^3) dont la taille varie autour de 7 µm. Ils sont issus d'érythroblastes de la moelle osseuse, et de la maturation finale du réticulocyte, cellule précédent le stade d'érythrocyte lors de l'érythropoïèse. Les hématies ont une durée de vie moyenne de 120 jours. Elles sont anucléées et ont un

aspect de disque biconcave, morphologie adaptée à la déformabilité. L'hémoglobine intra-érythrocytaire permet le transport de l'oxygène depuis les poumons vers les cellules et tissus de l'organisme. C'est également à la surface de leur membrane que l'on retrouve les antigènes permettant de déterminer les différents groupes sanguins.



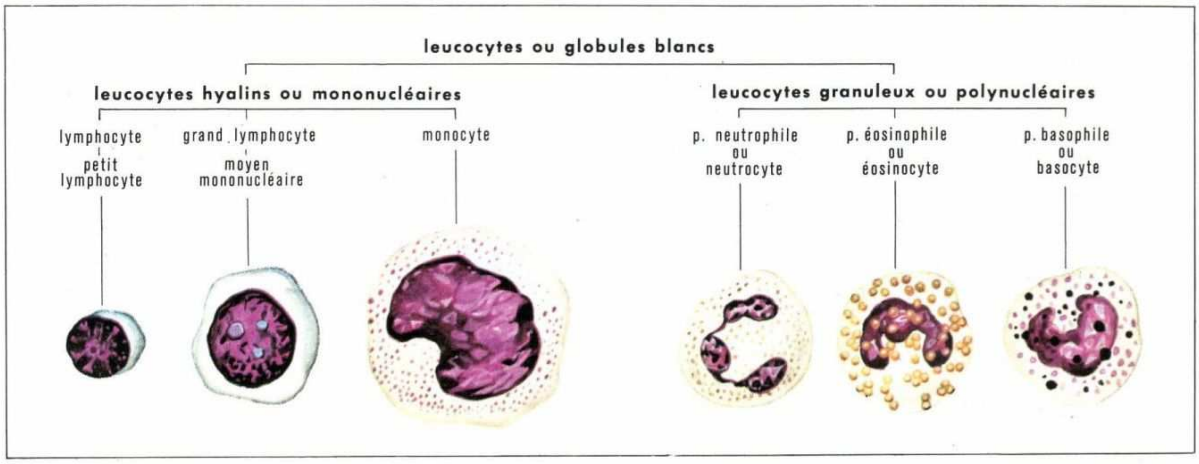
(Mehta et Hoffband, 2003)

Figure 7 : Schéma explicatif de l'hématopoïèse

L'hématopoïèse est un processus de multiplication cellulaire au cours duquel des cellules souches, présentes dans la moelle osseuse, s'autorégénèrent et se différencient en plusieurs lignées cellulaires. À la fin d'un processus de maturation, les cellules matures passent de la moelle vers le sang circulant et constituent les différentes cellules sanguines.

- **Les globules blancs** (Figure 8) : ou leucocytes, sont des cellules possédant un noyau et un cytoplasme. Elles sont présentes dans le sang à hauteur de 4 000 à 10 000/mm^3 et agissent dans le cadre de la défense immunitaire de l'organisme. On peut en distinguer deux catégories selon que leur noyau soit monolobé (lymphocytes et monocytes) ou plurilobé (polynucléaires).

  Les globules blancs sont reconnaissables sur frottis sanguin en lecture microscopique, grâce à la coloration de May Grünwald Giemsa (MGG) qui associe deux colorants : le May-Grünwald, contenant un colorant acide, l'éosine, et un colorant basique, le bleu de méthylène. Le Giemsa contient de l'éosine et un colorant basique, l'azur de méthylène. L'association de ces deux colorants permet de différencier les éléments figurés du sang grâce à leurs propriétés distinctes. Le May-Grünwald colore les éléments acidophiles et les granulations neutrophiles des leucocytes. Le Giemsa colore le cytoplasme des monocytes, des lymphocytes et la chromatine des noyaux.



(Larousse médical Ed. 1971-1976)

Figure 8 : Schéma des différents leucocytes présents dans le sang

Les leucocytes peuvent être repartis en deux catégories, ceux présentant un noyau monolobé, les lymphocytes et monocytes, et ceux présentant un noyau plurilobé, les polynucléaires. Ces derniers sont également nommés granulocytes. Sur ce schéma on peut retrouver les 3 sous-catégories de polynucléaires, les neutrophiles, les éosinophiles et les basophiles.

Les polynucléaires constituent environ 70% des leucocytes, ont une taille d'environ 13 µm et peuvent être de trois types, neutrophiles, basophiles et éosinophiles :

- 99% sont des **polynucléaires neutrophiles** (soit 2 à 7,5 G/L) dont la durée de vie est de 24h. Ils sont caractérisés par un noyau à 3 ou 4 lobes et un cytoplasme granuleux sensible aux colorants neutres. Ils remplissent leurs fonctions immunitaires en s'attaquant aux micro-organismes étrangers, aux bactéries et aux levures. Les polynucléaires neutrophiles sont capables d'ingérer les particules étrangères (phagocytose), de libérer des substances antibactériennes et des enzymes protéolytiques (qui digèrent les bactéries tuées), ce qui est à l'origine des inflammations. L'accumulation de polynucléaires neutrophiles morts dans les zones infectées provoque l'apparition de pus.
- **Les basophiles** représentent environ 0,3% des polynucléaires (soit moins de 1% des leucocytes). Leur cytoplasme se caractérise par la présence de nombreuses grosses granulations et est sensible aux colorants basiques. Ils interviennent dans les réactions allergiques et inflammatoires. Ils sont en surnombre (basocytose) dans des hémopathies type syndromes myéloprolifératifs tels la leucémie myéloïde chronique.
- **Les polynucléaires éosinophiles** sont représentés à hauteur de 0,7% des polynucléaires (soit 1 à 3% des leucocytes et moins de 0,4 G/L). Ils présentent un noyau à 2 lobes, de nombreuses granulations sensibles aux colorants acides. Ils jouent un rôle dans les réactions allergiques et ont une fonction antiparasitaire par la libération des composants des granulations. De même que pour les polynucléaires neutrophiles, ils ont la capacité de phagocyter.

Les monocytes représentent 2 à 10% des globules blancs (soit 0,2 à 1 G/L). Leur taille, assez grande, varie de 15 à 22 µm. Leur noyau est réniforme et leur cytoplasme gris/bleu présente de fines granulations. Les monocytes sont présents 1 à 3 jours dans le sang, puis migrent et se localisent dans les différents tissus de l'organisme (on parle alors d'histiocytes à fonction macrophagique et immunitaire) où ils peuvent résider jusqu'à plusieurs années. Les monocytes contribuent à l'élimination de bactéries, de particules étrangères et des hématies en fin de vie (phagocytose).

Les lymphocytes constituent les éléments essentiels de la réponse immunitaire et représentent environ 25% des leucocytes soit 1,5 à 4 G/L chez un adulte. Ils sont formés de petits et grands lymphocytes. Les petits lymphocytes mesurent entre 9 et 12 µm de diamètre, ont un noyau arrondi ou ovalaire, un cytoplasme réduit sans granulations cytoplasmiques et leur rapport nucléo-cytoplasmique est relativement élevé. Les grands lymphocytes mesurent entre 12 et 15 µm de diamètre, ont un noyau ovalaire et un rapport nucléo-cytoplasmique plus faible que le petit lymphocyte. Certains grands lymphocytes présentent des granulations cytoplasmiques rouges (azurophiles).

Lors de la réalisation de la formule leucocytaire, tous les lymphocytes sont comptés en une seule et même catégorie, indépendamment de leur taille.

Parmi les lymphocytes, différentes sous-classes sont distinguées selon leurs fonctions et leurs marqueurs de surface membranaire (40). On en dénombre trois, B, T et Natural Killer, qui représentent respectivement environ 60 à 80%, 5 à 15% et 10 à 15% des lymphocytes de l'adulte. Ces sous populations lymphocytaires peuvent être caractérisées par des méthodes de cytométrie en flux après réalisation d'immunomarquages utilisant des anticorps spécifiques de lignée cellulaire.

- **Les lymphocytes B**, se développent à partir de la moelle osseuse (B pour *Bone marrow*). Ils sont impliqués dans la réponse immunitaire humorale, c'est à dire grâce aux anticorps qu'ils produisent (immunoglobulines) et qui serviront à la reconnaissance spécifique et à la destruction de l'antigène pathogène. L'élément fonctionnel majeur de cette lignée est le B cell receptor (BCR), une molécule d'immunoglobuline présente à sa surface en plusieurs exemplaire mais sur laquelle ne peut se fixer qu'un seul type d'antigène. Chaque lymphocyte B n'exprime donc qu'un seul type d'immunoglobuline à sa surface. Après être entrés en contact avec l'antigène correspondant, les lymphocytes B se transforment en plasmocytes et sécrètent des immunoglobulines anticorps de même spécificité que celles présentes à leur surface, et les diffusent dans l'organisme pour neutraliser les antigènes pathogènes. D'autres molécules sont présentes à leur surface, permettant de les identifier telles le CD19 (CD pour Cluster de Différenciation), le CD22, le CD20 et le CD23.

- **Les lymphocytes T** se développent à partir d'un progéniteur lymphoïde commun qui migre de la moelle osseuse vers le thymus où leur maturation se termine (T pour thymus). Les lymphocytes T sont impliqués dans la réponse immunitaire de type cellulaire. Tous les lymphocytes T peuvent reconnaître un antigène grâce au T cell receptor (TCR). À leur surface, différentes molécules sont présentes, le CD3, le CD2, le CD5, et le CD7. Les lymphocytes T peuvent être subdivisés en deux types, les lymphocytes T CD4 (molécule présente sur 2/3 des lymphocytes T), dits auxiliaires, sont responsables de la sécrétion de molécules, les cytokines, permettant la multiplication des lymphocytes T et B et la production d'anticorps. Les lymphocytes T CD8 (molécule présente sur 1/3 des lymphocytes T) peuvent être cytotoxiques ou suppresseurs et ont pour rôles l'élimination de cellules étrangères, cancéreuses ou infectées, et la régulation des réponses immunitaires.

- **Les lymphocytes Natural Killer** ont un précurseur commun avec les lymphocytes T et à ce titre ils présentent certains marqueurs membranaires identiques, mais également des antigènes leurs étant spécifiques CD56, CD57 et CD16. Ces cellules jouent un rôle important dans l'immunité naturelle puisqu'elles sont dotées de propriétés cytotoxiques naturelles, pouvant être augmentées par certaines cytokines, et qui sont employées spontanément envers les virus et les cellules cancéreuses.

Les polynucléaires, monocytes et lymphocytes Natural Killer sont impliquées dans la réponse immunitaire naturelle (innée), rapide et non spécifique. En revanche, les lymphocytes T et B sont impliqués dans l'immunité acquise, lente mais spécifique. Les lymphocytes B jouent un rôle majeur dans l'immunité humorale par leur capacité à sécréter des immunoglobulines.

Ces différentes sous-classes sont le plus souvent non différentiables au microscope comme les lymphocytes T CD8 et les lymphocytes Natural Killer qui sont morphologiquement proches, sous forme de grands lymphocytes granuleux.

a) Techniques de numération et d'identification des cellules sanguines en pratique quotidienne

Les automates d'hématologie cellulaire, habituellement utilisés dans les laboratoires de biologie médicale, permettent de décompter les cellules sanguines et de détecter des anomalies (quantitatives et qualitatives) des différents éléments figurés du sang. Ces automates permettent la réalisation d'un hémogramme (appelé également numération formule leucocytaire) qui dénombre les différentes cellules par volume de sang (plaquettes, hématies, et les différentes catégories leucocytaires). En fonction de critères quantitatifs et qualitatifs, des frottis sanguins sont réalisés par méthode manuelle ou automatique dans le but d'être analysés en microscopie. En cas d'anomalies qualitatives portant notamment sur les leucocytes, cette analyse morphologique microscopique permet d'orienter le biologiste vers une pathologie réactionnelle (la plus fréquente) ou une hémopathie. En cas de suspicion d'hémopathie et en fonction du type d'hémopathie suspectée, différents examens complémentaires sont réalisés. Ces examens peuvent nécessiter une analyse de la moelle (après ponction dans l'os sternal ou iliaque), ou d'autres examens sanguins tels l'immunophénotypage (par cytométrie en flux) des cellules suspectes immunomarquées, ou une étude cytogénétique voire moléculaire du sang et/ou de la moelle.

b) Les maladies du sang ou hémopathies (38, 41)

Ces maladies se traduisent par une diminution des cellules normales, ou cytopénie (pour les hématies on parlera d'anémie, diminution du taux d'hémoglobine, et pour les plaquettes de thrombopénie), ou par une augmentation des cellules normales (polyglobulie pour les hématies) ou par la présence de cellules anormales (cas des leucémies).

Plus de 80 maladies du sang sont répertoriées et peuvent toucher n'importe quelle catégorie cellulaire (hématies, leucocytes ou plaquettes). Ces maladies peuvent être classées en trois catégories, i) les cancers, parmi lesquels les leucémies, les lymphomes et les myélomes sont les plus représentés, ii) les maladies dues à une déficience de la moelle osseuse (myélodysplasie), entraînant un dysfonctionnement dans le processus d'hématopoïèse comme dans le cas d'une leucémie myélo-monocytaire chronique, iii) ou encore les maladies génétiques qui sont le plus souvent héréditaires, telles l'hémophilie, la drépanocytose (mutation de l'hémoglobine) ou encore la thalassémie (défaut de synthèse d'hémoglobine).

Dans le cadre de ces travaux de thèse nous nous sommes intéressés à l'une des formes cancéreuses des maladies sanguines, **la leucémie**.

Il existe différents types de leucémies, toutes caractérisées par une prolifération de leucocytes anormaux, et qui se distinguent selon les lignées cellulaires à partir desquelles elles se sont développées (42). On distingue les leucémies « aiguës » et « chroniques » :

c) Les leucémies aiguës

Elles évoluent rapidement et peuvent engager le pronostic vital du patient si elles ne sont pas traitées. Les leucémies aiguës se caractérisent par une prolifération anarchique dans la moelle osseuse de cellules cancéreuses immatures, issues de précurseurs anormaux de leucocytes. Cette prolifération anormale empêche le développement et la différenciation des cellules sanguines normales, entraînant une diminution des hématies et des plaquettes.

Il existe globalement deux types de leucémies aiguës, la leucémie aiguë lymphoblastique (issue de précurseurs de lymphocytes), et la leucémie aiguë myéloblastique (issue de précurseurs de polynucléaires ou monocytes). L'Institut de veille sanitaire (InVS) estime les taux d'incidence standardisés mondiaux de 2011 à 4,9/100 000 personnes par année pour les hommes et à 4,1/100 000 pour les femmes (43).

d) Les leucémies chroniques

Elles sont caractérisées par une prolifération de cellules matures, ou d'aspect mature. On en distingue deux types, la leucémie myéloïde chronique, caractérisée par un excès de leucocytes de la lignée myéloïde, et la leucémie lymphoïde chronique (LLC), dont les cellules tumorales sont issues de la lignée des lymphocytes. La leucémie lymphoïde chronique est la plus fréquente des leucémies de l'adulte et présente une incidence croissante avec l'âge. Parmi les LLC, 5% affectent les lymphocytes de type T, et 95% les lymphocytes de type B. En 2011, l'InVS estime l'incidence standardisée mondiale des LLC à 3,6/100 000 pour les hommes et à 2,1/100 000 pour les femmes (43). Au diagnostic, l'âge moyen est de 72 ans et 27% des patients ont moins de 65 ans.

Le choix de la LLC de type B à tout d'abord été motivé par l'accès facile et en grande quantité aux cellules sanguines de patients, mais également par la possibilité d'accéder aux échantillons cryopréservés. Enfin, le travail d'équipe réalisé depuis plusieurs années ainsi

que les compétences humaines disponibles sur place ont contribué au choix de la leucémie lymphoïde chronique de type B comme sujet d'étude pour ces travaux de thèse.

2) La Leucémie Lymphoïde Chronique (44-48)

Les patients souffrant de LLC sont initialement asymptomatiques, les symptômes survenant avec l'évolution de la maladie. De ce fait, la LLC est souvent découverte devant une augmentation des lymphocytes sanguins, de façon fortuite à l'occasion de la réalisation systématique d'un hémogramme et ce, dans plus de 80% des cas.

Chez une personne non atteinte, la durée de vie des lymphocytes B présents dans le sang varie de plusieurs jours à plusieurs mois. Au terme de cette période, ces cellules entrent en apoptose et sont remplacées par de nouveaux lymphocytes B selon le processus de production continue de l'hématopoïèse. Dans le cas d'une personne atteinte de LLC, une partie des lymphocytes B n'entre plus en apoptose tandis qu'une autre partie se multiplie anarchiquement, provoquant une accumulation de ces cellules B clonales d'aspect mature dans le sang mais également dans les ganglions lymphatiques, la rate ou encore la moelle osseuse. Jusqu'en 2005, c'est ce mécanisme de défaut d'apoptose qui a été considéré comme le mécanisme principal de l'accumulation du clone leucémique. Depuis cette période, il a été démontré que la prolifération *in vivo* était à l'origine de cette accumulation (49). La Figure 9 permet d'effectuer une comparaison de deux frottis sanguins obtenus chez un volontaire pour le premier (A) et chez un patient atteint de LLC de type B très hypercytaire pour le second (B). Sur cette figure on constate la présence d'une hyperlymphocytose évidente (B) correspondant à l'accumulation des lymphocytes dans le sang.

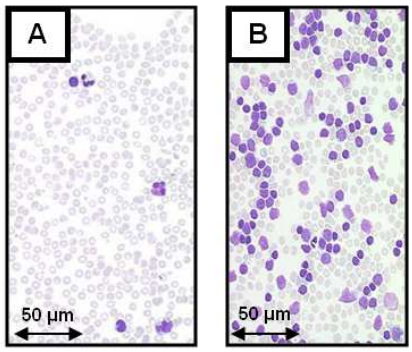


Figure 9 : Vues au microscope de frottis sanguins colorés

Vues à l'objectif 40x de frottis sanguins colorés au May-Grünwald Giemsa d'un individu sain (A) et d'un patient LLC-B présentant une lymphocytose (B).

a) Diagnostic

Toute lymphocytose supérieure à 4 G/L (soit 4 000/mm^3) détectée lors de la réalisation d'un hémogramme, et persistant au-delà de 3 mois chez un adulte, nécessite des investigations biologiques complémentaires. Si l'hyperlymphocytose est importante et l'aspect du frottis sanguin est évocateur, ces investigations sont réalisables immédiatement. Une analyse microscopique de frottis sanguins sur lame de verre standard, colorés au MGG permet d'apprécier la morphologie des cellules et de détecter des anomalies éventuelles sur les lymphocytes. Ensuite, un immunophénotypage de ces derniers par cytométrie en flux, basé sur l'étude des marqueurs de surface lymphocytaires permet :

- D'obtenir un dénombrement des sous catégories lymphocytaires (T, B, NK)
- De mettre en évidence une prolifération clonale B
- D'établir un score immunologique, nommé score de Matutes (50) (ou score Royal Marsden Hospital) en fonction du profil phénotypique obtenu (Figure 10)
- Enfin, d'éliminer d'autres syndromes lymphoprolifératifs de pronostic et de traitement différent (51)

La détection d'une quantité de lymphocytes B clonaux supérieure à 5 000/mm^3 et un score de Matutes supérieur à 3, sont nécessaires pour poser le diagnostic de la LLC-B.

Après confirmation de la maladie, la détermination de son stade d'évolution est nécessaire avant toute indication thérapeutique (42, 52). Cette classification est réalisée *via* les classifications de Rai (53) ou de Binet (54).

Cotation	1	0
CD5	+	-
CD23	+	-
Expression sIg monotypique	Faible	Forte
FMC7	-	+
Expression de CD79b	Faible	Forte

(Cymbalista, 2013)

Figure 10 : Score de Matutes

Le calcul du score de Matutes est réalisé d'après cinq critères. Un point de ce score est obtenu pour chaque critère rempli.

b) Classification de Binet (42, 44, 54)

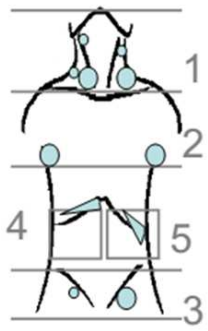
En France, la classification de Binet est la plus couramment utilisée (Figure 11). Elle tient compte des informations fournies par l'hémogramme et du nombre d'aires lymphoïdes palpables.

Stade A : Environ 80% des patients sont classés au stade A. Le nombre d'aires lymphoïdes palpables est inférieur à 3 et il n'existe pas de cytopénie associée à la lymphocytose. À ce stade, aucun traitement n'est nécessaire, mais une surveillance clinique et biologique est réalisée tous les 6 à 12 mois afin de prévenir une éventuelle progression. Progression qui pour plus de la moitié de ces patients n'a jamais lieu, ne nécessitant donc aucun traitement.

Stade B : L'augmentation du volume de la rate, ou des ganglions du patient dans au moins 3 aires, permet de classer le patient au stade B de la maladie. Ce changement de volume est notamment dû à l'accumulation des lymphocytes tumoraux dans ces organes dits « réservoirs ». Le stade B concerne environ 15% des patients.

Stade C : Le stade C représente environ 5% des cas au diagnostic et se définit par la présence d'une anémie (Hémoglobine < 10 g/dL) et/ou une thrombopénie (plaquettes inférieures à 100 G/L) quel que soit le nombre de ganglions atteints.

L'évolution de la maladie chez les patients de stade A est très variable et difficile à déterminer. L'étude de facteurs pronostiques peut donc être réalisée dans le but de détecter les cas pouvant évoluer dans des formes plus graves, et ceci dès leur diagnostic.



Stade	Aires lymphoïdes palpables	Hémoglobine < 10 g / 100 ml Ou Plaquettes < 100 000 / mm³
A	< 3	Non
B	≥ 3	Non
C	Quel que soit le nombre d'aires lymphoïdes palpables	Oui

(Haute Autorité de Santé, 2011)

Figure 11 : Schéma des 5 aires lymphoïdes palpables (gauche) et tableau de classification de Binet

La détermination par le clinicien du nombre de zones lymphoïdes palpables associée au taux d'hémoglobine ou de plaquettes dans le sang permet d'attribuer le stade d'évolution de la maladie du patient (A, B ou C)

c) Les principaux facteurs pronostiques (48, 55, 56)

Pour les patients ne nécessitant pas la mise en place d'un traitement, trois facteurs pronostiques biologiques simples sont répertoriés :

- Le temps de doublement des lymphocytes : qui représente le temps nécessaire au doublement du nombre de cellules tumorales chez le patient. Chez les patients présentant un temps de doublement inférieur à 12 mois, la médiane de survie est de 5 ans (période au-delà de laquelle la moitié des patients sera décédée).
- Le taux de Beta 2 microglobuline sérique, de mauvais pronostic s'il est supérieur à la normale.
- L'expression du CD38 à la surface des lymphocytes B clonaux (effectuée lors de la réalisation de l'immunophénotypage diagnostique initial)

Ces trois facteurs permettent de déterminer la périodicité de la surveillance.

D'autres marqueurs sériques de risques de progression de la maladie (chez les patients de stade A) sont utilisables mais non obligatoires. Ils évaluent la multiplication des cellules anormales par le taux d'expression d'une enzyme, la thymidine kinase, ainsi que l'augmentation du taux de CD23 soluble.

Concernant les autres facteurs pronostiques tels le statut mutationnel, l'expression du gène ZAP-70 et l'analyse cytogénétique, ils ne sont toujours pas nécessaires à l'heure actuelle lors de l'établissement du bilan, mais des études prospectives sont en cours.

- Mutations des immunoglobulines (récepteurs d'antigènes dont font partie les anticorps) : les patients ne présentant pas de telles mutations ont une médiane de survie estimée à 8 ans contre 25 ans pour ceux présentant un profil muté. Ceci quelque soit le stade diagnostiqué dans la classification de Binet.

- L'expression d'un gène, ZAP-70 : son expression dans les lymphocytes B tumoraux est corrélée au profil non muté des gènes des immunoglobulines et de ce fait, représente un facteur de mauvais pronostic. Cependant, ce marqueur n'apparaît pas toujours fiable.

- Les anomalies cytogénétiques (57) : relatives au matériel génétique des cellules (anomalies chromosomiques) propres à l'hémopathie permettent d'identifier des cas pathologiques de stade A avec des évolutions différentes. Leur valeur pronostique a été étudiée sur de larges séries de patients. Elles ont été mises en évidence dans plus de 80 % des cas. Il existe 4 types d'anomalie possibles : les délétions 11q, 13q et 17 p (délétion = perte d'un fragment de chromosome) et la trisomie 12. La délétion du 13q, localisée sur le chromosome 13, est retrouvée dans 55% des cas, et est de bon pronostic si elle est isolée. La trisomie 12 (15% des cas) est associée à une résistance au traitement par fludarabine et à une réévolutivité rapide après traitement. La technique FISH (Fluorescence In Situ Hybridization) est l'examen de référence pour déterminer les facteurs pronostiques cytogénétiques.

Il est à noter que la complexité et le coût élevé de ces tests ne permettent pas de les réaliser « en routine » pour tous les patients diagnostiqués au stade A. Seuls ceux intégrant un protocole thérapeutique ou un projet thérapeutique curateur en bénéficient.

Concernant les stades B et C de la maladie, ils nécessitent tous deux des traitements adaptés à chaque patient. Ces traitements sont fonction de critères d'évolutivité de la maladie et de symptômes systémiques (44) tels une perte de poids d'au moins 10% dans les 6 derniers mois, une fièvre supérieure à 38°C pendant au moins deux semaines (non due à une infection), un temps de doublement des lymphocytes inférieur à 6 mois, une apparition/aggravation d'une anémie ou d'une thrombopénie par exemple.

d) Traitements de la LLC (44, 52)

Les traitements administrés sont dans tous les cas discutés en réunion de concertation pluridisciplinaire. L'inclusion dans un protocole d'étude prospectif est prioritaire. Le but de ces traitements est de contenir l'évolution de la maladie, d'en faire disparaître certaines manifestations ou encore de préserver la qualité de vie des patients. Cependant, même si ces traitements peuvent, pour certains, considérablement faire diminuer le nombre de cellules cancéreuses, ils ne les éradiquent jamais totalement.

La leucémie lymphoïde chronique peut donc être traitée par :

- **Chimiothérapie** dans la majorité des cas, *via* différents médicaments comme le Chlorambucil ou le Cyclophosfamide (agents alkylants), les Anthracyclines comme la Doxorubicine (antibiotique cytotoxique) ou encore les analogues des purines comme la Fludarabine et la Cladribine (qui inhibent la synthèse d'ADN entraînant un arrêt de la croissance cellulaire). Les analogues des purines sont très efficaces envers les LLC mais touchent l'ensemble des lymphocytes (sains y compris) entraînant une immunodéficience des patients. L'autogreffe de cellules souches est également rendu difficile après ces traitements.

- L'administration d'**anticorps monoclonaux** permet une destruction ciblée de cellules tumorales. Ces anticorps comme le rituximab et l'alemtuzumab (commercialisés sous les noms de MABTHERA® et MABCAMPATH® respectivement), sont spécifiques aux antigènes présents à la surface des cellules cancéreuses et s'y fixent. Le rituximab est dirigé sur le CD20 à la surface des lymphocytes B et l'alemtuzumab sur le CD52. Ces cellules ciblées deviennent alors sujettes aux défenses immunitaires du patient.

- **Allogreffe** dans certains cas plus rares, comme par exemple lors d'une résistance du patient à la chimiothérapie, une greffe de cellules souches hématopoïétiques provenant d'un donneur compatible peut être envisagée.

À la suite de ces traitements, des effets secondaires peuvent apparaître comme différentes infections, une anémie (58) ou encore une thrombopénie. L'évolution de la LLC en lymphome (type syndrome de Richter) est également possible mais reste exceptionnelle.

Le diagnostic, l'établissement de facteurs pronostiques et la détermination du stade de la maladie sont donc de fort intérêt car ils jouent un rôle primordial dans le choix de la stratégie thérapeutique à adopter. Cependant, ces approches restent complexes et/ou coûteuses. Des approches diagnostiques et pronostiques basées sur des technologies différentes apparaissent de fort intérêt dans la mesure où ces technologies apporteraient des informations complémentaires potentiellement à moindre coût.

III) Spectroscopies vibrationnelles

1) Généralités

La spectroscopie consiste à mesurer et interpréter le spectre électromagnétique produit par une substance qui émet ou absorbe un rayonnement. La spectroscopie optique repose sur la séparation de la lumière par un prisme ou un réseau de diffraction. Un spectre représente l'intensité du rayonnement absorbé ou émis en fonction de sa fréquence.

Il existe différents types de spectroscopies optiques, la spectroscopie d'émission de fluorescence et les spectroscopies vibrationnelles, toutes basées sur l'interaction lumière-matière.

Dans ce travail, les spectroscopies vibrationnelles ont été utilisées pour l'analyse de matériel biologique. Ces techniques présentent l'avantage de n'être ni invasives ni destructives, et ne nécessitent ni marquage ni coloration des échantillons. Leur principe est basé sur la mesure des fréquences de vibration des liaisons atomiques de molécules simples, ou de systèmes plus complexes (tels les tissus), lorsqu'ils sont soumis à l'application d'une onde électromagnétique excitatrice. En fonction de la fréquence de cette onde, notée v_0, les liaisons atomiques de l'échantillon analysé vont subir des élongations ou des déformations d'angle (dans le plan ou dans l'espace), nommées modes de vibration (Figure 12).

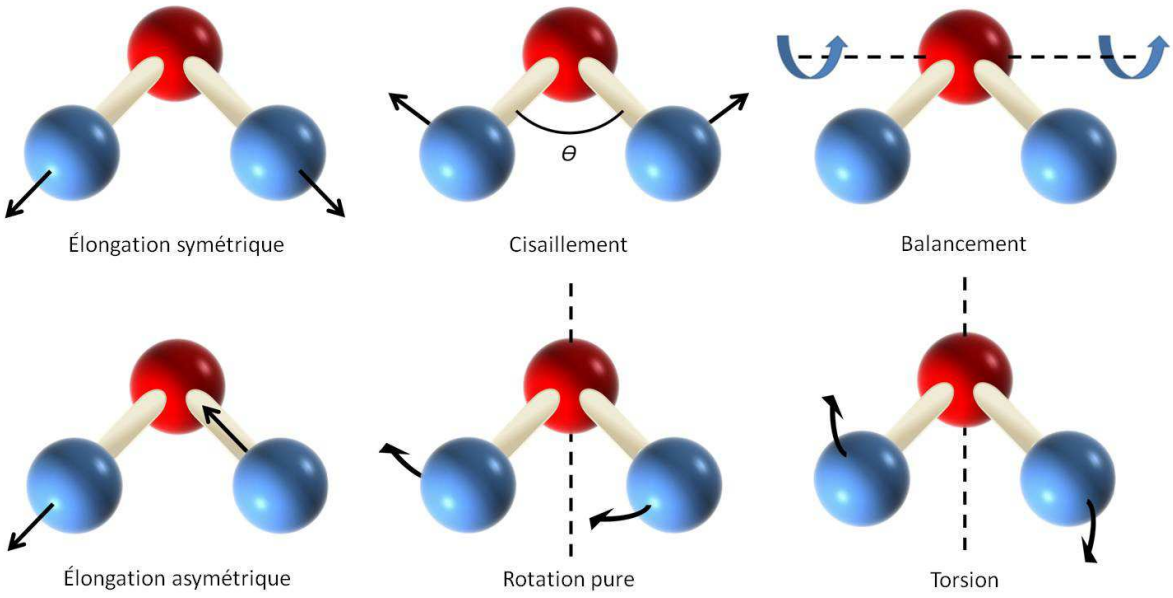


Figure 12 : Exemples de modes fondamentaux de vibration moléculaire

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).

Dans le cas d'une molécule diatomique, un seul mode de vibration est possible, l'élongation. Cependant, dans le cas de molécules polyatomiques, les liaisons deviennent plus nombreuses et leur géométrie plus complexe. Ainsi, en considérant N atomes, chacun possède 3 degrés de liberté en translation et 3 degrés de liberté en rotation dans l'espace. Si ces atomes sont reliés entre eux avec des liaisons de différentes longueurs et différents angles, la molécule possède 3N degrés de liberté. Parmi eux, 6 degrés correspondent soit à une translation soit à une rotation de la molécule entière. Les modes de vibrations d'une molécule à N atomes sont donc au nombre de 3N-6. On notera un cas particulier que sont les molécules linéaires (exemple : 3 atomes formant un angle de 180°), pour lesquelles l'un des modes correspond à une rotation autour de l'axe de cette molécule. On ramènera donc leur nombre de modes à 3N-5.

Ainsi, l'enregistrement des variations d'énergie entre l'onde électromagnétique excitatrice et l'onde électromagnétique obtenue après interaction avec l'échantillon, va permettre d'estimer la fréquence et l'intensité de chacun de ces modes. Ces informations sont représentées sous la forme d'un spectre, traduisant la nature et la structure des différentes molécules dont l'échantillon est composé. En d'autres mots, un spectre représente une véritable signature moléculaire de l'échantillon analysé.

On notera que l'unité de référence pour les spectroscopies vibrationnelles est le nombre d'onde, représentant le nombre d'oscillations d'une onde lumineuse par unité de longueur. Il est exprimé comme l'inverse de la longueur d'onde (exprimée en cm). Son unité est donc le cm^{-1}.

Il existe différents types de spectroscopies vibrationnelles. Dans ce travail, les spectroscopies d'absorption infrarouge (IR) et de diffusion Raman ont été appliquées à l'étude de cancers cutanés et sanguin.

2) Principe des spectroscopies vibrationnelles

a) Spectroscopie d'absorption infrarouge (59)

La région de rayonnement infrarouge s'étend de 13000 à 10 cm^{-1} et est subdivisée en 3 domaines, le « proche infrarouge » (13000-4000 cm^{-1}), le « moyen infrarouge » (4000-400 cm^{-1}) et le « lointain infrarouge » (400-10 cm^{-1}). Par la suite, nous nous intéresserons

exclusivement à la région du « moyen infrarouge » puisque c'est dans cette région que la majorité des molécules présente leurs fréquences de vibration. Cette région apparaît donc la plus adaptée à l'analyse moléculaire d'échantillons.

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. Notons toutefois que les niveaux d'énergie et la polarisation du rayonnement incident doivent vérifier certaines règles dites « règles de sélection » (cf infra) afin que l'absorption ait lieu.

Un spectre infrarouge présente donc des valeurs d'absorbance (ou de transmittance) en fonction de nombres d'onde. Les valeurs d'absorbance et de transmittance sont calculées par la loi de Beer-Lambert suivant l'équation suivante :

$$I = I_0 \times e^{-\varepsilon c l} \qquad (1)$$

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

La transmittance T est définie par :

$$T = \frac{I}{I_0} \qquad (2)$$

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

$$A = \log\left(\frac{1}{\mathrm{T}}\right) = \log\left(\frac{I_0}{I}\right) \qquad (3)$$

Un spectre infrarouge apporte donc une information qualitative quant à la composition moléculaire de l'échantillon analysé puisque chaque bande du spectre correspond à un groupe de vibrations caractéristiques des molécules de l'échantillon. Il apporte également une information quantitative puisque l'absorbance, en un nombre d'onde donné, est proportionnelle à la concentration des molécules ayant vibré. Cependant, cette proportionnalité avec la concentration est limitée à des valeurs d'absorbance situées en général entre 0,3 et 2.

b) Spectroscopie de diffusion Raman (60)

L'effet de diffusion Raman est constaté lors de l'application d'un faisceau lumineux monochromatique de forte énergie (envoi de photons) sur l'échantillon analysé. Au cours de l'interaction entre les molécules de l'échantillon et les photons incidents, d'énergie hv_0 (où h est la constante de Planck = 6,6262x10^{-34} J.s et v_0 la fréquence de l'onde excitatrice), les molécules passent d'un niveau réel d'énergie à un niveau virtuel. Puis redescendent à leur niveau d'énergie originel.

Ce changement d'état énergétique, schématisé par le diagramme de Jablonski en Figure 13, permet d'observer une faible diffusion **élastique** des photons dont l'ordre de grandeur est de 1 sur 10^4 photons. Il est à noter qu'ainsi diffusés, ils ne présentent pas de variation d'énergie.

Ce phénomène est appelé **la diffusion Rayleigh**.

Plus rarement encore, le niveau d'énergie originel des molécules est légèrement modifié après l'excitation. Des photons sont alors diffusés de façon **inélastique**, avec un ordre de grandeur de 1 sur 10^8 photons. C'est à dire qu'un échange d'énergie a lieu entre le rayonnement et la matière, se traduisant par une différence d'énergie ($hv_0 \pm hv_m$) entre le photon incident de fréquence v_0 et le photon diffusé de fréquence ($v_0 \pm v_m$).

Cette différence d'énergie hv_m correspond à un mode de vibration de fréquence v_m d'une molécule composant l'échantillon.

Ce phénomène est nommé **diffusion Raman**.

On observe deux types de diffusion Raman inélastique :

- **Stokes :** lorsque la matière absorbe de l'énergie aux photons incidents, l'énergie des photons diffusés diminue et vaut ($hv_0 - hv_m$).

- **Anti-Stokes :** lorsque la matière cède de l'énergie aux photons incidents, l'énergie des photons diffusés augmente et vaut ($hv_0 + hv_m$).

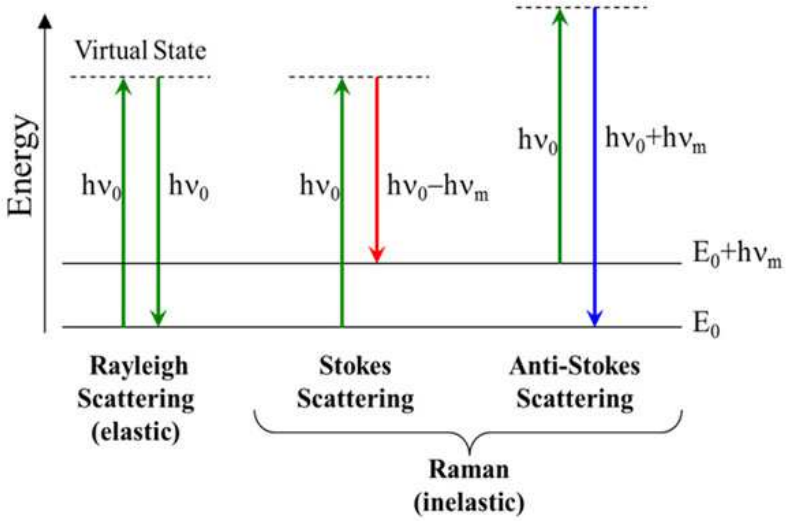


Figure 13 : Diagramme de Jablonski

Le diagramme de Jablonski présente la différence, en termes d'énergie, entre la diffusion élastique (effet RayLeigh) et la diffusion inélastique (effet Raman Stokes et Anti-Stokes) de la lumière. Avec $h\nu_0$ l'énergie du photon incident, $h\nu_m$ la différence d'énergie entre le photon incident et le photon diffusé et E_0 l'état énergétique fondamental de la molécule.

Un spectre enregistré par spectroscopie Raman représente donc des intensités (nombres de photons diffusés par unité de temps) en fonction de nombres d'onde.

c) Règles de sélection

Les variations des niveaux d'énergie des photons sont directement liées à la polarisabilité des liaisons moléculaires. En d'autres mots, sous l'effet du champ électrique E_0 du rayon incident, le nuage électronique de la liaison moléculaire va se déformer et acquérir un moment dipolaire μ_{tot} tel que :

$$\mu_{tot} = \mu_p + \alpha E_0 \cos(v_0 t) \qquad (4)$$

Avec μ_p le moment dipolaire permanent, α le tenseur de polarisabilité, E_0 le champ électrique, v_0 la fréquence du rayon incident et t le temps.

En écrivant le moment dipolaire permanent et le tenseur de polarisabilité autour de la position d'équilibre on a :

$$\mu_p = \mu_{p(0)} + \sum_{n=1}^{3N-6} \left(\frac{d\mu_p}{dQ_n}\right) Q_n \qquad (5)$$

Et $$\alpha = \alpha_0 + \sum_{n=1}^{3N-6} \left(\frac{d\alpha}{dQ_n}\right) Q_n \quad (6)$$

Avec Q_n la coordonnée normale du mode de vibration n de la molécule (n = 1, 2,…,3N-6).

Dans une approximation harmonique et pour une fréquence v_n et une amplitude Q_0 données on a :

$$Q_n = Q_0 \cos(v_n t) \quad (7)$$

Ainsi, une expression générale du moment dipolaire total peut être obtenue selon les équations précédentes :

$$\mu_{tot} = \mu_{p(0)} + \sum_{n=1}^{3N-6} \left(\frac{d\mu_p}{dQ_n}\right) Q_0 \cos(v_n t) + \alpha_0 E_0 \cos(v_0 t) + \frac{1}{2} E_0 \sum_{n=1}^{3N-6} \left(\frac{d\alpha}{dQ_n}\right) Q_0 \{\cos[(v_0 - v_n)t] + \cos[(v_0 + v_n)t]\} \quad (8)$$

D'après cette équation générale :

Le terme $\sum_{n=1}^{3N-6} \left(\frac{d\mu_p}{dQ_n}\right) Q_0 \cos(v_n t)$, correspond à l'absorption infrarouge d'un dipôle vibrant à la fréquence v_n. Pour que ce terme soit non nul, une variation du moment dipolaire μ_p doit avoir lieu.

Le terme $\alpha_0 E_0 \cos(v_0 t)$, est à l'origine de la diffusion Rayleigh (diffusion élastique).

Enfin le terme $\frac{1}{2} E_0 \sum_{n=1}^{3N-6} \left(\frac{d\alpha}{dQ_n}\right) Q_0 \{\cos[(v_0 - v_n)t] + \cos[(v_0 + v_n)t]\}$, représente la diffusion Raman d'un dipôle oscillant à la fréquence $(v + v_n)$ (anti-Stokes) ou à la fréquence $(v - v_n)$ (Stokes). Pour que ce terme soit non nul, une variation de la polarisabilité α doit avoir lieu.

Lorsqu'une molécule possède un centre de symétrie et que des vibrations symétriques par rapport à ce centre ont lieu (*i.e.* moment dipolaire constant et variation de polarisabilité), alors ces vibrations ne seront actives qu'en Raman. En revanche, si ces vibrations sont antisymétriques, elles ne seront actives qu'en infrarouge.

Même si ces règles apparaissent simples dans le cas de molécules diatomiques, il devient difficile de savoir si un signal Raman ou infrarouge peut exister dans le cas de systèmes plus

complexes. On notera également que ces règles ont été déterminées dans une approximation harmonique et que dans des cas réels, on peut observer sur les spectres enregistrés des intensités faibles dues à des harmoniques de fréquence nv_n ou encore à des combinaisons de fréquences telles $v_n + v'_n$.

3) Instrumentation

a) Le spectromètre infrarouge à transformée de Fourier

Un dispositif expérimental de spectromètre infrarouge à transformée de Fourier (IRTF) est présenté en Figure 14.

La source de rayonnement infrarouge est obtenue en portant à haute température (1500°C) un solide, tel le Globar (pour Glowing-bar), constitué de carbures de silicium. Le Globar rayonne par incandescence et permet d'émettre dans le domaine spectral du moyen-infrarouge (de 1 à 100 µm ou 50 à 5000 cm^{-1}). Cependant, son utilisation génère une grande quantité de chaleur et nécessite un refroidissement.

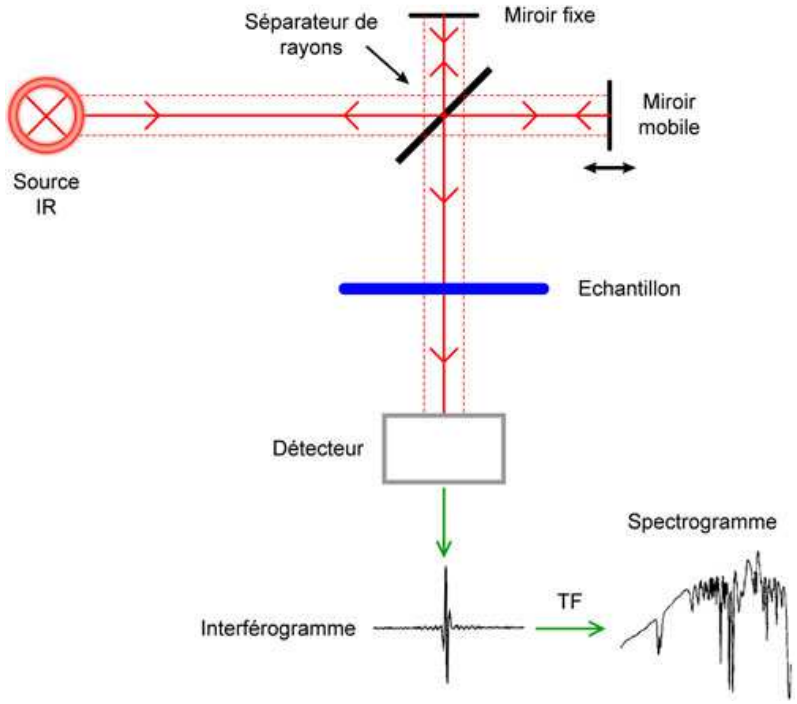
L'interféromètre de Michelson est un dispositif permettant de générer des interférences au sein d'un rayonnement. Par ce procédé, il devient possible d'analyser les différentes longueurs d'onde de la source polychromatique simultanément.

Cet interféromètre comprend une lame séparatrice semi-transparente vers laquelle est orientée la source lumineuse polychromatique. Cette lame divise le rayon incident en deux faisceaux d'égales intensités (50% du rayon incident chacun). Le premier faisceau est réfléchi vers un miroir fixe, alors que le second est transmis vers un miroir mobile perpendiculaire au premier. Les deux faisceaux sont alors réfléchis par les miroirs et recombinés au niveau de la lame séparatrice. Notons toutefois que chacun des deux faisceaux est de nouveau séparé en deux faisceaux d'intensité réduite de moitié lors du second passage au travers de la lame séparatrice. Ainsi, le faisceau recombiné arrivant sur l'échantillon représente 50% de l'intensité du rayon initial (25% + 25%).

Le miroir mobile s'éloigne de la lame séparatrice à une vitesse constante, ajoutant ainsi une distance optique supplémentaire pour le second faisceau. En fonction de cette distance, les deux faisceaux recombinés vont présenter les mêmes amplitudes mais avec des différences de phase, et seront susceptibles d'interférer entre eux. On distingue deux types

d'interférence, constructive (les ondes se cumulent) et destructive (les ondes s'annulent), qui sont fonction de la différence de trajet optique par rapport aux différentes longueurs d'onde de la source polychromatique. Par la suite, le faisceau reconstitué arrive sur un détecteur MCT (Mercury Cadmium Telluride) où l'ensemble des interférences constructives et destructives sont traduites sous la forme d'un interférogramme.

Remarques : l'interférogramme d'une source monochromatique est une fonction sinusoïdale. Celui d'une source polychromatique est construit comme la somme des fonctions sinusoïdales correspondant à chacune des longueurs d'onde de la source. Ainsi, chaque point de l'interférogramme correspond à une position du miroir mobile, et représente une information propre à toutes les longueurs d'onde de la source lumineuse.



(Thèse Mathias Borella, 2006)

Figure 14 : Schéma instrumental d'un spectromètre IRTF

Un spectromètre infrarouge à transformée de Fourier est composé d'une source lumineuse infrarouge, d'un interféromètre de Michelson et d'un dispositif de calcul.

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

b) Le spectromètre Raman

Un spectromètre Raman est composé d'une source lumineuse monochromatique orientée et filtrée par différents dispositifs, tels des miroirs, filtre notch, fente et trou confocal. Un réseau dispersif permet, après excitation de l'échantillon, de séparer les différentes ondes Raman diffusées en fonction de leur fréquence respective. Enfin, un détecteur CCD (Charged-Couple Device) permet l'enregistrement simultané des intensités de ces ondes, représentées sous la forme d'un spectre Raman.

Un dispositif expérimental est présenté en Figure 15.

Contrairement à la spectroscopie infrarouge à transformée de Fourier, la source lumineuse utilisée dans un spectromètre Raman est monochromatique, et peut être choisie à différentes longueurs d'onde. Il est cependant à noter que plus la longueur d'onde est basse (fréquence élevée), plus la diffusion et la fluorescence enregistrées au niveau de l'échantillon seront intenses. Le choix de la longueur d'onde d'excitation est donc dépendant de l'échantillon (et parfois du support) considéré pour l'étude.

La source lumineuse est orientée vers l'échantillon *via* différents miroirs et filtres atténuateurs permettant, si besoin, de réduire son intensité afin de ne pas dégrader l'échantillon étudié.

Un filtre coupe-bande, ou filtre notch, et utilisé afin de filtrer le rayonnement en fonction de ses différentes longueurs d'onde. Ainsi, le rayonnement monochromatique issu de la source est intégralement réfléchi par ce filtre pour être redirigé vers l'objectif du microscope.

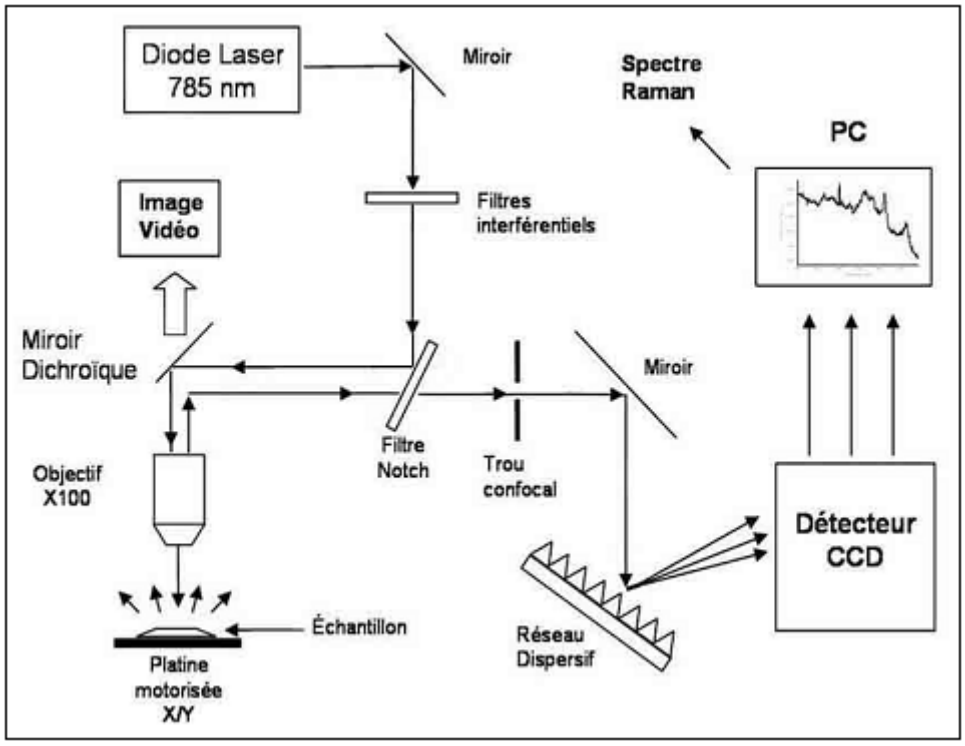
Après excitation de l'échantillon, les rayons diffusés (Rayleigh et Raman) sont renvoyés, *via* l'objectif, vers le filtre notch qui ne transmettra que les ondes issues de la diffusion Raman. En effet, la diffusion Rayleigh présentant la même fréquence que la source excitatrice, sera réfléchie.

Le trou confocal et la fente d'entrée du spectromètre sont les prochaines étapes dans la progression des ondes Raman. Le trou confocal permet de réduire le signal propre à un volume donné de l'échantillon, alors que la largeur de la fente (non représentée sur la Figure 15) détermine la résolution spectrale de l'appareil. Il est à noter qu'une réduction de

l'ouverture de la fente traduit une meilleure résolution mais également une intensité moindre du rayon.

Le réseau dispersif est un dispositif permettant de séparer les ondes d'un rayon lumineux en fonction de leur fréquence, et est caractérisé par un nombre de traits par millimètre. Plus ce nombre de traits est élevé plus la résolution spectrale sera haute. Cependant, cette augmentation de la résolution entraine une diminution de la fenêtre spectrale analysable.

Après séparation des ondes par le réseau dispersif, leurs intensités sont simultanément enregistrées *via* un détecteur CCD, permettant l'obtention d'un spectre Raman.



(Thèse Florence Draux, 2009)

Figure 15 : Schéma instrumental d'un spectromètre Raman

Un spectromètre Raman est composé d'une source lumineuse monochromatique, dirigée par différents miroirs et filtres vers l'échantillon. Les ondes Raman diffusées par ce dernier sont séparées par un réseau dispersif en fonction de leur fréquence respective et enregistrées simultanément sous la forme d'un spectre Raman.

4) Applications des spectroscopies vibrationnelles en oncologie

Comme précisé précédemment, les spectroscopies vibrationnelles sont des technologies non destructives, non invasives, ne nécessitant ni coloration ni marquage, et permettant de sonder la composition moléculaire des échantillons analysés. De ce fait, beaucoup d'études employant ces technologies ont été menées dans l'optique d'apporter de nouveaux outils d'aide au diagnostic différentiel de cancers, comme par exemple les cancers gastro-intestinaux, du poumon, du col de l'utérus, du sein,… (61, 62).

Les cancers cutanés ont fait l'objet de nombreuses études par spectroscopies vibrationnelles (63-65). Ces études ont permis de définir des marqueurs spectroscopiques, différenciant les tumeurs des tissus sains, et traduisant des variations de structure ou de contenu en lipides, protéines et acides nucléiques. De plus, l'examen histologique sur une exérèse complète de la tumeur demeurant le gold standard pour poser le diagnostic, des études *in vivo* et « temps réel » ont été menées *via* des sondes utilisant la spectroscopie vibrationnelle afin d'apporter des informations propres à la tumeur, en amont de toute intervention chirurgicale (63, 66).

En ce qui concerne la leucémie lymphoïde chronique, des études ont également été menées afin d'identifier des marqueurs spectroscopiques caractérisant les lymphocytes B néoplasiques (67-71). Ces études ont montré des différences biochimiques entre les cellules saines et pathologiques, notamment sur leur contenu en ADN, ARN, lipides et protéines.

La spectroscopie vibrationnelle montre donc un réel potentiel dans la discrimination cellulaire et tissulaire des échantillons étudiés (sains et pathologiques). Cependant dans beaucoup d'études, ce potentiel discriminatoire est renforcé en utilisant des outils chimiométriques, des méthodes statistiques multivariées (72), et des algorithmes de classification (supervisée ou non). L'utilisation de ces outils mathématiques a plusieurs buts, tels la purification du signal enregistré ou encore l'extraction d'informations pertinentes quant à la différenciation voulue. Un des avantages de cette approche algorithmique de traitement et de classification des spectres, est qu'elle apporte une notion d'automatisation dans le processus de discrimination sain/pathologique.

IV) Analyse numérique des données

1) Prétraitements

En fonction de l'échantillon étudié, du support utilisé ou encore de la calibration des machines, les spectres peuvent être composés de bruit, de décalages en fréquence, de signaux parasites ou peuvent être inexploitables.

Qu'ils soient enregistrés par spectroscopie infrarouge ou Raman, plusieurs prétraitements doivent être appliqués à ces spectres avant qu'ils puissent être considérés dans un processus de classification. Ces prétraitements constituent une étape essentielle (73) puisqu'ils permettent :

- Une augmentation de la robustesse des classifications,
- Une meilleure compréhension des données par l'homme mais également par les machines,
- La suppression des données dites « aberrantes » (ou « outliers »),
- Et dans certains cas, la réduction de dimensionnalité des données, en cherchant à n'extraire de celles-ci que l'information pertinente.

On trouve différentes catégories de prétraitements, chacune ayant un but précis (72, 74) :

a) Suppression des outliers

La détection et la suppression d'outliers dans un ensemble de données brutes, sont réalisées par des tests qualité. Ces tests sont généralement appliqués en amont des prétraitements et permettent de s'affranchir de certaines données présentant des informations très bruitées ou résultant d'un problème matériel lors de leur acquisition. Le calcul d'un rapport « signal sur bruit » pour chaque spectre est un exemple de test qualité communément employé dans ce domaine. Ce test calcule le ratio entre une gamme spectrale connue pour être très informative et une autre gamme connue pour être pauvre en signal. Une valeur minimale de ce ratio permet de ne conserver que les données les plus intenses. D'autres tests s'appuient également sur des valeurs seuils, et permettent de rejeter certaines données si leurs intensités ne sont pas suffisantes en certains nombres d'onde du spectre (test de l'épaisseur de l'échantillon, test d'une bande spécifique, etc).

Il existe des techniques plus complexes comme l'analyse en composantes principales (ACP) (75), dont la fonction première est l'extraction d'informations et la réduction de dimensionnalité, et qui a pour principe de représenter chaque spectre comme une combinaison linéaire de composantes principales pondérées. Le rejet d'un spectre peut alors être réalisé en fonction de ses poids sur certaines composantes principales. L'Extended Multiplicative Signal Correction (EMSC) (76), qui approxime également chaque spectre par une combinaison de signaux (détails dans le chapitre II), est une autre méthode couramment employée en spectroscopie pour détecter et éliminer les spectres aberrants.

b) Suppression des spikes

En spectroscopie Raman, des « spikes » sont susceptibles d'apparaître aléatoirement sur les spectres enregistrés. Ils se caractérisent par une très forte intensité, localisée en un seul nombre d'onde, et ont pour origine l'interaction entre des particules cosmiques de fortes énergies et le détecteur CCD du spectromètre Raman. Des méthodes ont été proposées pour corriger ce phénomène. Elles sont basées sur la comparaison entre un spectre présentant un spike et un de ces spectres voisins (77), sur une approximation linéaire locale des intensités à corriger (78), ou encore sur l'utilisation d'un filtre médian pour corriger ces intensités (79). Cependant, de par la rareté de leur apparition, il est généralement plus rapide (lorsque cela est possible) de recommencer l'acquisition au même point de l'échantillon.

c) Lissage

Le lissage des données est réalisé dans le but de supprimer le bruit de mesure sur les spectres et ainsi augmenter le rapport signal sur bruit. Cependant, le degré de lissage appliqué aux données doit être adapté, car un lissage trop intense aura pour effet de supprimer de l'information. Il existe différents filtres permettant cette correction parmi lesquels on retrouve les filtres passe-bas, les filtres binomiaux ou les filtre moyens coulissant (« sliding average filter »). En spectroscopie vibrationnelle, le filtre de Savitsky-Golay (80) est le plus communément employé puisqu'il permet une bonne réjection du bruit tout en conservant les caractéristiques des spectres telles les extrémums et les largeurs de pics, contrairement aux filtres précédemment cités. Cette méthode, basée sur une estimation polynomiale locale le long d'une fenêtre glissante, est simple à programmer et rapide à exécuter.

d) Correction de ligne de base

Des facteurs instrumentaux ou la nature intrinsèque de l'échantillon peuvent engendrer une ligne de base sur les spectres acquis. Par exemple, en spectroscopie infrarouge, la ligne de base est générée par des effets de diffusion par l'échantillon. En spectroscopie Raman, la ligne de base trouve principalement naissance dans la fluorescence générée par l'échantillon. Cette ligne de base se traduit généralement sous la forme d'une courbe à variation lente dans les spectres. Ceux-ci sont donc corrigés en soustrayant à chacun d'eux une approximation de sa propre ligne de base. Cette approximation peut être représentée par une droite, une combinaison de segments de droites ou une courbe, et est déterminée manuellement ou automatiquement selon différentes méthodes.

Les estimations par fonction polynomiale (recherche d'un polynôme de degré choisi afin de « coller » au mieux à chaque spectre) (81-84), ou par la technique des ondelettes (représentant un signal comme une somme pondérée d'approximations qui modélisent le fond de fluorescence, et de détails qui approximent le signal d'intérêt et le bruit) (85, 86), sont les plus communément employées. Cependant, il en existe beaucoup d'autres souvent adaptées au type de données auxquelles elles sont appliquées (87-89).

e) Normalisation

La normalisation est l'un des prétraitements les plus importants en spectroscopie vibrationnelle. Elle permet de s'affranchir des variations d'intensités dues par exemple aux différences d'épaisseurs de l'échantillon ou aux variations de focalisation de l'objectif. La normalisation est donc utilisée dans un but d'homogénéisation en ramenant les spectres à une échelle commune. Différentes fonctions peuvent être utilisées. La normalisation « Min-Max » amène le minimum du spectre à 0 par translation, puis le maximum à 1 par contraction/dilatation. La normalisation « 1-norm » centre le spectre puis divise chaque intensité par la somme des valeurs absolues de toutes les intensités de façon à ce que cette somme soit égale à 1. La normalisation « vectorielle » (ou « 2-norm ») centre le spectre, puis divise chaque intensité par la racine carrée de la somme du carré de toutes les intensités de façon à ce que cette somme soit égale à 1. La fonction « Standard Normal Variate » (90) centre le spectre puis divise chaque intensité par l'écart-type de toutes les intensités, de façon à ce que le spectre normalisé est une moyenne égale à 0 et un écart-type égale à 1.

f) Correction de signaux contaminants

L'extraction de signaux dits « contaminants » peut être nécessaire selon que le milieu dans lequel est placé l'échantillon ou le support utilisé contribuent ou non au signal enregistré. Le principe consiste à estimer la proportion d'un signal de référence (le spectre du contaminant pur) dans les spectres enregistrés. Ce spectre contaminant est ensuite soustrait du spectre de donnée à hauteur de sa contribution. Cette correction est rapide, efficace, et est réalisée pour chaque spectre indépendamment des autres. Différentes méthodes existent comme l'estimation par moindres carrés classiques ou contraints à la non négativité.(91), ou l'estimation par optimisation non linéaire (92).

Par exemple, la méthode Nonnegatively Constrained Least Squares (NCLS) a été appliquée en spectroscopie Raman pour réaliser un déparaffinage numérique des spectres en estimant la contribution des composantes spectrales de la paraffine dans les spectres enregistrés sur des coupes paraffinées de naevi et de mélanomes (93).

g) Extended Multiplicative Signal Correction (EMSC)

En spectroscopie infrarouge, la méthode « Extended Multiplicative Signal Correction » (76), qui est une extension de la méthode « Multiplicative Signal Correction » (94), vise à minimiser l'influence de facteurs contaminants dans un signal. Son principe repose sur la décomposition des données en une combinaison linéaire de signaux pondérés. Chaque spectre est alors représenté par la somme d'un spectre représentatif de l'échantillon (en pratique c'est le spectre moyen), de plusieurs spectres de contaminants purs, d'une fonction polynomiale et d'un résidu. Les contributions estimées des signaux contaminants peuvent alors être retirés. De même, la ligne de base est corrigée pour chaque spectre *via* son estimation par la fonction polynomiale. Des règles d'exclusion peuvent également être déterminées par l'utilisateur en considérant les poids de ces signaux. Les spectres de données aberrantes (ou outliers) peuvent alors être reconnus comme ceux présentant un faible poids sur le spectre moyen et un fort résidu. Enfin, une normalisation des spectres est réalisée.

De par sa capacité à réaliser simultanément l'ensemble de ces corrections aux données, l'EMSC apparaît donc de fort intérêt en spectroscopie vibrationnelle et est de ce fait

communément utilisée dans la littérature (95, 96). En effet, on trouve des applications de l'EMSC à des données acquises par spectroscopie vibrationnelle sur des échantillons de cancer du colon (97), des gliomes (98), des cancers cutanés (99) ou encore sur des cellules leucémiques (100). L'EMSC est plus amplement détaillée dans le chapitre « Matériels et Méthodes » de ce manuscrit.

h) Sélection de variables

Lorsqu'un jeu de données est composé de plusieurs centaines ou milliers de variables, toutes ces variables ne sont pas utiles pour répondre à une question précise de prédiction. Seul un sous-ensemble de celles-ci explique le phénomène étudié. Leur identification est donc primordiale pour proposer un modèle optimal de prédiction du phénomène. Les méthodes de sélection de variables (101-103) sont des techniques supervisées qui ont pour objectif de réduire les données à leurs dimensions les plus pertinentes afin de répondre au problème posé, sans les transformer. Différentes méthodes de sélection de variables existent, empiriques ou automatiques, et parmi lesquelles on trouve (101) :

- La sélection manuelle, basée sur des connaissances *a priori* par l'utilisateur des dimensions potentiellement explicatives du phénomène étudié. En spectroscopie vibrationnelle, un exemple simple et communément employé de sélection de variables est la réduction des spectres à une gamme spectrale, connue pour être la plus informative quant à un type d'échantillon étudié. Couramment, en spectroscopie infrarouge, la région 900-1800 cm^{-1}, appelée région de l'empreinte digitale, est sélectionnée puisqu'elle est réputée pour être la plus informative pour l'analyse de tissus (104).

- Les filtres, basés sur les caractéristiques intrinsèques aux données. Ces méthodes calculent généralement un score de pertinence par variable. Puis les variables sont ordonnées sur ce score, de la plus pertinente à la moins pertinente. Seules les variables les plus pertinentes sont conservées et utilisées pour entraîner un classifieur supervisé. Ces méthodes sont réputées simples, rapides et indépendantes du classifieur utilisé. Cette dernière caractéristique peut cependant être vue comme un inconvénient car les tâches de sélection et de classification sont dissociées. Un

autre inconvénient de ces méthodes est qu'elles sont souvent basées sur des tests statistiques univariés, c'est-à-dire que chaque variable est considérée indépendamment des autres. Le test de Student et le test du χ^2, sont des exemples de ces filtres. Pour pallier le problème de l'univariance liée à ces tests, des méthodes multivariées ont été introduites pour prendre en compte la corrélation entre les variables. Parmi ces méthodes, la sélection de variable basée sur la corrélation (Correlation-based Feature Selection ou Fast Correlation-Based Feature Selection) et les filtres de Markov (Markov Blanket filter) sont les plus utilisés.

- Les méthodes englobantes ou « Wrapper », basées sur l'évaluation de sous-ensembles de variables par une méthode de classification choisie. Un avantage de ce type de méthode est que la sélection des variables est intimement liée à la méthode de classification choisie. Un autre avantage est l'aspect multivarié de ces méthodes, permettant ainsi de tenir compte de la dépendance entre les variables. Les inconvénients de ces techniques par rapport aux filtres sont leur plus grande exposition au surapprentissage et leur lourdeur informatique. Ces méthodes se divisent en deux catégories, à savoir les méthodes déterministes (Sequential Forward Selection, Sequential Backward Elimination) et les méthodes aléatoires (recuit simulé, algorithmes génétiques).

- Les méthodes embarquées ou « Embedded » construisent simultanément les sous-ensembles de variables et le classifieur. Les avantages sont les mêmes que pour les méthodes englobantes mais leur complexité informatique est beaucoup moins grande. Des exemples de ces techniques sont les arbres de décision ou les classifieurs naïfs Bayésiens pondérés.

Les méthodes de sélection automatique de variables ont été largement appliquées en spectroscopie vibrationnelle.

Une méthode de sélection de variable par classement et basée sur une mesure de covariance a été appliquée en spectroscopie infrarouge pour sélectionner 81 nombres d'onde spécifiques de la maladie de la vache folle (105). Dans (106), des tests de Student et

de Mann-Whitney ont été appliqués à des spectres infrarouge pour déterminer les pics spécifiques de carcinomes thyroïdiens. Ces pics discriminants ont ensuite été considérés en entrée d'une analyse discriminante canonique pour proposer un classifieur différenciant goitre et carcinome. Dans (99), un test de Kruskal-Wallis a permis de déterminer les nombres d'onde spécifiques au diagnostic de tumeurs cutanées.

Des algorithmes génétiques ont également été couplés à une analyse discriminante linéaire (Linear Discriminant Analysis, LDA) afin de sélectionner les nombres d'onde spécifiques de la maladie de la vache folle (107) ou de lésions cutanées (99). En spectroscopie Raman, ils ont été couplés à la PLS-DA (Partial Least Square, PLS – Discriminant Analysis, DA) pour proposer un modèle de diagnostic *in vivo* du précancer du col de l'utérus (108) ou à la LDA pour estimer un modèle diagnostic du syndrome de Barrett dans l'œsophage (109).

i) Extraction de variables

Comme pour la sélection de variables, les méthodes d'extraction de variables ont pour but de réduire la dimensionnalité des données et de mettre en avant l'information pertinente à une classification donnée. Cependant, les méthodes d'extraction de variables cherchent à combiner les variables des objets, de façon linéaire ou non, afin de créer de nouvelles variables plus discriminantes, appelées caractéristiques et représentées dans un espace de dimensions différentes. On parle alors d'espace des caractéristiques, ou « feature space » en anglais. Différentes méthodes d'extraction de variables existent (103, 110) :

- Les méthodes non-supervisées : elles ont pour but de réaliser la meilleure reconstruction possible des données dans l'espace des caractéristiques, sans tenir compte des classes d'appartenance des objets. Principalement, deux grandes catégories d'algorithmes sont utilisées en pratique :
    - Les algorithmes de classification non-supervisée qui vont chercher à grouper les individus les plus similaires dans une même classe, et les individus les plus différents dans des classes distinctes. Chaque classe est représentée par un prototype ou centroïde qui est classiquement calculé comme la moyenne des objets appartenant à la classe. Chaque centroïde devient alors une caractéristique. Les algorithmes K-Means et HCA (Hierarchical Clustering Analysis) sont généralement privilégiés pour réaliser cette tâche.

- Les factorisations matricielles qui vont chercher à estimer une sous-structure des données. Généralement, un modèle linéaire est supposé, c'est-à-dire que tout objet peut être considéré comme la somme pondérée de plusieurs composantes sous-jacentes aux données et qui permettent de mieux expliquer, représenter les données. Une fois les composantes estimées, chaque objet est projeté sur ces composantes afin d'obtenir ses nouvelles coordonnées dans ce nouvel espace. Parmi ces méthodes, les plus connues pour estimer ces composantes sous-jacentes sont l'ACP (111), l'analyse en composantes indépendantes (112), la factorisation en matrice non-négative (113), la transformation en ondelettes (114).

- Les méthodes supervisées : en s'appuyant sur l'appartenance connue des objets d'un jeu d'entraînement, ces méthodes vont rechercher une nouvelle représentation des données de façon à optimiser la séparation entre les classes. Parmi les méthodes linéaires, la LDA (115) et les moindres carrés partiels (PLS) (116) sont les plus utilisées en spectroscopies vibrationnelles. Parmi les méthodes non-linéaires, on peut citer la méthode Maximum Representation and Discrimination Feature (MRDF) (117).

Dans la littérature, l'ACP a permis d'extraire les composantes nécessaires à l'amélioration du diagnostic par spectroscopie infrarouge de cancer du sang périphérique (118) ou cutanés (99), et par spectroscopie Raman de lésions colorectales (119) et cutanées (63).

Dans (120), une décomposition en ondelettes de spectres infrarouge a été réalisée afin d'améliorer les résultats de prédiction d'un réseau de neurones artificiels construit pour différencier colon normal, adénocarcinomes et carcinomes.

La méthode Maximum Representation and Discrimination Feature (MRDF) suivi d'une classification probabiliste Sparse Multinomial Logistic Regression (SMLR) a largement été appliquée en spectroscopie Raman pour extraire des caractéristiques diagnostiques spécifiques de tumeurs du cerveau (121) ou du cancer du col de l'utérus (122).

Bien que d'autres méthodes de corrections de données ou de sélection/extraction d'informations pertinentes peuvent être nécessaires à des études particulières, ces prétraitements constituent une étape primordiale à tout processus de classification et sont les plus communément employés en spectroscopie vibrationnelle (72).

2) Classifications

a) Généralités

Les algorithmes de classification sont des outils mathématiques généralement utilisés dans le but de faciliter la gestion ou l'analyse de données de grandes quantités, pouvant présenter de grandes similarités et/ou de grandes dimensions. Ces algorithmes sont nombreux et de différentes natures (123).

Leur principe repose sur la distinction de groupes d'objets en différentes classes selon des critères de proximité entre eux, calculés différemment selon l'algorithme considéré.

La quantité de données à classer, le type, la dimension, ou encore l'homogénéité sont autant d'éléments à considérer afin de choisir l'algorithme de classification le mieux adapté.

On distingue deux catégories d'algorithmes de classification, supervisée et non supervisée.

b) Algorithmes de classification non supervisée

Un algorithme de classification non supervisée est basé sur le regroupement d'individus sans considérer d'autres informations que celles propres aux individus eux-mêmes. En d'autres termes, si l'appartenance des données à une classe A ou B est *a priori* connue, cette information n'interviendra pas dans le processus de classification d'un algorithme de classification non supervisée. Ainsi, les résultats de classification obtenus sont dépendants des individus donnés en entrée. Il existe différents algorithmes de classification non supervisée :

Les algorithmes hiérarchiques ascendants (124) sont itératifs et nécessitent le calcul de distances entre les objets deux à deux. Au fur et à mesure des itérations successives, un objet ou groupe d'objet est rattaché à un autre objet ou groupe d'objets. Lorsque tous les objets sont réunis en un même groupe, le résultat se présente sous la forme d'un arbre hiérarchique dont les nœuds représentent les liaisons successives. L'ordre et la hauteur de ces nœuds dans l'arbre déterminent les proximités entre les objets.

Les algorithmes hiérarchiques descendant (125), sont basés sur une dichotomie successive de l'ensemble des données. Lors d'un processus itératif, chaque sous-ensemble de données créé à partir de l'ensemble initial est subdivisé jusqu'à obtenir un objet par sous ensemble.

De même que pour la classification hiérarchique ascendante, les résultats sont visualisables sous la forme d'un arbre hiérarchique.

Les algorithmes réalisant des partitions, ne sont pas basés sur le principe d'agrégation itératif. Les K-means (126), et les Fuzzy C-Means (FCM) (127), sont des algorithmes dits « de centres mobiles » où les individus à classer sont rattachés à des centroïdes (données représentatives de chaque classe). Ces rattachements sont mis à jour au cours d'un processus itératif jusqu'à convergence d'une fonction objectif calculée à chaque étape.

Les algorithmes hiérarchiques, les K-Means et les FCM sont les méthodes de classification non supervisée les plus largement appliquées en spectroscopie vibrationnelle. Par exemple, les HCA ont permis de différencier des lésions cutanées (128), les K-means ont été appliqués à la caractérisation de cancer de la vessie ou de la peau (129, 130), les FCM ont permis de caractériser des cancers cutanés (131) ou du foie (132). Les FCM sont décrits plus en détails dans le chapitre « Matériels et Méthodes » de ce manuscrit.

c) Algorithmes de classification supervisée

Dans le cas d'une classification supervisée, un modèle de prédiction est construit en amont de toute classification. Dans ce but, un jeu d'individus étiquetés et appartenant au moins à 2 classes différentes est nécessaire. Un tel jeu de données, où les objets y figurant sont représentatifs des classes auxquelles ils appartiennent, est nommé « jeu d'entrainement ». À partir de ce jeu, un modèle de prédiction est calculé, différemment selon l'algorithme choisi, et permet l'étiquetage de tout nouvel objet non labellisé. Le label attribué correspond à celui de la classe du modèle la plus proche de l'individu.

La détermination du modèle de prédiction, par le choix du jeu d'entrainement et du paramétrage de l'algorithme utilisé, est donc primordiale. En pratique, plusieurs modèles sont calculés et chacun est confronté à un jeu de données externe (différent du jeu d'entrainement) nommé « jeu de validation ». Après comparaison entre les labels réels et ceux attribués par chaque modèle, des valeurs de sensibilité et spécificité sont calculées. Le modèle présentant les meilleurs résultats est retenu. Un test supplémentaire est parfois réalisé pour à la fois confirmer la véracité du modèle retenu et pour s'assurer que celui-ci n'est pas spécifique aux données du jeu de validation. Pour cela on teste un troisième jeu de

données dit « jeu de test ». De la même manière que pour le jeu de validation, les prédictions sont vérifiées, les résultats sont calculés, et leur analyse confirme ou non les premières tendances obtenues.

Il est à noter que les étiquetages sont réalisés indépendamment d'un objet à l'autre. Ainsi, le résultat d'une classification supervisée ne dépend pas de l'ensemble des objets à classer sinon du modèle de prédiction utilisé. Il existe beaucoup d'algorithmes de classification supervisée. Les réseaux de neurones artificiels (ANN : Artificial Neural Network)(133) calculent un modèle de prédiction en suivant un schéma inspiré du réseau neuronal biologique. Les K plus proches voisins (KNN : K Nearest Neighbor) (134) ne calculent pas de modèle de prédiction mais attribuent à un nouvel individu le label majoritaire parmi les K individus les plus proches de lui. La LDA (115) recherche une combinaison linéaire des variables décrivant les individus, et conduisant à une discrimination optimale des classes. La LDA est très souvent précédée d'une ACP. Les machines à vecteurs supports (Support Vector Machines, SVM) (135, 136) calculent un modèle de prédiction à partir des individus les plus à la frontière des classes à représenter.

Dans la littérature, on trouve différentes études couplant ces algorithmes à des données acquises par spectroscopies vibrationnelles et appliquées au domaine du biomédical et plus précisément à la détection de cancers. Parmi ces applications on retrouve la prédiction de spectres Raman par les ANN appliqués à la détection de mélanome (137), les KNN appliqués au diagnostic du cancer de la prostate (138), la LDA pour le diagnostic de cancer du larynx (139), les SVMs pour la détection de cancers gastriques (140) ou de métastases dans le cerveau (141). Une étude dédiée au diagnostic du cancer du sein par spectroscopie compare les résultats obtenus par plusieurs de ces algorithmes de classification à savoir ANN, KNN, LDA et SVM (142).

V) Objectifs de thèse

Le but de ces travaux de thèse est d'améliorer et d'apporter de nouveaux outils d'aide au diagnostic de cancers par l'utilisation des spectroscopies vibrationnelles infrarouge et Raman, en les couplant à des algorithmes de classification (supervisée ou non).

Deux axes de recherche ont été menés en parallèle dans cette thèse. Le premier relève du développement algorithmique en classification non supervisée pour une application en imagerie infrarouge de cancers cutanés. Le second concerne l'application d'outils numériques à des données Raman, pour le diagnostic de la LLC à l'échelle cellulaire.

1) Aide au diagnostic de cancers cutanés par spectroscopie infrarouge

Au sein du laboratoire, des études ont montré le potentiel discriminatoire de la spectroscopie infrarouge pour diagnostiquer les cancers cutanés lorsque cette technologie est couplée à des algorithmes de classification non supervisée tels K-means (99, 143) et FCM (144). L'algorithme FCM a montré des résultats comparables à ceux obtenus par K-means à la différence qu'il apporte des informations supplémentaires, comme les zones de transition entre différentes classes par exemple.

Néanmoins, l'algorithme FCM nécessite la gestion de deux paramètres, le nombre de classes C et le paramètre de flou m. Le choix de ces paramètres reste empirique et constitue une limitation à son utilisation. Un algorithme basé sur les FCM et nommé RBA (Redundant Based Algorithm) (145) a donc été développé pour choisir automatiquement ces deux paramètres de façon optimale. Cet algorithme étant basé sur la recherche de redondance entre les classes estimées, un paramétrage de C et m induisant l'estimation de classes non-redondantes sera considéré comme optimal. Ce nouvel algorithme, appliqué à des échantillons de cancers cutanés, est capable de faire la distinction entre les différentes structures tissulaires des échantillons et met en avant les zones de transitions entre ces structures. Cependant, l'algorithme RBA a été spécifiquement développé pour une application à des données spectrales infrarouge, et de par sa structure interne, nécessite un temps de calcul conséquent.

Le premier objectif de ces travaux de thèse concernait deux améliorations algorithmiques de cette technique. La première visait la réduction du temps de calcul en interrompant

prématurément les FCM au cours de leurs itérations lorsque des classes redondantes sont estimées. De plus, une limitation du nombre de couples (C, m) testés a été apportée. La deuxième amélioration avait pour but de généraliser l'utilisation de l'algorithme RBA à d'autres types de données. Pour cela, une nouvelle mesure de similarité a été incorporée.

2) Aide au diagnostic de la LLC par spectroscopie Raman

Le second axe de recherche de ces travaux a été mené dans le cadre d'un projet ANR-TECSAN IHMO visant à développer une solution d'imagerie hybride combinant en un même instrument, microscopie multispectrale et microspectroscopie Raman. L'objectif de ce projet est à terme de compléter l'analyse multidisciplinaire classique utilisée jusqu'à présent pour le diagnostic et l'établissement de facteurs pronostiques d'une tumeur sur un patient.

Dans ce projet, la LLC a été retenue comme pathologie d'étude en raison de plusieurs arguments :

- L'accès facile et en grande quantité au matériel tumoral, en particulier aux cellules lymphoïdes sanguines de patients présentant une LLC.
- La disponibilité du matériel cryopréservé suivant les recommandations médicales habituelles.
- L'expertise médicale et non médicale dans le domaine des syndromes lymphoprolifératifs chroniques, avec la possibilité d'accéder sur place aux différentes techniques d'explorations biologiques.
- Le travail d'équipe réalisé dans ce domaine depuis plusieurs années.

Dans le but d'une valorisation clinique, l'étude est réalisée sur des frottis sanguins étalés sur lames de verre, c'est-à-dire dans des conditions très proches de celles pratiquées en routine.

Le but de l'étude est donc de mettre en place un nouvel outil autonome d'aide au diagnostic de la LLC, *i.e.* capable de distinguer un échantillon sain d'un échantillon pathologique. Cet outil nécessite au préalable le choix de traitements numériques adaptés aux données Raman enregistrées, et la mise en place d'une procédure de classification automatique.

La construction et l'adaptation de cet outil pour son intégration au module développé dans le cadre du projet IHMO ont fait l'objet du second axe de recherche de cette thèse.

Chapitre 2 : Matériels et Méthodes

Dans ce chapitre sont détaillées les caractéristiques et configurations du matériel utilisé pour ces études. Les algorithmes de traitement et de classification des spectres IR et Raman enregistrés y sont également décrits plus en détails.

I) Préparation des échantillons

1) Coupes tissulaires

Pour l'analyse infrarouge, des coupes de 10 µm d'épaisseur sont réalisées avec un microtome et sont déposées sur lames de fluorure de calcium (CaF_2) sans aucun agent fixateur, déparaffinage chimique ou coloration. Seule une goutte d'eau distillée est déposée entre la lame et la coupe afin de permettre une meilleure adhésion. Des supports (ou fenêtres) de CaF_2 ont été choisis car ils présentent l'avantage de ne pas absorber dans l'infrarouge et permettent ainsi de ne pas apporter de signal contaminant dans les spectres enregistrés. Enfin, la fenêtre est déposée sur une plaque chauffante afin de faire évaporer l'eau (qui présente un signal en infrarouge) et de liquéfier la paraffine, permettant ainsi de détendre la coupe et d'éviter les plis éventuels.

En parallèle, des coupes adjacentes à celles utilisées pour l'analyse infrarouge sont préparées de façon conventionnelle, avec une coloration H&E, afin d'obtenir un diagnostic dit « de référence » par un anatomopathologiste expérimenté. Par la suite, ce diagnostic sera confronté aux résultats d'histologie spectrale infrarouge. Pour cette étude, un jeu de 16 coupes tissulaires incluses en paraffine est considéré. Ce jeu est composé de 4 coupes de BCC, 3 coupes de SCC, 4 coupes de maladie de Bowen et 5 coupes de mélanomes.

2) Frottis sanguins

Concernant les analyses par spectroscopie Raman, les frottis sanguins sont obtenus en étalant une goutte de sang sur une lame de verre standard qui est ensuite séchée à température ambiante pendant quelques minutes. Aucun traitement du sang en amont ni de fixation des cellules sur le support ne sont réalisés, ce qui rend la préparation de ces échantillons simple, rapide et commune aux préparations réalisées en routine clinique. L'unique différence entre ces préparations est la nécessité en clinique de colorer les lames, en général par une coloration MGG, car l'observation au microscope est une étape nécessaire lors de l'établissement du diagnostic de la LLC par un hématologue expérimenté.

Pour cette étude, les échantillons sanguins de 27 volontaires et 49 patients atteints d'une LLC de type B hypercytaire (pourcentage de lymphocytes B supérieur à 80%) ont été considérés.

II) Instrumentation

1) Microspectroscopie infrarouge à transformée de Fourier

L'enregistrement des données IR est effectué *via* l'imageur Spotlight 300 (Perkin Elmer, Courtabœuf, France) couplé à un spectromètre infrarouge à transformée de Fourier, Spectrum One (Perkin Elmer, Courtabœuf, France) (Figure 16). Lors d'une acquisition, l'échantillon est traversé en différents points (suivant un quadrillage (X, Y) régulier) par la source lumineuse excitatrice de type Globar, focalisée *via* des objectifs Cassegrain.

La résolution spatiale choisie pour cette étude est de 6,25µm, se traduisant par une taille des pixels de l'image spectrale finale de 6,25 µm x 6,25 µm. Les spectres infrarouge sont enregistrés sur la gamme spectrale 400-4000 cm^{-1} avec une résolution spectrale de 4 cm^{-1} entre deux nombres d'onde consécutifs, et une accumulation de 16 scans pour chaque spectre. Un interféromètre de Michelson permet la mesure, en différents nombres d'onde, des intensités transmises ou absorbées. Une transformée de Fourier de l'interférogramme ainsi obtenu est réalisée *via* un ordinateur dédié.

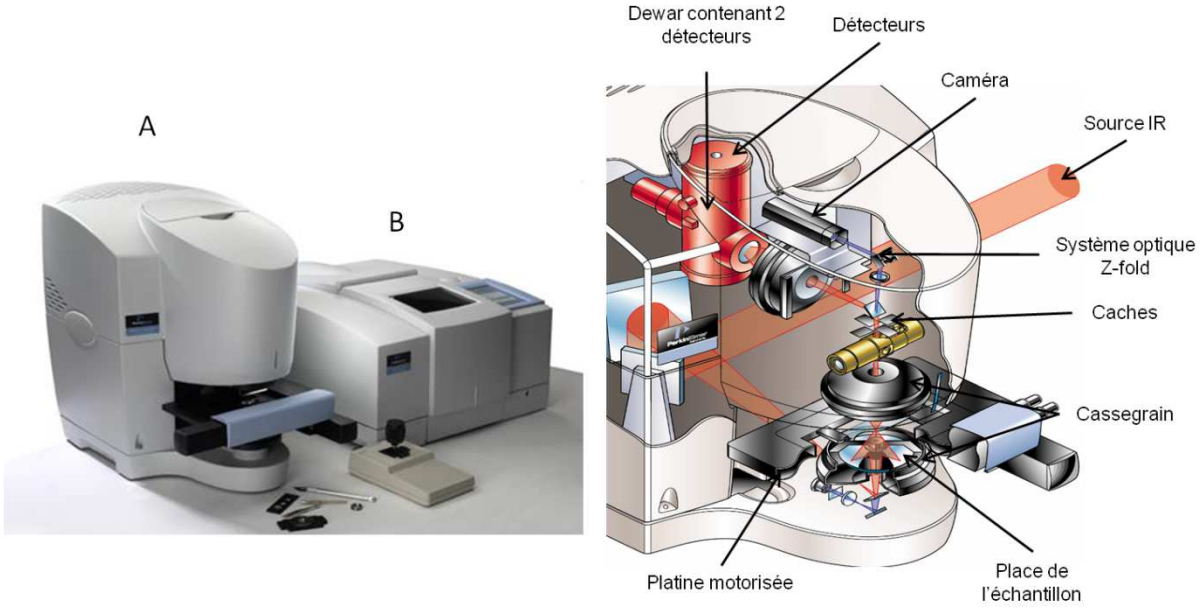


Figure 16 : Imageur Spotlight IRTF

Micro-imageur composé de (A) le système d'imagerie (Microscope Spotlight 300, Perkin Elmer) associé à (B) un spectromètre infrarouge à transformée de Fourier (Spectrum One, Perkin Elmer).

Les données enregistrées constituent une image spectrale infrarouge se présentant sous la forme d'un cube d'intensités (Figure 17), avec en *X* et *Y* les coordonnées géographiques des différents spectres enregistrés sur l'échantillon, et en Λ l'ensemble des nombres d'onde de ces spectres (à chaque pixel correspond un spectre infrarouge). Une image représentant les intensités d'absorbance (ou de transmittance) de tous les pixels en un nombre d'onde donné, peut être visualisée en chaque tranche du cube suivant l'axe Λ.

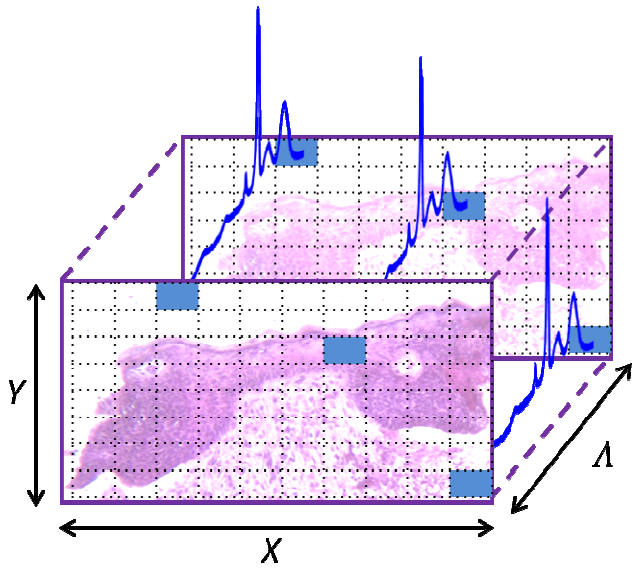


Figure 17 : Cube de données spectrales

Cube de données infrarouge représentant en X et en Y les coordonnées géographiques des spectres, et leurs intensités suivant les différents nombres d'onde en .

2) Microspectroscopie Raman

Une machine multimodale (Figure 18) a été développée par la société Tribvn (Châtillon, France), combinant un spectromètre Raman (Horiba Jobin Yvon, Villeneuve d'Ascq, France), un microscope conventionnel (ECLIPSE FN1 Nikon SA, Champigny Sur Marne, France) équipé de deux objectifs à sec (Nikon SA, Champigny Sur Marne, France), de grossissement 40x et 100x (ouverture numérique de 0,6 et 0,9 respectivement) et une plateforme motorisée en X, Y (Ludl Electronic Products, Ltd). L'éclairage de l'échantillon est réalisé par le bas de l'instrumentation *via* une source lumineuse à led blanche (Schott, KL 1500 Led), traversant une roue à filtres polychromatique (x10 couleurs) (Figure 19). L'image visible de l'échantillon peut être observée *via* deux caméras couleur ou noir & blanc (Stemmer Imaging, Suresnes, France).

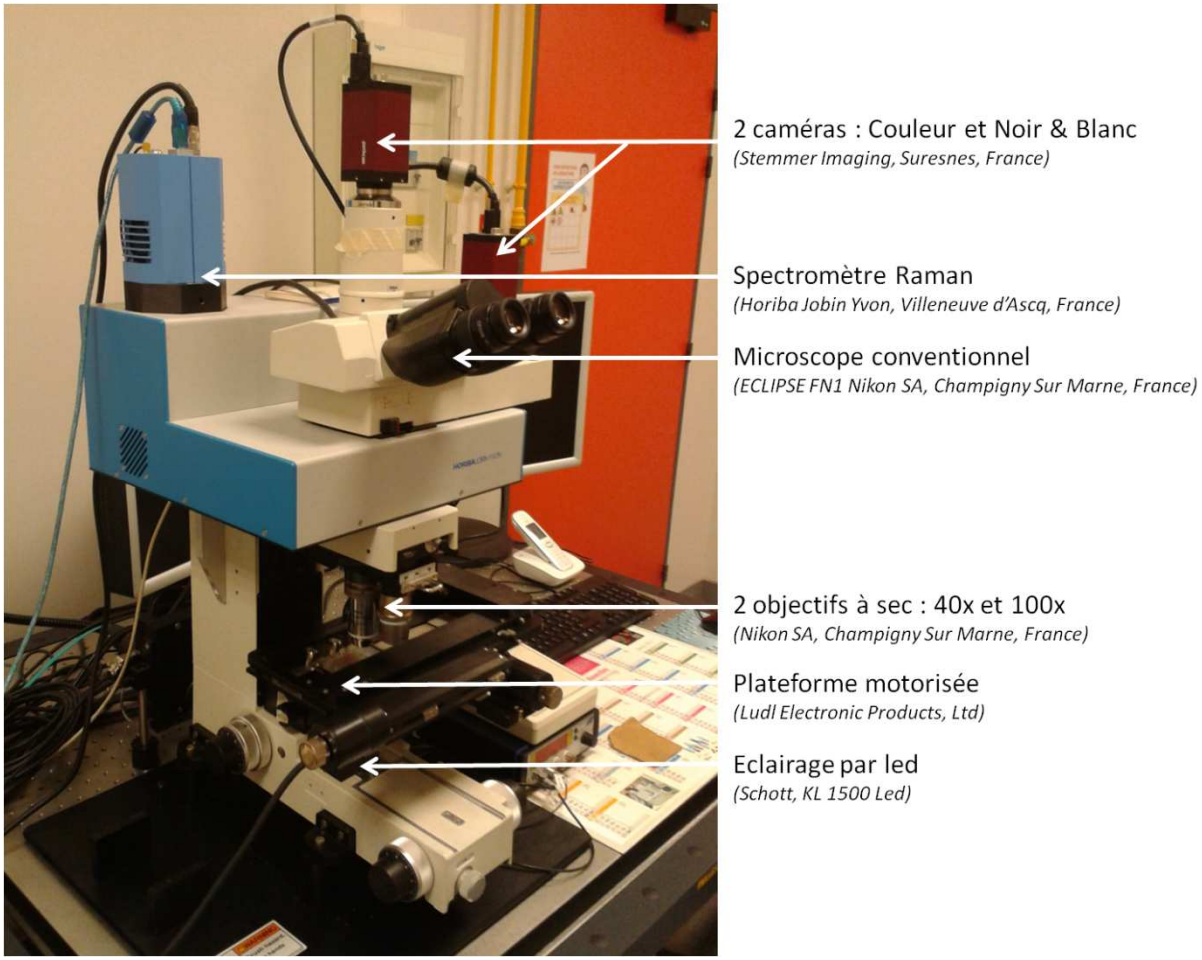


Figure 18 : Module Raman

Plateforme d'acquisition de spectres Raman IHMO composée d'un microscope conventionnel, d'un spectromètre Raman (diode laser 532nm), d'un éclairage par led et d'une plateforme motorisée.

Le spectromètre Raman est équipé d'une source laser excitatrice de longueur d'onde 532 nm (Toptica Photonics, Munich, Germany), délivrant une puissance laser de 22 mW en sortie et 13,5 mW sur l'échantillon, d'un trou confocal de 100 µm, apportant une résolution spatiale de 1 µm en (X, Y) et de 2 µm en Z, d'un réseau de dispersion 1200 traits par millimètre et d'un détecteur CCD de 1024x256 pixels (Technologies Andor, South Windsor, CT, Etats-Unis) refroidi à -70°C par effet Pelletier. Les spectres Raman sont enregistrés sur la gamme spectrale 703-3187 cm^{-1} avec une résolution spectrale de 4 cm^{-1}, une résolution spatiale de 2 µm et une accumulation de 2 fois 10 secondes pour chaque spectre.

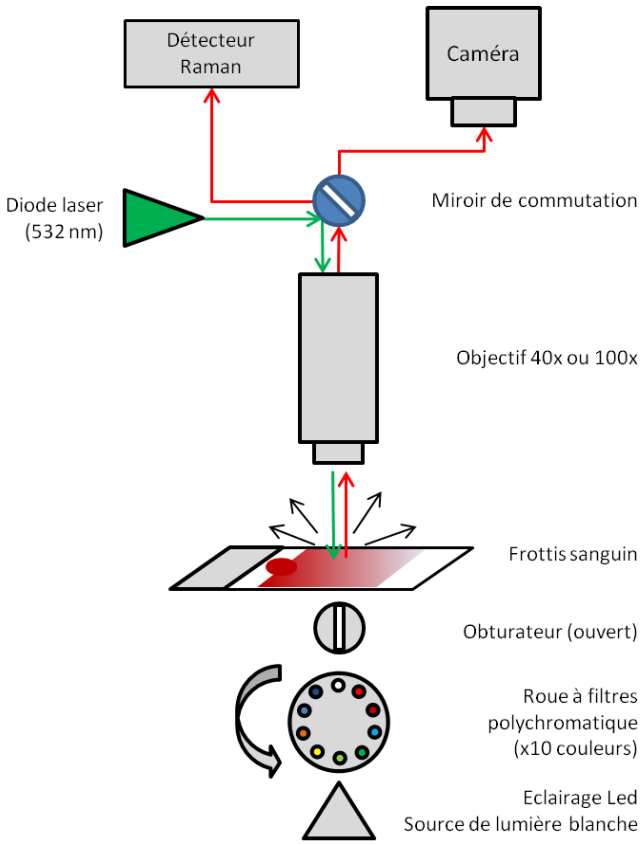


Figure 19 : Processus d'acquisition Raman
Schéma descriptif du processus d'acquisition d'un spectre Raman sur un frottis sanguin non coloré

Dans cette étude le protocole d'acquisition des spectres Raman consiste dans un premier temps à scanner (par photographie) une zone du frottis sanguin à l'aide de l'objectif 40x. Les leucocytes sont automatiquement localisés et dénombrés sur cette zone par une technique de traitement d'image permettant la détection d'objet. Chacune de ces cellules est ainsi géographiquement localisée sur le frottis (coordonnées en X et Y), autorisant un retour sur leur position après coloration du frottis. Dans un second temps, les acquisitions de spectres Raman sont réalisées sur le noyau des leucocytes préalablement détectés, à l'aide de l'objectif 100x.

III) Prétraitements des spectres

1) Prétraitements des spectres infrarouge

Les coupes tissulaires considérées dans cette étude sont incluses en paraffine. Lors du processus d'acquisition des spectres infrarouge, la paraffine apporte une contribution au signal enregistré, contribution qu'il faut éliminer ou à défaut neutraliser. Pour ce faire, une méthode de déparaffinage numérique est utilisée, l'Extended Multiplicative Signal Correction (EMSC) (146). Cette méthode est à l'origine utilisée en spectroscopie infrarouge pour séparer les effets physiques de la dispersion de la lumière des effets chimiques induisant l'absorption de la lumière (76). Récemment, cette technique a été adaptée pour neutraliser la contribution de la paraffine dans les spectres enregistrés(147, 148).

Tout spectre infrarouge $\mathbf{s_i}$ composé de n nombres d'onde v_n, peut être écrit sous la forme :

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

Avec :

$\hat{\mathbf{s}}$	$\in \mathbb{R}^{1\times n}$	un spectre de n nombres d'onde, représentatif des spectres de l'échantillon (spectre moyen en général)
$\mathbf{I}$	$\in \mathbb{R}^{k\times n}$	la matrice des k spectres représentatifs de chaque contaminant à retirer des données
$\mathbf{P} = \begin{pmatrix} v_1^0 & \cdots & v_1^p \\ \vdots & \ddots & \vdots \\ v_n^0 & \cdots & v_n^p \end{pmatrix}$	$\in \mathbb{R}^{(p+1)\times n}$	la transposée de la matrice de Vandermonde des n nombres d'onde v_n. Elle est utilisée ici pour calculer $\mathbf{c_i P}$, une fonction polynomiale d'ordre p représentant la déformation de la ligne de base des spectres enregistrés
$\mathbf{e_i}$	$\in \mathbb{R}^{1\times n}$	un vecteur d'erreur du model EMSC
a_i	scalaire	un coefficient d'approximation de $\mathbf{s_i}$ par $\hat{\mathbf{s}}$
$\mathbf{b_i}$	$\in \mathbb{R}^{1\times k}$	un vecteur de coefficients d'approximation de $\mathbf{s_i}$ par $\mathbf{I}$
$\mathbf{c_i}$	$\in \mathbb{R}^{1\times(p+1)}$	un vecteur de coefficients d'approximation de $\mathbf{s_i}$ par $\mathbf{P}$

Pour construire la matrice des contaminants $\mathbf{I}$, un spectre moyen de plusieurs spectres enregistrés sur la paraffine pure n'est pas suffisant. En effet, en raison de leur hétérogénéité, il est nécessaire de réaliser en premier lieu une analyse en composantes principales sur ces spectres. On notera cependant qu'un test qualité est réalisé sur les spectres de paraffine en amont de l'analyse en composantes principales. Ce test, basé sur des rapports d'amplitude de pics, permet d'écarter les spectres trop bruités ou ne présentant pas un signal de paraffine bien défini. Le spectre moyen ainsi que les composantes principales les plus représentatives de la paraffine (ayant les plus grandes valeurs propres donc expliquant la plus grande variabilité du signal de la paraffine) constituent la matrice des signaux contaminants $\mathbf{I}$.

Les trois coefficients, a_i, b_i et c_i sont estimés par la méthode des moindres carrés dans le but de minimiser la somme des erreurs calculées sur chaque spectre $\sum_{j=1}^{n} \mathbf{e_{ij}}^2$, avec n le nombre de données à corriger. Les spectres corrigés de la contribution des contaminants et de la déformation de leur ligne de base peuvent alors s'écrire sous la forme :

$$\mathbf{s_{icorr}} = a_i \hat{\mathbf{s}} + \mathbf{e_i} \tag{10}$$

Afin d'analyser les différences biochimiques existantes entre les différents pixels, et non les différences d'épaisseur de la coupe tissulaire, les spectres sont normalisés autour du spectre moyen $\hat{\mathbf{s}}$ selon l'équation :

$$\mathbf{s_{icorr}^N} = \hat{\mathbf{s}} + \frac{\mathbf{e_i}}{a_i} \tag{11}$$

À la suite de la correction par EMSC, les spectres infrarouge peuvent entrer dans un processus de classification.

2) Prétraitements des spectres Raman

Les spectres Raman étant enregistrés avec une résolution spectrale de 4 cm^{-1}, la première étape de prétraitement consiste à les interpoler linéairement afin d'obtenir un pas unitaire entre deux nombres d'onde consécutifs.

Comme mentionné précédemment, certains signaux parasites peuvent contribuer aux signaux Raman enregistrés sur l'échantillon. Dans cette étude, les lames de verre sur lesquelles les frottis sanguins sont étalés possèdent un signal propre qui pollue les spectres

enregistrés. Un deuxième prétraitement consiste donc à estimer la proportion d'un spectre de verre pur dans chacun des spectres Raman enregistrés. Cette proportion est estimée par moindres carrés puis soustraite de chaque spectre indépendamment des autres, et proportionnellement à leur estimation.

Afin de retirer le bruit lié à la mesure, les spectres sont ensuite lissés par un filtre de Savitzky-Golay (80), communément utilisé en spectroscopie. Ce filtre présente les avantages d'être efficace, simple d'utilisation et permet d'éliminer le bruit tout en préservant les pics, contrairement à d'autres lissages (comme un lissage par moyenne). Son principe consiste à faire glisser une fenêtre le long du spectre, et d'approximer la portion de ce spectre par une fonction polynomiale. La valeur de ce polynôme au point central de la fenêtre constitue la valeur lissée du spectre en ce point. Ce processus est ensuite répété pour chaque point du spectre. Dans cette étude un polynôme de premier ordre et une taille de la fenêtre fixée à 9 points ont été appliqués.

Trois tests qualité sont ensuite réalisés sur les données. Le premier calcule un rapport signal sur bruit en considérant d'une part la gamme spectrale 2800-3150 cm^{-1} présentant un fort signal cellulaire, et d'autre part la gamme 1800-2200 cm^{-1} représentative d'une zone de bruit. Ce test qualité ne retient que les spectres dont le ratio signal/bruit est supérieur à trois fois leur écart type.

Le deuxième test qualité évalue un ratio entre la gamme spectrale propre au signal leucocytaire 1656-1720 cm^{-1} (Amide I) et celle relative au signal érythrocytaire 1540-1656 cm^{-1} (Hémoglobine). En effet, lorsqu'un globule blanc, sur lequel on réalise une acquisition Raman, est proche ou superposé à une hématie, un signal propre à l'hémoglobine de cette dernière peut apparaître dans le spectre enregistré. Ainsi, les spectres présentant un ratio signal globule blanc / globule rouge inférieur à 1 sont éliminés de l'étude. Cette valeur seuil a été déterminée par plusieurs tests considérant différentes combinaisons de spectres issus de leucocytes et d'hématies isolés.

Le troisième test concerne la détection de spike dans les spectres enregistrés. Une comparaison entre chacun d'eux et la moyenne des spectres d'une lame permet de facilement les détecter. Les spectres présentant un spike sont systématiquement éliminés de l'étude.

Une correction supplémentaire est réalisée afin de corriger les distorsions de ligne de base de chaque spectre. Dans ce but, une estimation polynomiale d'ordre 2 et une soustraction de cette ligne de base estimée sont réalisées pour chacun.

Finalement, une normalisation vectorielle sur l'ensemble de la gamme spectrale est réalisée *via* la fonction Standard Normal Variate (90). Cette normalisation a pour effet de s'affranchir des variations d'échelles des spectres, et permet ainsi de les rendre comparables entre eux. Après normalisation, la moyenne de chaque spectre est nulle et son écart type égal à un, présentant ainsi l'avantage pour les spectres d'avoir la même moyenne et dispersion.

3) Sélection de variables sur les spectres Raman

Dans le but de maximiser la performance des classifications sur les spectres Raman, une méthode multivariée englobante de sélection de variables, *randfeatures* (The MathWorks, Inc), est utilisée.

Soit un jeu d'entrainement composé de N objets de D variables, et répartis en C classes distinctes. La pertinence de ces variables quant à la classification des objets en ces C classes est déterminée par le calcul de scores lors du processus itératif suivant :

1. Choix aléatoire d'un sous-ensemble de p variables, $p \epsilon [1, 2, \dots, D]$.
2. Réduction de la dimension des objets à ces p variables.
3. Réalisation d'une classification par LDA.

Après classification, chaque objet présente un pourcentage d'appartenance à chacune des C classes (la somme de ces pourcentages est égale à 100%). Chaque objet sera considéré comme correctement classé si son pourcentage d'appartenance maximal est supérieur à un seuil *ConfidenceThreshold* fixé par l'utilisateur, et si ce pourcentage est attribué à la véritable classe de l'objet.

4. Les scores des p variables aléatoirement sélectionnées sont incrémentés si la proportion d'objets bien classés est supérieure à la valeur d'un seuil *PerformanceThreshold* fixé par l'utilisateur.

Ce processus est répété depuis l'étape 1) jusqu'à obtenir un nombre minimum de tirages vérifiant la condition 4). Ce nombre minimum *PoolSize* doit être fixé par l'utilisateur et choisi suffisamment grand afin que chaque variable puisse être statistiquement tirée plusieurs fois.

Une fois ce processus itératif achevé, les variables décrivant les objets sont ordonnées par ordre décroissant de leur score, donc de la plus pertinente à la moins significative.

L'application de cette méthode requière donc le choix de quatre paramètres par l'utilisateur, la taille des pools de variables à tester (p), les deux valeurs seuils (*ConfidenceThreshold* et *PerformanceThreshold*) et le nombre de tirage *Poolsize*.

Dans cette étude le jeu d'objets utilisé est constitué de 47 spectres Raman moyens, chacun composé de 1580 nombres d'onde (1580 variables). Deux classes sont représentées *via* 30 spectres moyens obtenus à partir de 1957 spectres acquis sur 30 échantillons leucémiques, et 17 spectres moyens calculés à partir de 831 spectres acquis sur 17 échantillons témoins.

La quantité de nombres d'onde p aléatoirement sélectionnés à chaque itération est fixée à 15. Les seuils *ConfidenceThreshold* et *PerformanceThreshold* sont respectivement fixés à 50% et 95%, et le nombre de tirages minimal *PoolSize* vérifiant la condition 4) est fixé à 1000.

À la suite de différents tests, les nombres d'onde les plus pertinents retenus pour cette étude sont ceux présentant un score supérieur à 15 (74 nombres d'onde).

IV) Classification des spectres

1) Classification non supervisée des spectres infrarouge : Fuzzy C-Means

a) Principe

L'algorithme de classification Fuzzy C-means (127) (FCM), est une généralisation de l'algorithme K-means (126). Il est non supervisé et est couramment utilisé en classification de données.

Le premier paramètre nécessaire à l'algorithme FCM est le nombre de classes, noté C, qui est fixé par l'utilisateur.

Étant donné X un ensemble de N objets $x_i, i \in [1, 2, \dots, N]$, le but de FCM est de trouver la partition de X qui minimise la fonction objectif suivante :

$$J(U, C) = \sum_{i=1}^{N} \sum_{j=1}^{C} (u_{i,j})^m \left\| x_i - c_j \right\|^2 \quad (12)$$

Où U est la matrice des degrés d'appartenance, C le nombre de classes, $u_{i,j}$ le degré d'appartenance de l'objet x_i au centroïde c_j, $j \in [1, 2, \dots, C]$.

Les degrés d'appartenance doivent respecter deux règles :

$$u_{i,j} \in [0,1] \; \forall \, i \in [1, 2, \dots, N], \forall \, j \in [1, 2, \dots, C]$$

$$\sum_{j=1}^{C} u_{i,j} = 1, \forall \, i \in [1, 2, \dots, N] \quad (13)$$

Chaque objet appartient donc à toutes les classes avec une probabilité $u_{i,j}$. Contrairement à d'autres algorithmes de classification dits « durs » (tels K-means ou HCA), où un objet appartient à une et une seule classe, les FCM permettent d'introduire la notion de flou, *i.e.* un objet est partagé entre plusieurs classes. Le contrôle du degré de flou est géré par le second paramètre de l'algorithme FCM, le paramètre de flou, noté $m, \in]1, +\infty[$.

Lorsque la valeur de m tend vers 1, alors la partition obtenue par FCM tend vers une partition obtenue par K-means, *i.e.* une classification dure où chaque objet est pratiquement attribué en totalité à une seule classe. Lorsque m croît, la quantité de flou augmente et les objets se partagent (non équitablement) entre les classes. En revanche, lorsque m devient

grand, le flou est complet se traduisant par des objets uniformément répartis sur l'ensemble des classes. Les classes sont alors toutes identiques est deviennent non informatives.

Dans la littérature, ce paramètre est généralement fixé à 2 lorsque l'algorithme FCM est appliqué à des données réelles (149-152).

Pour minimiser la fonction objectif (12), Bezdek a proposé un algorithme itératif et convergent, composé des étapes suivantes (127) :

1. Détermination du nombre de classes $C \in [2, 3, \dots, N]$
2. Initialisation aléatoire des degrés d'appartenance $u_{i,j}$ tels qu'ils respectent les deux règles citées en (13).
3. Calcul des centroïdes c_j, en considérant tous les objets pondérés par leurs degrés d'appartenance. On parle de repositionnement de centroïde selon la formule :

$$c_j = \frac{\sum_{i=1}^{N} (u_{i,j})^m x_i}{\sum_{i=1}^{N} (u_{i,j})^m}, \forall\, j \in [1, 2, \dots, C] \qquad (14)$$

4. Mise à jour des degrés d'appartenance suivant :

$$u_{i,j} = \frac{1}{\sum_{k=1}^{C} \left(\frac{d_{i,j}}{d_{i,k}}\right)^{(2/(m-1))}}, \forall\, i \in [1, 2, \dots, N], \forall\, j \in [1, 2, \dots, C] \qquad (15)$$

Où $d_{i,j}$ est la distance euclidienne entre l'objet x_i et le centroïde c_j.

5. Calcul de la fonction objectif J selon l'équation 12
6. Boucle sur les étapes 3, 4 et 5, jusqu'à convergence de l'algorithme, *i.e.* jusqu'à atteindre l'un des deux critères d'arrêt.

Le premier critère d'arrêt consiste à comparer les valeurs de la fonction objectif calculées à chaque itération. Si la différence de deux valeurs consécutives ne varie plus au-delà d'un certain seuil (fixé), alors la classification est considérée comme convergente et le résultat final est retourné. Le second critère d'arrêt représente un nombre d'itérations limité. Ce critère est nécessaire dans certains cas, minoritaires, où les valeurs objectives consécutives oscillent et ne tendent à converger que très lentement.

b) Avantages et limites

De manière générale, l'algorithme FCM présente l'avantage d'être rapide et basé sur des calculs mathématiques et une logique simples. C'est également un algorithme de classification dit « flou », puisque chaque objet est attribué à l'ensemble des classes avec une pondération, contrairement aux algorithmes de classification dits « durs », tels les K-means, pour lesquels un objet appartient à une et une seule classe. L'analyse de ces attributions pondérées permet de mettre en avant une hétérogénéité intra et inter-classes pouvant fortement augmenter le potentiel informatif du résultat obtenu. Un exemple illustratif serait l'interprétation d'une classification en deux clusters, où un objet pratiquement équidistant des deux centroïdes se verrait attribué à 100% au plus proche dans le cas d'une classification dure, alors qu'une classification par FCM mettrait en avant des degrés d'appartenance de 55% au premier centroïde et de 45% au second.

Les objets à la frontière entre deux classes ou plus peuvent être mis en évidence et ainsi faire apparaître des zones de transition ou de chevauchement entre ces classes.

Cependant, l'algorithme FCM reste sensible aux objets aberrants qui ont pour effet de fausser les repositionnements des centroïdes à chaque itération. Une seconde remarque majeure peut également être faite quant à l'initialisation aléatoire des centroïdes lors de la première itération. Il faut savoir que l'exécution d'un FCM sur un même jeu de données mais avec deux initialisations distinctes (car aléatoires) peut mener à des résultats totalement différents. Cette initialisation peut donc être considérée comme un avantage ou un inconvénient suivant que l'on désire de la robustesse dans les résultats obtenus après plusieurs exécutions de l'algorithme, ou que l'on cherche à explorer différentes possibilités de classification du jeu d'objets testé.

Des solutions à ces faiblesses sont apportées dans la littérature. Que ce soit par des calculs préalables afin de déterminer les meilleurs individus à choisir au démarrage de l'algorithme et ainsi s'affranchir du choix aléatoire (153, 154); ou encore par des modifications intrinsèques aux calculs eux-mêmes (155, 156).

Il est également à noter que le choix empirique de *C* est une limite à l'algorithme FCM, puisque le nombre de classes optimal ne peut être déterminé par l'algorithme seul.

Concernant le degré de flou m, le choix d'une valeur trop faible ou trop forte influe considérablement sur les résultats obtenus (classification dure ou non informative). Le paramétrage des FCM est donc de premier intérêt. En effet, lors d'une étude appliquée à un adénocarcinome colique, Lasch *et al* ont montré qu'un mauvais paramétrage des FCM pouvait mener à une classification moins pertinente en comparaison des résultats obtenus par un algorithme de classification « dur » (157).

L'estimation automatique des valeurs optimales de *C* et de *m* apparaît donc de fort intérêt.

c) Indices de validité et m-rules

Dans le but d'automatiser le choix du nombre de classes C, des méthodes objectives, appelées indices de validité (IV), ont été développées. Les indices de validité existent en nombre dans la littérature (158), et ont pour principe de calculer une valeur d'indice pour une classification FCM donnée. Ainsi, plusieurs valeurs d'indice peuvent être obtenues en faisant varier le paramétrage de C et de m lors de différentes classifications FCM. Généralement, ces valeurs d'indice sont définies comme le rapport entre la compacité et la séparation des classes obtenues. La valeur minimale (ou maximale selon l'indice considéré) parmi celles calculées détermine la classification optimale *i.e.* les valeurs optimales de C et m.

En ce qui concerne le degré de flou m, peu de méthodes existent pour estimer sa valeur optimale (159-161). Ces méthodes, nommées « règles en *m* » ou « *m*-rules » déterminent la valeur optimale du degré de flou en considérant, soit les résultats obtenus pour une classification donnée (comme pour les indices de validité), soit les données elles-mêmes. Ces méthodes s'appuient sur des calculs de compacité des classes, sur le nombre d'erreur de classification ou encore sur l'estimation d'une valeur maximale à partir de laquelle la valeur optimale de m est calculée.

La difficulté principale concernant ces méthodes d'estimation est que les indices de validité supposent *m* fixé à sa valeur standard 2, alors que les règles en *m* supposent le nombre de classes optimal connu. Or, C et m sont intimement liés. Il est donc indispensable de considérer des méthodes d'estimation simultanée de ces deux paramètres.

De telles méthodes ont été proposées dans la littérature. L'algorithme proposé par Schwämmle et Jensen (162) est une version optimisée de l'algorithme de Futschik et Carlisle (163), où des valeurs de m successives et croissantes sont considérées pour un nombre de classe fixé. Au cours de cette augmentation de la valeur de flou, les données sont de plus en plus distribuées sur les centroïdes et l'apparition de deux centroïdes identiques déterminera la valeur optimale de m. Cette méthode est réalisée pour différents nombre de classes qui retournent tous une valeur de m optimale pratiquement identique. Le degré de flou est ensuite fixé à cette valeur, et des valeurs de C croissantes sont testées, jusqu'à obtenir une distance inter-centroïdes minimale, qui déterminera le nombre de classes optimal.

On notera toutefois que ces algorithmes n'estiment pas les deux paramètres des FCM simultanément mais consécutivement.

Au sein du laboratoire, un algorithme déterminant simultanément les valeurs optimales de ces deux paramètres à été développé et appliqué à des données d'imagerie spectrale infrarouge de cancers cutanés (145). Cet algorithme est basé sur le calcul de redondance entre les classes et introduit à ce titre un paramètre seuil permettant de gérer les valeurs de redondance calculées.

Différentes valeurs de seuil sont testées, et pour chaque couple (m, seuil) une valeur optimale de C est calculée.

Pour une valeur donnée du degré de flou, un nombre de classes trop grand engendre une classification par FCM présentant des classes doublons ou redondantes. Deux classes seront considérées comme telles si le coefficient d'intercorrélation des degrés d'appartenance des objets à ces classes s'avère supérieur à une valeur seuil donnée. En incrémentant le nombre de classes pour chaque valeur de m testée, un nombre de classes optimal est obtenu, *i.e.* au delà de ce nombre de classes, le résultat présente des classes redondantes.

À partir de ces résultats, des courbes sont tracées, représentant les valeurs de C optimales pour tous les degrés de flou testés. Les valeurs de C et m optimales finales sont obtenues en analysant les paliers de ces courbes.

Cet algorithme présente d'excellents résultats sur les images infrarouge testées mais demeure sensible à d'autres jeux de données (comme artificiels ou simulés). On notera également que sa structure interne itérative nécessite un temps d'exécution considérable.

2) Classification des spectres Raman : Support Vector Machine

a) Principe

L'algorithme de classification Support Vector Machine (135, 136) est un puissant algorithme de classification supervisée bien connu en classification de données, notamment lorsqu'il est appliqué à la reconnaissance de formes (164), de textes (165), mais également dans le domaine de la bioinformatique (166-168).

Pour expliquer les SVMs, nous allons nous placer dans le cas le plus simple, c'est-à-dire un problème de discrimination entre deux populations d'objets à deux dimensions. Par la suite, on parlera d'objet positif (respectivement négatif) si cet objet appartient à la classe + (respectivement -). Comme tout algorithme de classification supervisée, l'algorithme SVM nécessite un jeu d'entrainement, constitué d'objets labellisés dans les deux classes, pour estimer un modèle de prédiction.

Pour résoudre ce problème, les SVMs se basent sur deux grands concepts : les séparations à vaste marge et les fonctions noyaux (ou fonctions kernel).

b) Les SVMs linéaires

Pour introduire le concept de séparation à vaste marge, considérons le jeu de données présenté sur la Figure 20. Intuitivement, une façon simple de classer les objets est de tracer une droite entre les deux groupes. Ce jeu de données est ainsi qualifié de linéairement séparable. De plus, les deux populations d'objets étant complètement séparées, on cherchera à tracer la droite la plus éloignée de ces deux groupes, donc la droite qui laissera la plus grande marge possible entre les objets de chaque groupe. Cette idée est à la base du développement des SVMs.

Schématiquement, les objets positifs et négatifs les plus à la frontière inter-population sont sélectionnés et constituent les vecteurs supports. Un exemple est proposé sur la Figure 20 où les vecteurs supports sont encadrés. À partir de ces vecteurs, un hyperplan *(H)* est calculé

pour séparer au mieux les deux populations, *i.e.* on recherche l'hyperplan générant la plus grande marge entre les deux populations. Les SVMs linéaires ont pour caractéristique principale de trouver un hyperplan séparateur des populations sans transformation préalable des objets.

On note H_+ l'hyperplan parallèle à H et contenant les objets + les plus proches de H (c'est-à-dire les vecteurs supports +). De même, on note H_-, l'hyperplan parallèle à H et contenant les objets - les plus proches de H (vecteurs supports -). Il existe une infinité d'hyperplans H séparant deux populations linéairement séparables, l'estimation de l'hyperplan optimal se fait donc sur un critère de maximisation de l'écart entre H_+ et H_-. La distance entre H_+ et H étant identique à celle entre H_- et H, ces distances sont appelées marge (Figure 20).

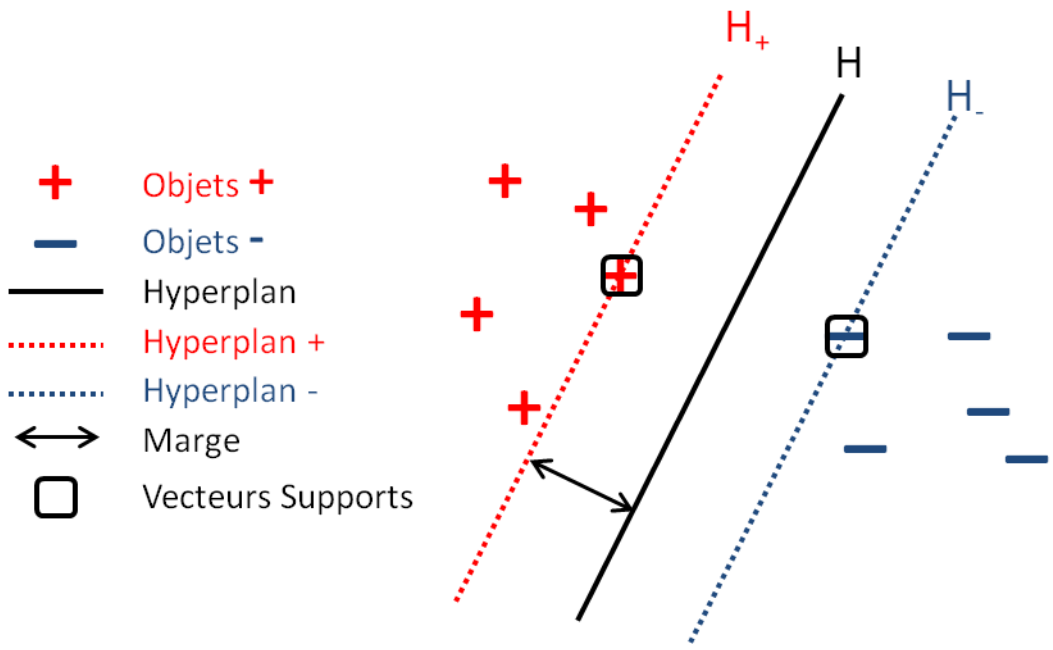


Figure 20 : Hyperplan et populations linéairement séparables

Schéma illustratif du calcul d'un hyperplan dans le cas de populations linéairement séparables (hyperplan non optimal) en considérant un individu de chaque population comme vecteurs supports.

Nous allons maintenant décrire les SVMs linéaires d'un point de vue mathématique.

Soit X un ensemble d'objets $x_i, \epsilon\ \mathbb{R}^D, i\ \epsilon\ [1, \dots, N]$, ayant pour label $y_i = -1$ si x_i appartient à la classe – et $y_i = +1$ si x_i appartient à la classe +.

Tout point x dans un espace pouvant être représenté par un vecteur $\vec{x}$ dans cet espace, l'équation d'un hyperplan H vérifie l'équation :

$$H(x) = \vec{w}.\vec{x} + b = 0 \quad (16)$$

où $\vec{w}\ \epsilon\ \mathbb{R}^D$ est le vecteur normal à l'hyperplan, $\vec{x}\ \epsilon\ \mathbb{R}^D$ est parallèle à l'hyperplan, $b\ \epsilon\ \mathbb{R}$ une constante et « . » le produit scalaire entre deux vecteurs. Le but des SVMs linéaires est de trouver la fonction linéaire discriminante f, donc d'estimer $\vec{w}$ et b, telle que pour tout objet $\vec{x}$ de classe y :

$$\begin{cases} f(\overrightarrow{x_i}) = \vec{w}.\overrightarrow{x_i} + b\ \geq 1\ si\ y_i = +1 \\ f(\overrightarrow{x_i}) = \vec{w}.\overrightarrow{x_i} + b\ \leq 1\ si\ y_i = -1 \end{cases} \quad (17)$$

Ces équations peuvent également s'écrire :

$$y_i(\vec{w}.\overrightarrow{x_i} + b) \geq 1, \forall\ \overrightarrow{x_i}\ \epsilon\ X \quad (18)$$

L'hyperplan H défini par l'équation 16 sépare donc linéairement les deux populations et est appelé frontière de décision.

Le but est de maximiser la marge. Si la fonction discriminante f est normalisée de façon à ce que $f(\overrightarrow{x_i}) = +1$ pour les vecteurs supports de la classe + et que $f(\overrightarrow{x_i}) = -1$ pour les vecteurs supports de la classe -, alors la marge est égale à $\frac{1}{\|\vec{w}\|}$. Maximiser la marge $\frac{1}{\|\vec{w}\|}$ revient à minimiser $\|\vec{w}\|$. Cependant, par commodité mathématique, l'expression suivante équivalente, nommée *formulation primale*, est préférée :

$$\begin{cases} Minimiser\ \frac{1}{2}\|\vec{w}\|^2 \\ Sous\ contrainte\ y_i(\vec{w}.\overrightarrow{x_i} + b) \geq 1, \forall\ \overrightarrow{x_i}\epsilon\ X \end{cases} \quad (19)$$

Le problème d'optimisation non linéaire avec contrainte ainsi obtenu peut être résolu par la méthode de Lagrange, permettant de le transformer en un problème d'optimisation sans contrainte, appelé *formulation primale du lagrangien :*

$$L_P = \frac{1}{2}\|\vec{w}\|^2 - \sum_{i=1}^{N} \alpha_i y_i(\vec{w}.\overrightarrow{x_i} + b) + \sum_{i=1}^{N} \alpha_i \quad (20)$$

où $\alpha_i\ \geq 0$ est une constante pour chaque contrainte, ou « multiplicateur de Lagrange ».

En minimisant L_P par rapport à $\vec{w}$ et b et en maximisant par rapport à α_i, on en déduit la *formulation duale du lagrangien* :

$$\begin{cases} Maximiser\ L_D = \sum_{i=1}^{N} \alpha_i - \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N} \alpha_i \alpha_j y_i y_j\ \vec{x_i}.\vec{x_j} \\ Sous\ contrainte \sum_{i=1}^{N} y_i \alpha_i = 0, 0 \leq \alpha_i \end{cases} \quad (21)$$

Le vecteur $\|\vec{w}\|$ s'exprime alors sous la forme :

$$\vec{w} = \sum_{i=1}^{N} y_i \alpha_i \vec{x_i} \quad (22)$$

Le maximum de L_D et le minimum de L_P sont obtenus pour les mêmes valeurs de $\vec{w}$, b et α_i.

En pratique, seuls les objets présents sur les hyperplans H_+ et H_- présentent un multiplicateur de Lagrange non nul. Ces objets constituent les vecteurs supports.

L'hyperplan optimal peut être écrit :

$$h(x) = \vec{w}^*.\vec{x} + b^*$$

ou encore

$$h(x) = \sum_{i=1}^{N} \alpha_i{}^* y_i.\vec{x}.\vec{x_i} + b^* \quad (23)$$

avec $\alpha_i{}^*$ les solutions de l'équation 21, et b^* obtenu par l'équation :

$$b^* = y_k - \sum_{i=1}^{N} \alpha_i y_i \vec{x_i}.\vec{x_k} \quad (24)$$

où $\vec{x_k}$ est un vecteur support.

Dans certains cas les populations peuvent présenter un chevauchement de leurs données. Aucune séparation parfaite entre elles ne peut donc être estimée, *i.e.* il n'existe pas de marge et donc pas d'hyperplan. On parle alors d'objets non linéairement séparables.

Pour pallier ce problème, les contraintes définies par l'équation 18 sont modifiées afin d'autoriser des erreurs de classification suivant :

$$y_i(\vec{w}.\vec{x_i} + b) \geq 1 - \xi_i, \forall\, \vec{x_i}\, \epsilon\, X \tag{25}$$

où $\xi_i \geq 0$ est appelé variable ressort et permet à un objet soit d'appartenir à la marge soit d'être mal classé. Afin de limiter la quantité de tels objets, l'expression $\frac{1}{2}\|\vec{w}\|^2$ à minimiser dans la formulation primale (équation 19) est pondérée par l'ajout d'un terme $C\sum_{i=1}^{N}\xi_i$.

Le nouveau problème d'optimisation devient donc :

$$\begin{cases} Minimiser\ \frac{1}{2}\|\vec{w}\|^2 + C\sum_{i=1}^{N}\xi_i \\ Sous\ contrainte\ y_i(\vec{w}.\vec{x_i} + b) \geq 1 - \xi_i, \forall\, \vec{x_i}\epsilon\, X \end{cases} \tag{26}$$

La constante $C > 0$ permet d'introduire un compromis entre la maximisation de la marge et la minimisation du nombre de d'erreur de classification. Plus la valeur de C est grande, moins les erreurs de classification sont tolérées lors de la recherche de l'hyperplan optimal. L'introduction du terme $C\sum_{i=1}^{N}\xi_i$ dans le problème d'optimisation modifie la formulation duale de l'équation 21 uniquement par l'extension de la contrainte $0 \leq \alpha_i$ qui devient $0 \leq \alpha_i \leq C$. Les ${\alpha_i}^*$ ayant été déterminés en tenant compte de cette nouvelle contrainte, $\vec{w}^*$ et b^* sont estimés en utilisant les équations 22 et 24.

La Figure 21 illustre trois hyperplans calculés pour un même jeu de données, en prenant en compte différents vecteurs supports. D'après cette figure, on constate que la recherche d'un nombre d'erreur minimal peut également minimiser la marge entre les deux populations, rendant le modèle de prédiction calculé moins généralisable.

L'application d'un algorithme SVM linéaire sur un jeu de données non linéairement séparable va donc retourner un hyperplan optimal permettant de séparer celles-ci au mieux (nombre d'erreur minimal). Cependant, selon le jeu de données considéré, le meilleur résultat obtenu peut ne pas être suffisant, comme le montre la Figure 22 où deux populations d'objets sont imbriquées de façon concentrique.

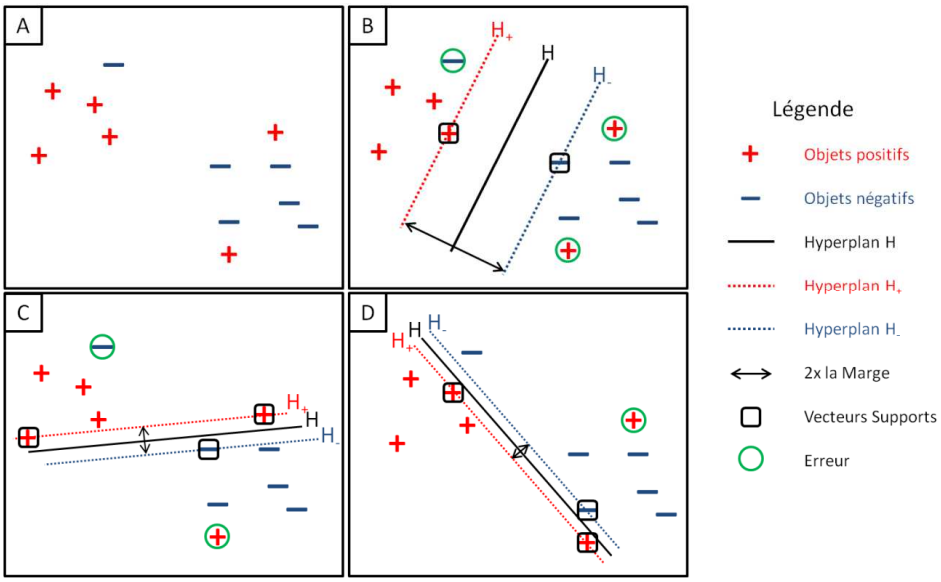


Figure 21 : Hyperplans et populations non linéairement séparables

Schéma illustratif de différents calculs d'hyperplans séparateurs dans le cas de (A) deux populations non linéairement séparables, (B) hyperplan estimé pour un $\boldsymbol{C}$ petit comptant 3 erreurs de classification et une forte marge, (C) hyperplan estimé pour un $\boldsymbol{C}$ moyen comptant 2 erreurs de classification et une marge réduite, (D) hyperplan estimé pour un $\boldsymbol{C}$ grand comptant une seule erreur de classification et une faible marge.

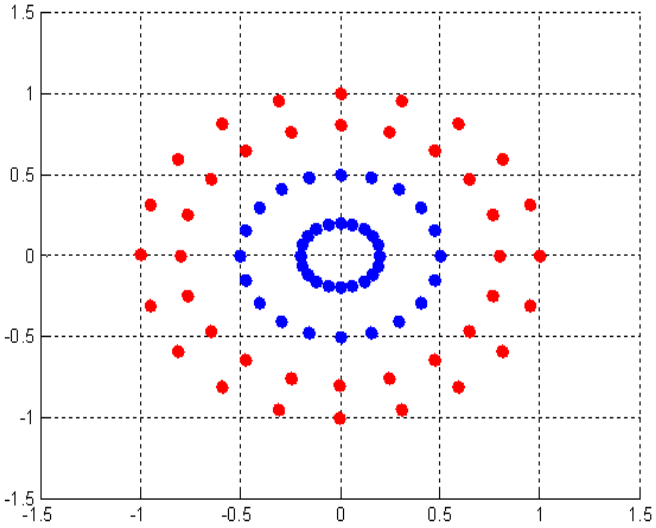


Figure 22 : Populations d'objets imbriquées de façon concentrique

Illustration de deux populations de données de dimension 2 et non linéairement séparables. Tout hyperplan calculé par un SVM linéaire comptera près de la moitié des objets d'une des deux populations comme des erreurs de classification.

Dans ce cas, tout hyperplan entraine un nombre d'erreur de classification de près de la moitié des objets pour l'une des deux classes. La recherche d'un hyperplan par une méthode de classification non linéaire apparaît donc plus pertinente.

c) Les SVMs non linéaires

Les SVMs non linéaires, appliqués à des objets non linéairement séparables, sont basés sur le théorème de Cover (169) démontrant qu'une transformation de l'espace des données, généralement en un espace de dimension supérieure pouvant aller jusqu'à l'infini, peut rendre linéairement séparables les objets projetés dans ce nouvel espace. Cet espace est communément appelé « espace des caractéristiques » ou « feature space » en anglais. Après projection des objets dans l'espace des caractéristiques, une classification par SVM linéaire peut être réalisée.

Cette projection des objets depuis l'espace des données $\mathbb{R}^D$, vers un espace des caractéristiques $\mathcal{F}$, est réalisée *via* une fonction non linéaire, Φ telle que :

$$\begin{array}{rccc} \Phi : & \mathbb{R}^{\mathrm{D}} & \rightarrow & \mathcal{F} \\ & x & \rightarrow & \Phi(x) \end{array} \qquad (27)$$

La Figure 23 illustre une projection non linéaire des deux populations concentriques présentées en Figure 22, vers un espace de dimension supérieure où il devient possible de trouver un hyperplan, linéaire dans cet espace, séparant complètement ces deux populations.

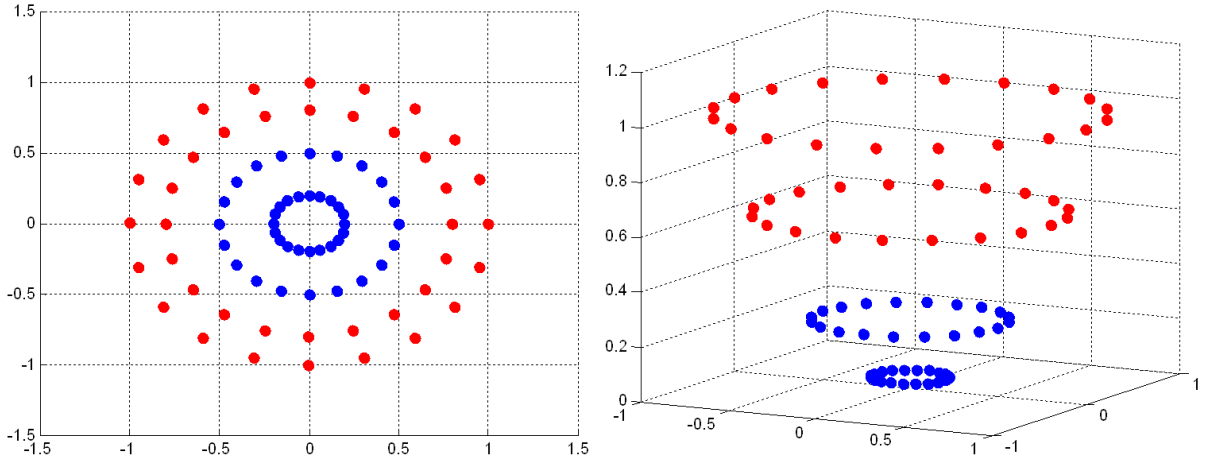


Figure 23 : Projection non linéaire de données

Exemple d'une projection d'objets non linéairement séparables dans l'espace des données ($\mathbb{R}^2$) vers un espace des caractéristiques $\mathcal{F}$ de dimension supérieure ($\mathbb{R}^3$) par une fonction polynomiale de degré 2.

La fonction discriminante de l'équation 17 s'écrit alors dans le cas d'une transformation non linéaire sous la forme :

$$\begin{cases} f(\overrightarrow{x_i}) = \vec{w}.\Phi(\overrightarrow{x_i}) + b \geq 1 \; si \; y_i = +1 \\ f(\overrightarrow{x_i}) = \vec{w}.\Phi(\overrightarrow{x_i}) + b \leq 1 \; si \; y_i = -1 \end{cases} \quad (28)$$

Dans l'espace des caractéristiques $\mathcal{F}$, la fonction discriminante f est une fonction linéaire. Dans l'espace des données $\mathbb{R}^D$, elle est non linéaire si Φ est une fonction non linéaire.

Le problème d'optimisation dans l'espace $\mathcal{F}$ devient alors :

$$L_D = \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i \alpha_j y_i y_j \; \Phi(\overrightarrow{x_i}).\Phi(\overrightarrow{x_j}) \quad (29)$$

et l'équation de l'hyperplan optimal :

$$h(x) = \sum_{i=1}^{N} {\alpha_i}^* y_i.\Phi(\vec{x}).\Phi(\overrightarrow{x_i}) + b^* \quad (30)$$

La solution du problème d'optimisation ne dépend donc plus du produit scalaire $\vec{x}.\overrightarrow{x_i}$ dans l'espace des données, mais du produit scalaire $\Phi(\vec{x}).\Phi(\overrightarrow{x_i})$ dans l'espace des caractéristiques. Il n'est donc pas nécessaire de connaître explicitement la forme de la transformation par la fonction Φ, qui peut engendrer des données transformées de très grande dimension difficilement calculables, puisque seule la connaissance de son produit scalaire est nécessaire. Cette astuce est connue sous le nom d'« astuce du noyau » ou « kernel trick » en anglais.

Supposons que l'on dispose d'une fonction K, dite « fonction noyau », telle que :

$$K(x_i, x_j) = \Phi(\mathrm{x_i}).\Phi(\mathrm{x_j}) \quad (31)$$

La possibilité de calculer un produit scalaire dans l'espace $\mathcal{F}$ par l'intermédiaire de la fonction noyau permet de s'affranchir de la fonction Φ.

L'hyperplan optimal peut ainsi être obtenu en remplaçant $\Phi(\mathrm{x_i}).\Phi(\mathrm{x_j})$ par $K(x_i, x_j)$ dans les équations 29 et 30.

Parmi les fonctions noyaux les plus usitées, on trouve le noyau polynomial :

$$K(x_i, x_j) = (a(\overrightarrow{x_i}.\overrightarrow{x_j}) + b)^p \quad (32)$$

avec p le degré du polynôme.

Et le noyau gaussien :

$$K(x_i, x_j) = exp\left(-\frac{\left\|\mathrm{x_i} - \mathrm{x_j}\right\|^2}{2\Theta^2}\right) \quad (33)$$

avec $\Theta > 0$ qui permet de gérer l'écart type de la gaussienne. Une valeur minimale de Θ permet d'approcher des formes très irrégulières alors qu'une valeur plus élevée produira des formes plus lisses.

d) Avantages et limites

L'un des points forts de l'algorithme de classification Support Vector Machine est sa capacité à traiter rapidement des données de grande dimension. De plus, il rend possible une séparation linéaire des objets ne l'étant pas dans leur espace de représentation d'origine en apportant une solution optimum globale. De ce fait, la présence de données aberrantes ou non homogènes parmi le jeu d'entrainement n'empêche pas nécessairement la création d'un modèle de prédiction efficace. En effet, dans le cas d'une classification de données non linéairement séparables le taux d'erreur pondérable autorise une modulation entre l'importance accordée à ces erreurs et la complexité du classifieur (selon la fonction noyau choisie).

En ce qui concerne le calcul de l'hyperplan séparateur, que ce soit pour des populations linéairement séparables ou non, il est réalisé à partir de vecteurs supports considérablement inférieurs en nombre par rapport à la totalité des données du jeu d'entrainement, concentrant ainsi le calcul de cet hyperplan sur les éléments les plus à la frontière entre les deux populations d'objets, et simplifiant la complexité et le temps de calcul des modèles de prédiction.

Cependant, la gestion des paramètres reste délicate et peut influencer grandement la pertinence du modèle de prédiction.

Prenons le cas du sur-apprentissage (overfitting en anglais) où seuls les objets ayant servi à la construction du modèle sont correctement discriminés, et où tout autre individu est labellisé comme l'une des deux classes exclusivement. Ce phénomène rend le modèle de prédiction spécifique au jeu d'entrainement utilisé pour le construire, et se présente lors d'une trop forte estimation du paramètre d'erreur.

Chapitre 3 : Résultats et discussions

Dans ce chapitre sont présentés sous forme d'articles les résultats obtenus au cours de ces différentes études.

Préambule

Les résultats obtenus au cours de cette thèse s'articulent autour de deux thèmes de recherche et sont présentés ici sous forme de publications scientifiques.

Le premier thème de recherche de cette thèse était axé autour d'une étude combinant la technologie de microspectroscopie infrarouge à l'optimisation d'un nouvel algorithme de classification non supervisée. Cet algorithme a permis d'une part, d'apporter un supplément informatif dans un but diagnostic de cancers cutanés, et d'autre part, a montré un fort potentiel autant sur le plan du temps de calcul que dans la qualité des résultats obtenus. Ces travaux ont fait l'objet de trois publications.

La **publication n°1** propose une comparaison entre les résultats obtenus *via* 23 indices de validité, 4 règles en *m* et une amélioration de l'algorithme de classification Fuzzy C-Means Redundant Based Algorithm (FCM-RBA), développé au sein du laboratoire. Ces classifications ont été réalisées sur 16 images spectrales infrarouge de cancers cutanés.

La **publication n°2** est une application de cet algorithme amélioré, nommé ici FCM-URBA (Upgraded), à des jeux de données simulés et réels classiquement utilisés dans la littérature. 5 indices de validité récemment publiés y sont également testés ainsi qu'une méthode permettant elle aussi d'estimer à la fois les valeurs optimales des deux paramètres à gérer, *m* et *C*.

La **publication n°3** est une revue des différents indices de validité et règles en *m* présents dans la littérature, appliqués à des images spectrales infrarouge et des jeux de données classiquement utilisés dans la littérature, qu'ils soient réels ou artificiels.

Le second thème de recherche a été mené dans le cadre du projet ANR TECSAN IHMO, où une machine multimodale couplant plusieurs technologies de pointe a pu être mise au point. Ce module a permis de réaliser une étude visant à améliorer le diagnostic de la LLC au moyen de la technologie de microspectroscopie Raman et d'un algorithme de classification supervisée. Ces travaux ont fait l'objet de trois publications.

La **publication n°4** met en avant le projet IHMO, ainsi que les résultats de classification Raman préliminaires obtenus dans ce cadre.

La poursuite de cette étude préparatoire a permis de considérablement augmenter la base de données disponible. Les résultats de classification obtenus ont tout d'abord conforté ceux de la première publication, et ont fait l'objet d'une présentation dans un congrès international puis d'une publication sous la forme d'abstract étendu (**publication n°5**).

Une description détaillée des traitements de données, des méthodes statistiques multivariées et de l'algorithme de classification utilisés a fait l'objet de la **publication n°6**.

En perspective de ce projet, le module IHMO a également permis de réaliser des tests exploratoires dans un tout autre domaine d'application, dont l'objet était de détecter la présence du parasite *Plasmodium falciparum* (responsable de la malaria) dans les hématies d'un échantillon de sang contaminé. Ces premiers travaux ont été présentés lors d'un congrès international et ont fait l'objet de la **publication n°7**.

Article #I

Publication n°1

FCM parameter estimation methods: Application to infrared spectral histology of human skin cancers

T. Happillon, D. Sebiskveradze, V. Vrabie, O. Piot, P. Jeannesson, M. Manfait, C. Gobinet.

Proceedings of the 20th European Signal Processing Conference (EUSIPCO 2012) **2012**, pp. 1831-1835

Présentation de la publication n°1

Contexte

En oncologie, une grande attention est portée aux techniques qui permettraient de diagnostiquer de plus en plus précocement les cancers cutanés. En effet, un diagnostic de la maladie lors de ses premiers stades améliore les effets thérapeutiques des traitements médicaux administrés, et augmente ainsi l'espérance de vie des patients. Une analyse histologique conventionnelle est basée sur l'observation au microscope par un anatomo-pathologiste, d'une coupe de tissu (biopsie ou exérèse) colorée à l'Hématoxyline et Eosine (H&E). Les critères morphologiques sont donc fortement considérés lors d'un tel diagnostic. Cependant, une analyse moléculaire de l'échantillon par microspectroscopie infrarouge (IR) à transformée de Fourier, sans coloration ni destruction de celui-ci, peut apporter un réel surplus d'informations pertinentes pour aider le spécialiste à établir son diagnostic. L'application d'algorithmes de classification floue non supervisée (comme les FCM) permet d'accéder à des informations nouvelles invisibles par histologie conventionnelle, telles les régions d'interconnectivité entre tissus.

Objectif

Cependant, le comportement des FCM est contrôlé par deux paramètres : le nombre de classes C, et le paramètre de floue, m. Des méthodes, appelées indices de validité, ont été développées dans la littérature pour estimer le nombre de classes pour un paramètre de floue fixé. Inversement, des méthodes, nommées règles en *m*, permettent d'estimer le paramètre de floue pour un nombre de classes donné. Notre équipe a récemment développé une méthode innovante d'estimation simultanée de ces deux paramètres spécifiquement pour l'imagerie spectrale infrarouge. Bien que donnant d'excellents résultats lorsqu'appliquée sur des images spectrales infrarouge de tumeurs cutanées, cette méthode dénommée Fuzzy C-Means Redundant Based Algorithm (FCM-RBA) souffre de deux problèmes. En effet, elle nécessite un temps de calcul conséquent, et ces résultats n'ont jamais été comparés à ceux d'indices de validité récents ou de règles en *m*.

Le premier objectif de ce travail est donc de réduire le temps de calcul de cet algorithme appliqué sur des images spectrales infrarouge. Le deuxième objectif est de montrer la

supériorité de cette version accélérée de FCM-RBA vis à vis d'indices de validité récents et de règles en *m*. La comparaison sera faite sur la pertinence histologique des partitions floues estimées par ces différentes méthodes, lorsqu'elles sont appliquées à des images spectrales infrarouge de lésions cutanées.

Matériels et Méthodes

Dans cette étude, 16 échantillons de peaux cancéreuses incluses en paraffine ont été considérées. Parmi ces échantillons différents types de cancers cutanés sont représentés : 4 carcinomes basocellulaires, 3 carcinomes spinocellulaires, 4 maladies de Bowen et 5 mélanomes. Pour chacun, des coupes de 10µm d'épaisseur ont été réalisées à l'aide d'un microtome et ont été déposées sur un support de CaF_2 sans aucun agent fixateur. Des coupes adjacentes aux coupes imagées par infrarouge ont été colorées à l'H&E afin d'être analysées de façon conventionnelle par un spécialiste et ainsi déterminer les différentes catégories tissulaires présentes dans chaque échantillon. Ces analyses conventionnelles font office de références pour estimer la qualité des résultats obtenus au cours de cette étude. Les acquisitions de données ont été réalisées *via* un système d'imagerie (Microscope Spotlight 300, Perkin Elmer) associé à un spectromètre Infrarouge à Transformée de Fourrier (Spectrum One, Perkin Elmer) avec une résolution spatiale de 6,25 μm et une résolution spectrale de $4cm^{-1}$. Les données ainsi acquises ont ensuite été prétraitées par EMSC et normalisées avant d'être considérées pour les différentes classifications. L'algorithme FCM-RBA accéléré, 23 indices de validité et 4 règles en *m* ont été appliqués à chacun des 16 échantillons.

Résultats

Les résultats de classification obtenus au cours de cette étude montrent que pour chaque échantillon, le nombre de classe optimal estimé par les indices de validité est soit inférieur, soit très supérieur au nombre de structures diagnostiquées par l'anatomo-pathologiste sur les coupes colorées H&E. De plus, les valeurs de *m* (qu'elles soient estimées par les règles en *m* ou standard) ne permettent pas de mettre en avant les zones de transition entre les différents tissus (*m* trop faible) ou au contraire ne rendent pas les différentes classes distinguables (*m* trop élevé). Les résultats obtenus *via* l'algorithme FCM-RBA amélioré sont quant à eux en parfaite corrélation avec les analyses de coupes H&E. Le nombre de classes

optimal retourné par l'algorithme est en accord avec le nombre de structures tissulaires dans chaque coupe. La valeur optimale de *m* conjointement déterminée, n'est ni sous-estimée ni sur-estimée, et permet de mettre en avant les zones de transition entre les tissus. Ces résultats ont été constatés sur l'ensemble des 16 biopsies considérées dans cette étude.

Conclusion

On peut constater dans la littérature, et au vu des résultats de cette étude, que l'algorithme de classification Fuzzy C-Means est bien adapté aux analyses de données spectrales infrarouge. Cependant, le choix du paramétrage demeure très important puisque les résultats de classification en sont fortement dépendants, et c'est d'ailleurs dans le but d'obtenir une estimation optimale de ce paramétrage que les indices de validité et règles en *m* ont été développés. Dans cette étude, ces différents outils ont été comparés à une amélioration de la méthode FCM-RBA. Cette comparaison a montré que les résultats obtenus *via* cette dernière étaient meilleurs en termes de qualité et de temps de calcul que ceux obtenus par les indices de validité, combinés ou non aux règles en *m*. En effet, pour chaque échantillon analysé, les classes estimées par FCM-RBA amélioré représentent les différentes structures tissulaires, et la valeur de *m* associée permet de mettre en avant non seulement les possibles interactions entre elles, mais également l'hétérogénéité au sein d'une même structure. Les résultats obtenus sur les 16 échantillons de cancers cutanés testés ont ainsi permis de facilement localiser et caractériser les tumeurs.

Les apports de telles classifications (*i.e.* zones de transitions et hétérogénéités au sein des structures tissulaires), présentent un réel complément aux analyses classiques réalisées sur coupes colorées H&E et de ce fait, sont d'un grand intérêt d'un point de vue biomédical.

FCM PARAMETER ESTIMATION METHODS: APPLICATION TO INFRARED SPECTRAL HISTOLOGY OF HUMAN SKIN CANCERS

T. Happillon[1], D. Sebiskveradze[1], V. Vrabie[2], O. Piot[1], P. Jeannesson[1], M. Manfait[1], C. Gobinet[1]

[1]Unité MéDIAN, CNRS FRE 3481 MEDyC, UFR de Pharmacie, Université de Reims Champagne-Ardenne, 51 rue Cognacq-Jay, 51096 Reims Cedex, France
[2]CReSTIC, Université de Reims Champagne-Ardenne, Moulin de la Housse, BP 1039, 51697 Reims Cedex, France

ABSTRACT

Spectral histology of cancer can be achieved thanks to the analysis of infrared (IR) hyperspectral images. The Fuzzy C-Means (FCM) clustering is particularly well adapted since each object is attributed to all clusters with different membership values. Applied on IR hyperspectral images of human skin cancers, it can highlight fine transitions between tumor and surrounding tissues and/or tumor heterogeneities. However, to provide a biomedically interpretable clustering, the relevant values of the two FCM parameters, i.e. the number of clusters and the fuzziness parameter, must be judiciously selected for each analyzed tissue. In this paper, the performance of some classical cluster validity indices and *m*-rules previously presented in the literature are evaluated. A new heuristic method based on the redundancy of FCM clusters is also proposed. We show that our method highly improves the clustering quality and computational time when applied on IR images of human skin tumors.

Index Terms— Infrared spectral imaging, skin cancer, Fuzzy C-Means, validity indices, *m*-rules, spectral histology

1. INTRODUCTION

A challenge in oncology is to early diagnose tumors in order to increase the therapeutic effects of medical treatments and improve the life expectancy of patients. Traditionally, the diagnosis is realized by conventional histology which is based on the morphological analysis of biopsy slices stained with Hematoxylin and Eosin (HE). Since few years, novel and non-destructive biophotonic approaches are developed in order to achieve spectral histology as a complement to conventional histology.

Infrared (IR) spectroscopy has been shown to be a potential candidate to achieve spectral histology for cancer diagnosis. This technique is based on the interaction of an incident IR light beam with a point on the analyzed sample. When the energy of an incident photon is equal to the energy of a vibrational mode of the sample, this photon is absorbed by the sample, and a decrease of the transmitted light intensity is recorded. These intensity losses are recorded at different wavelengths (or wavenumbers) to give the absorbance spectrum of the acquisition point. A spectrum is thus a real molecular fingerprint giving information about the structure and metabolism of biomolecules at the acquisition point. IR spectroscopy has already been used to discriminate between different kind of tumors and also between healthy and tumoral tissues [1].

By repeating this operation with a scanning of the sample surface, a hyperspectral dataset is recorded. This huge amount of data requires the use of clustering methods to achieve spectral histology [2, 3], i.e. for the histopathological recognition of the sample structures, especially to highlight some structures invisible for the pathologist on HE histological images. Hierarchical analysis and K-Means (KM) clustering have been shown as very efficient to retrieve the analyzed tissue structures and to localize tumors [3, 4]. However, each spectrum belongs to only one cluster with these techniques.

Fuzzy C-Means (FCM) [5] seems to be better adapted since each spectrum can be attributed to all clusters thanks to membership values ranging from 0 to 1. This property is biologically relevant since transitional zones between tissue structures or even heterogeneities can be highlighted, while it is not always the case on the HE images.

FCM is controlled by two parameters, i.e. the number of clusters C and the fuzziness parameter m, that must be fixed by the user, and from which will depend the quality of the clustering results. When applied to IR spectroscopy, C is usually empirically chosen, whatever the clustering method used [3, 4], based on biological knowledge on the sample composition. On the contrary, m is classically fixed to 2 [3].

Some solutions have been developed to answer this parameterization problem. First, the automatic selection of C has been addressed by the so-called validity indices which optimize a criterion [6]. Second, *m*-rules estimate m by the optimization of a criterion [7-9] or by using empirical rules [10]. However, validity indices and *m*-rules have been developed independently from each other. Validity indices assume that m=2, while some applications on real world dataset show that this choice can lead to misclutering since it depends on the structure of the data. Moreover, some *m*-rules suppose that C must be *a priori* known.

© EURASIP, 2012 - ISSN 2076-1465

The development of methods able to simultaneously determine C and m is thus crucial. Recently, a new method, named FCM-Redundant Based Algorithm (FCM-RBA) has been proposed to solve this problem [11]. This method is based on the removal of redundant clusters. The first aim of this paper is to propose a faster version of this heuristic method. The second aim is to compare the efficiency of classical validity indices, *m*-rules and our method on IR images of human skin tumors.

2. FUZZY CLUSTERING TECHNIQUE

FCM [5] is a famous clustering algorithm which can be seen as a generalization of the well-known KM algorithm by introducing a fuzziness parameter m. This parameter must be higher than 1. Indeed, a value of m close to 1 makes FCM equivalent to the crisp KM clustering. On the contrary, as m increases, the fuzziness of the results increases as well. If m tends to infinity, the estimated clusters are identical.

With FCM, each IR spectrum can belong to every cluster with a membership coefficient ranging between 0 and 1. The only constraint is that the sum of these coefficients must be equal to 1 for each spectrum.

Intermediate spectra could thus be highlighted as their membership coefficients are equally distributed between several clusters. This kind of dataset clustering emphasizes a new type of information like transitional zones between two or more clusters. Such transitions are very informative here since they could match to a tumor progression into a surrounding healthy tissue and/or tumor heterogeneity.

In order to highlight those zones, it is necessary to determine the optimal values of C and m. That's why cluster validity indices and *m*-rules have been developed.

3. CLUSTER VALIDITY INDICES

Cluster validity indices estimate the optimal value of C by looking for the value of C that optimizes a criterion [6]. To construct these validity indices, several FCM with different values of C must be computed, being very time-consuming.

The first validity indices only consider the membership coefficients estimated by FCM [6]. We can cite the partition coefficient (V_{PC}) based on the sharing of objects between clusters, the partition entropy (V_{PE}) based on the fuzziness of the estimated partition, the modified partition coefficient (V_{MPC}) which is a transformation of V_{PC}, the Windham proportion exponent (V_{WPE}) based on the overlapping between clusters, the Kim-Kim-Lee-Lee coefficient (V_{KKLL}) based on the sharing between clusters, and the fuzzy partition measure (V_P) based on compactness and separation notions [6].

To overcome the observed sensitivity of these indices to m and their monotonic behavior in function of C, new validity indices were developed using membership coefficients and centroids [6]. A majority of these indices are constructed by melting two approaches, inter-cluster distances, which should be as biggest as possible, and the intra-cluster distances, which are preferred when they are as smallest as possible. The most efficient indices in the scientific literature are the following ones [6]: the Fukuyama-Sugeno index (V_{FS}), the Xie-Beni index (V_{XB}), the Kwon index (V_{Kw}), the Tang index (V_T), the partition coefficient and exponential separation (V_{PCAES}), the validity index of Tsekouras and Sarimveis (V_{VI}), the compose within and between scattering (V_{CWBS}), the Wang-Sun-Jiang index (V_{WSJ}), the Pakhira-Bandyopadhyay-Maulik index (V_{PBMF}), the separation and total compactness index (V_{STC}), the fuzzy separation and compactness index (V_{FSC}). Other indices are based on the hypervolume and the density (the fuzzy hypervolume index (V_{FHV}), the partition density index (V_{PD}) and the average partition density (V_{APD})), the granularity-dissimilarity measure (V_{GD}), the ratio of the overlapping to the separation (V_{OS}) or an optimality test (V_{OT}). Due to the lack of place, the interested reader can refer to [6] and references therein for a detailed presentation of these validity indices.

4. M-RULES

The aim of *m*-rules is to find the value of m which optimizes a criterion. *m*-rules can be decomposed into two groups. The *m*-rules belonging to the first group only use the data to assess the optimal value of m. Among these methods can be cited the Dembélé-Kastner empirical rule (R_{DK}) which is based on the computation of the variation coefficient, or the theoretical β rule (R_β) which computes upper bound for m.

The second group is composed of rules based on the results of the FCM algorithm, such as the fuzzy decision optimization rule (R_{FDO}) which maximizes the intersection between a fuzzy goal and a fuzzy constraint, or the Okeke-Karnieli rule (R_{OK}) which minimizes the reconstruction norm of the dataset with the FCM results.

5. FCM-RBA

In [11], a new heuristic method, named FCM-RBA, determining automatically the best values of C and m was proposed. Its main idea is to decrease the number of clusters if at least two clusters are similar for a given m.

The similarity here is measured by the Pearson correlation coefficient between the membership values. T is a new introduced parameter which represents a threshold for the correlation coefficients. It takes its values between 40% and 95%, with a step of 5%. If there is redundancy between two clusters, i.e. their correlation coefficient is superior to T, then one of them becomes useless.

In this article, FCM-RBA is modified in order to improve the computational time. First, an interruption inside the FCM algorithm itself occurs when the current results, during successive iterations, present redundancy between at least 2 clusters. Second, FCM results being independent on T, if, for a previous T value, a FCM clustering has already been computed for a given setting of C and m, then it is not necessary to

compute it again.

This redundant method proceeds as follows:

1. Initialization of parameters. C is empirically initialized at 15, m at 1.1 and T at 0.4.
2. Calculation of FCM (C, m, T).
3. Interruption of FCM, allowing a gain in time complexity, if redundant clusters are detected. C is revalued as follows: $C = C-1$, and FCM (C, m, T) is recomputed by going to step 2.
4. If no redundancy is detected in step 3, the best value of C is recorded into $C_{best}(m, T)$.
5. All those calculations are repeated from the second step with a new value of m, $m = m + 0.1$ and initialization of C as $C = C_{best}(m-0.1, T) + 2$, if the floor of 15 is respected, otherwise $C=15$. It is interesting to mention that this initialization of C induces a considerable gain in time of execution.
6. When all the values of m are gone over, all those calculations are run again with a new value of T as $T = T + 0.05$, till T is equal to 0.95.

For each T, a curve $C_{best}(m, T)$ is thus estimated. The interpretation of these curves to find the best m and C is left to section 6.4.

6. RESULTS AND DISCUSSION

6.1. Sample description

The previously described methods have been applied on an IR hyperspectral image acquired on a basal cell carcinoma (BCC) section. The image is composed of 104x61 pixels. In each pixel, a spectrum composed of 451 wavenumbers uniformly distributed between 900 and 1800 cm^{-1} was recorded. The spectra are considered as the objects to be clustered.

Figure 1 shows the HE image of the section adjacent to the one analyzed by IR imaging and annotated by a pathologist who is able to distinguish 4 different structures: stratum corneum (1), epidermis (2), dermis (3) and tumor nests (4). The dermis is known to be heterogeneous and can thus be decomposed into several clusters [4, 11].

6.2. Validity indices and m=2

We run validity indices with m=2 as traditionally described in literature. The estimated optimal numbers of clusters C_{opt} are presented in the second column of table 1. The majority of the validity indices estimates C_{opt}=2. Figure 2 shows the clusters estimated by FCM with such parameters. Cluster 2 represents mainly the dermis, while the epidermis and the tumor nests are merged in cluster 1. For 3 clusters such as estimated by V_{VI}, the results are the same except that the dermis is divided into two clusters (data not shown). The other validity indices are optimized for $C_{opt}\geq 11$ which induces a large number of redundant clusters (data not shown). FCM run with 11 clusters, such as estimated by V_T, leads to 4 identical clusters representing both tumor nests and epidermis (cluster 1, figure 3), 5 identical clusters of dermis (cluster 5, figure 3), and two supplementary different clusters of other dermis structures (clusters 10 and 11, figure 3). A m-value equal to 2 thus induces too much fuzziness for this spectral dataset in the FCM model and tumor nests cannot be separated from the epidermis. It is thus necessary to adapt the m-value to the studied dataset thanks to the use of the m-rules.

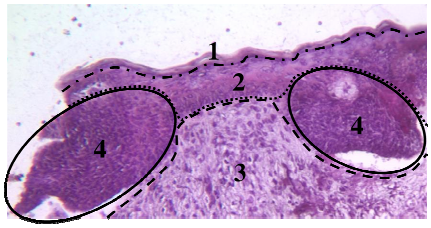


Figure 1: HE histological image of the BCC sample.

Validity indices	C_{opt} for m=2	C_{opt} for m=1.1	C_{opt} for m=1.3	C_{opt} for m=1.9	C_{opt} for m=4.4
V_{PC}	2	2	2	2	2
V_{PE}	2	2	2	2	2
V_{WPE}	2	2	2	2	2
V_{MPC}	2	4	5	2	3
V_{KKLL}	2	10	5	2	2
V_{P}	2	4	5	2	3
V_{FS}	15	14	13	15	15
V_{XB}	2	4	3	2	10
V_{Kw}	2	4	3	2	3
V_{T}	11	2	2	3	2
V_{FSC}	2	2	2	2	2
V_{FHV}	2	2	2	2	2
V_{APD}	2	2	2	2	2
V_{PD}	2	2	3	2	2
V_{PCAES}	2	2	2	2	8
V_{VI}	3	4	5	3	15
V_{CWBS}	14	4	5	11	6
V_{WSJ}	14	4	5	13	10
V_{PBMF}	15	2	2	15	15
V_{STC}	2	3	3	2	3
V_{GD}	2	4	3	2	3
V_{OT}	2	2	2	2	2
V_{OS}	2	5	5	2	11

Table 1: Optimal number of clusters C_{opt} estimated by the validity indices for different values of the fuzziness parameter m.

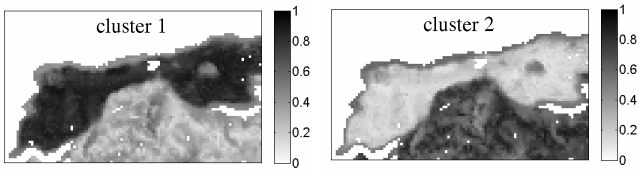


Figure 2: Membership values estimated by FCM for C=2 and m=2.

m-rules	R_{DK}	R_{OK}	R_{FDO}	R_{β}
m_{opt} for C=5	1.1	1.3	1.9	4.4

Table 2: Optimal m-value m_{opt} estimated by the m-rules.

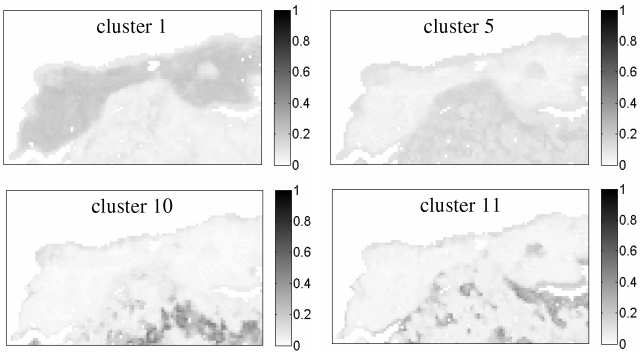


Figure 3: Membership values estimated by FCM for C=11 and m=2.

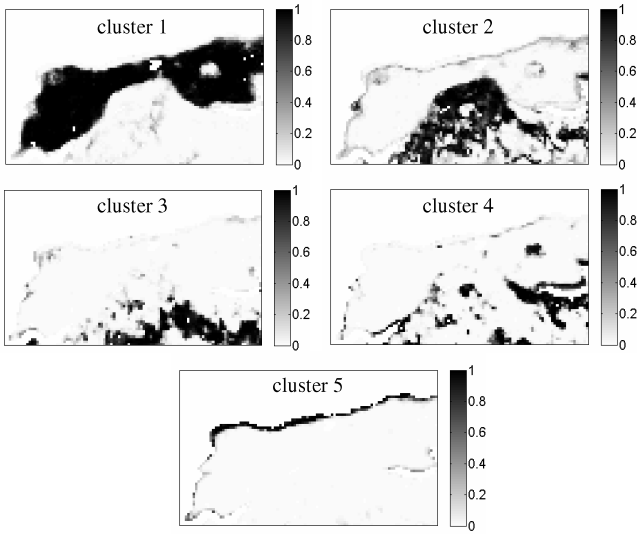


Figure 4: Membership values estimated by FCM for C=5 and m=1.3.

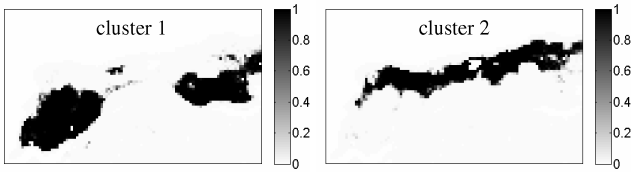


Figure 5: Membership values estimated by FCM for C=10 and m=1.1. Only the first 2 clusters are shown.

6.3. Validity indices and *m*-rules

R_{FDO} and R_{OK} being dependent on the number of clusters, we applied these rules by assessing that C=5 such as estimated by FCM-RBA and validated by a pathologist in another study [11]. The optimal m-values m_{opt} estimated by the m-rules are presented in table 2.

The optimal number of clusters C_{opt} was estimated by each validity index for the optimal m-value m_{opt} assessed by each m-rule. The results are pooled in columns 3 to 6 of table 1.

For m_{opt}=4.4 as estimated by R_β, it is clear that the results will be fuzziest than those estimated with m=2. For m_{opt}=1.9 as estimated by R_{FDO}, the same results as m=2 are obtained.

For m_{opt}=1.1 as estimated by R_{DK} or m_{opt}=1.3 as estimated by R_{OK}, the majority of the validity indices is optimized for $C_{opt}\leq$5. Figure 4 shows the FCM results for C_{opt}=5 and m=1.3 as estimated by R_{OK}. Tumor nests and epidermis are still merged (cluster 1). Clusters 2, 3 and 4 represent different dermis structures. Cluster 5 is the stratum corneum.

The other validity indices estimated $C_{opt}\geq$10. For such number of clusters, the tumor nests and the epidermis are isolated on different clusters as can be seen on figure 5 (clusters 1 and 2 respectively). However, the dermis is divided on 5 clusters, the stratum corneum on 2 clusters, and one cluster is composed of pixels at the interfaces stratum corneum/epidermis-tumor and epidermis-tumor/dermis (data not shown).

Only V_{KKLL} combined with R_{DK}, and V_{FS} combined with R_{DK} or R_{OK} are able to retrieve clusters associated to the tumor nests. However, the number of clusters necessary to isolate the tumor nests is high, and the optimal m-values estimated by R_{DK} and R_{OK} are near from 1, which means that the obtained results are comparable to those obtained with KM in another study [4] on the same sample with an empirical estimation of the number of clusters.

These results thus suggest that m-rules and validity indices are not adapted to this kind of data.

6.4. FCM-RBA

The curves $C_{best}(m,T)$ for the 12 different values of T obtained by FCM-RBA are presented on Figure 6. These curves rapidly decrease till a break which has been interpreted as a stabilization of the FCM solution. Thus, the best values of the parameters C and m are considered as being pointed out by the first break of each curve. Indeed, 7 curves present their curve break for C=5 and m=1.6, and 5 for C=5 and m=1.5 or 1.7. It must be known that making a clustering with a variation of 0.1 on the m-value doesn't change much the FCM results. Consequently, the best parameters are chosen as C=5 and m=1.6. Figure 7 presents the gray scale images obtained by FCM run with those parameters. Estimated clusters represent all the different tissue structures of the biopsy which were observed by the pathologist: the dermis, divided into three clusters because of its heterogeneity, the epidermis and the tumor nests. The interested reader can refer to [4] in order to access to the KM results obtained on this biopsy.

6.5. Discussion

The majority of validity indices used in this study is initialization dependent. On the contrary, our new method is more robust against initialization because the determination of the relevant FCM parameters is based on several curve breaks (one for each threshold value) and not on a unique maximum or minimum value such as the validity indices.

Furthermore, our method has been developed in order to simultaneously estimate m and C, while validity indices

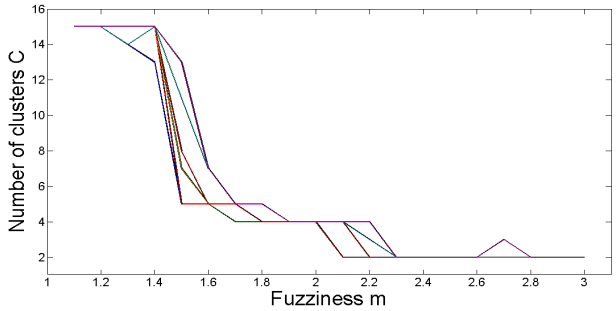


Figure 6: Curves $C_{best}(m,T)$ for the 12 different values of T estimated by FCM-RBA allowing identification of C_{opt} and m_{opt}.

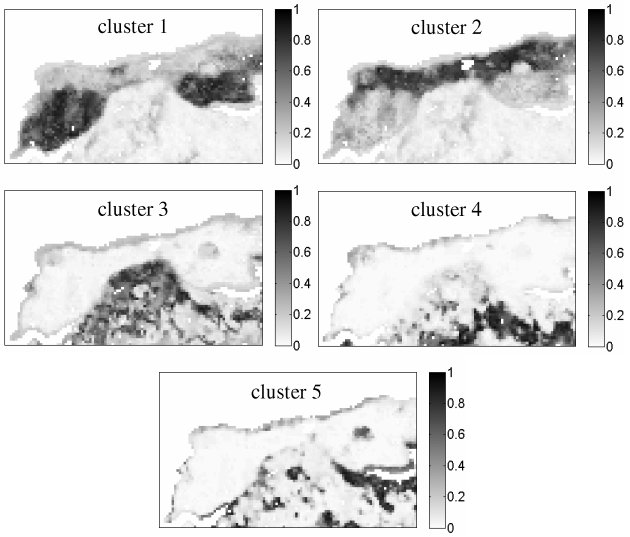


Figure 7: Membership values estimated by FCM for the parameters values given by FCM-RBA.

and *m*-rules have been developed ones independently from the others.

The main improvement is that, compared to the algorithm proposed in [11], the algorithm presented in this article is 8 times faster. This is due to a considerable diminution of the number of computed FCM. Indeed, many irrelevant settings of FCM are avoided, and interruption of FCM is imposed when redundant clusters are estimated.

These results were also validated on a group of 15 other biopsies representing several types of skin tumors (results not shown here due to the lack of place).

7. CONCLUSION

FCM is a clustering method that is well adapted to the analysis of IR spectral images acquired on human skin tumors. The FCM parameters, i.e. the number of clusters and the fuzziness parameter, must be carefully chosen. Several validity indices and *m*-rules have been proposed in the literature to overcome this selection problem. They have been compared in this paper to a new heuristic method based on the redundancy of clusters estimated by FCM. Our proposed method, named FCM-RBA, leads to better results in term of quality than those obtained with the validity indices combined to the *m*-rules. Indeed, this method is qualitatively more efficient as shown on a real dataset. Each cluster represents a unique tissue structure composing the biopsy, enabling an easy localization and characterization of the tumor. Moreover, the heterogeneity and the possible transitions between tissue structures could be highlighted, which cannot be realized with KM algorithm or, in most of cases, with a classical HE image analyzed by a pathologist. That's the reason why this new kind of information is of very important interest from a biomedical point of view.

8. REFERENCES

[1] R.K. Dukor, *Handbook of Vibrational Spectroscopy*, John Wiley and Sons Ltd., New York, 2001.

[2] B. Bird, M. Miljkovic, M. J. Romeo, J. Smith, N. Stone, M. W. George, and M. Diem, "Infrared microspectral imaging: distinction of tissue types in axillary lymph node histology," *BMC Clinical Pathology*, vol. 8, n°8, 2008.

[3] P. Lasch, W. Haensch, D. Naumann, and M. Diem, "Imaging of colorectal adenocarcinoma using FT-IR microspectroscopy and cluster analysis," *Biochimica et Biophysica Acta*, vol. 1688, pp. 176-186, 2004.

[4] 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*, vol. 133, pp. 197-205, 2008.

[5] J.C. Bezdeck, *Pattern recognition with fuzzy objectives function algorithms*, Plenum Press Ed, New York, 1981.

[6] W. Wang, and Y. Zhang, "On fuzzy cluster validity indices," *Fuzzy Sets and Systems*, vol.158, pp. 2095-2117, 2007.

[7] J. Yu, Q. Cheng, and H. Huang, "Analysis of the weighting exponent in the FCM," *IEEE Transactions on Systems, Man, and Cybernetics–Part B: Cybernetics*, vol. 34, pp. 634–639, 2004.

[8] H. Choe, and J. B. Jordan, "On the optimal choice of parameters in a Fuzzy C-Means algorithm," *Proceedings of the IEEE International Conference on Fuzzy Systems*, pp. 349–354, 1992.

[9] F. Okeke, and A. Karnieli, "Linear mixture model approach for selecting fuzzy exponent value in Fuzzy C-Means algorithm," *Ecological Informatics*, vol. 1, pp. 117–124, 2006.

[10] D. Dembélé, and P. Kastner, "Fuzzy C-Means method for clustering microarray data," *Bioinformatics*, vol. 19, pp. 973–980, 2003.

[11] 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*, vol. 91, pp. 799-811, 2011.

Article #II

Publication n°2

A new method for the simultaneous estimation of the number of clusters and the fuzziness parameter in FCM

T. Happillon, D. Sebiskveradze, V. Vrabie, O. Piot, P. Jeannesson, M. Manfait, C. Gobinet.

Article à soumettre à Pattern Recognition

Présentation de la publication n°2

Contexte

Comme expliqué dans le préambule de la publication n°1, un point faible de l'algorithme de classification non supervisée Fuzzy C-Means réside dans la subjectivité de son paramétrage. En effet, le nombre de classes *C* et la valeur du degré de flou *m* sont à déterminer par l'utilisateur en amont de toute classification. Afin d'optimiser le choix des valeurs de ces paramètres, différents outils ont été développés : les indices de validité et les règles en *m*. Cependant, dans la publication n°1, nous avons montré que de tels techniques sont incapables d'estimer des partitions pertinentes lorsque les FCM sont appliqués à des données spectrales infrarouge acquises sur des coupes de cancers cutanés. Dans la publication n°1, une méthode alternative a donc été proposée, appelée FCM-RBA amélioré, permettant d'estimer à la fois C et m. Appliquée sur 16 images spectrales infrarouge acquises sur des coupes de cancers cutanés, les partitions estimées par cette méthode étaient en parfait accord et même complémentaires à des analyses anatomo-pathologistes réalisées en parallèle.

Objectif

Bien que les résultats obtenus par notre méthode FCM-RBA amélioré sur des données spectrales infrarouge fussent de très bonne qualité, cette technique n'est pas efficace lorsqu'elle est appliquée à d'autres types de données. Ce manque de généralisation s'explique par la mesure de similarité entre deux classes définie dans notre algorithme par le coefficient de corrélation de Pearson. Or, cette mesure de similarité est sensible aux données aberrantes (outliers) dans le sens où elle ne détecte pas de corrélation entre deux classes identiques qui diffèrent uniquement d'un objet.

L'objectif de ce travail a donc été d'étendre l'applicabilité de cet algorithme amélioré, en modifiant la mesure de similarité. L'algorithme résultant de l'ensemble des modifications apportées à FCM-RBA est nommé Fuzzy C-Means Upgraded Redundant Based Algorithm (FCM-URBA). L'efficacité de ce nouvel algorithme a ensuite été testée sur des jeux de données synthétiques et réels bien connus de la littérature, puis comparée aux résultats obtenus par des indices de validité récents.

Matériels et Méthodes

La mesure de similarité choisie est une mesure de similarité floue qui n'est pas sensible aux données aberrantes et mieux adaptée à la comparaison de résultats issus de classifications non supervisées floues. De plus, l'algorithme FCM-URBA utilisant cette mesure de similarité n'a pas tendance à surestimer le nombre de classes comme le faisait l'algorithme FCM-RBA.

Les jeux de données utilisés dans cette étude sont au nombre de 9 et peuvent être subdivisés en deux catégories. Les 5 premiers sont des jeux de données simulées à deux dimensions et contenant des objets générés *via* une distribution gaussienne ou uniforme. La seconde catégorie concerne 4 jeux de données réelles, Ecoli, Iris, Serum et WBC (Wisconsin Breast Cancer), bien connus dans la littérature pour étudier les performances des algorithmes de classification.

Les indices de validité testés dans cette étude sont STCI (Separation and Total Compactness Index), SCI (Separation and Compactness Index), OSI (Overlap and Separation Index), FSI (Fuzzy Silhouette Width) et GIISI (Generalized Intra-Inter Silhouette Index). De plus, une méthode très récente d'estimation séquentielle de *C* et *m*, appelée méthode de Schwämmle-Jensen (SJM), a également été testée, en comparaison des méthodes FCM-RBA et FCM-URBA.

La comparaison des résultats n'est plus faite de façon subjective comme dans la publication n°1, mais de façon objective, en utilisant une méthode supervisée d'évaluation de la meilleure partition estimable par FCM sur un jeu de données. Cette méthode s'appelle la méthode de variation de l'information.

Résultats

Les résultats obtenus sur les jeux de données simulées montrent une certaine hétérogénéité dans les performances des différentes méthodes testées. En effet, les indices STCI, OSI, FSI et GIISI, estiment correctement le nombre de classes de 3 jeux sur 5, alors que l'indice SCI évalue correctement le nombre de classes de 4 jeux..

Les méthodes SJM et FCM-RBA, estiment correctement le nombre de classes de 2 jeux. Seule la méthode FCM-URBA estime correctement la structure des 5 jeux de données simulées.

Concernant les jeux de données réelles (*i.e.* Ecoli, Iris, Serum et WBC), les indices SCI, FSI et GIISI n'estiment correctement le nombre de classes que d'un seul jeu. L'indice OSI et l'algorithme FCM-RBA en partitionnent correctement 2, et l'indice STCI ainsi que la méthode SJM estiment correctement le nombre de classes de 3 jeux. Finalement, seul l'algorithme FCM-URBA partitionne correctement l'ensemble des 4 jeux de données réelles.

En ce qui concerne les valeurs du degré de flou, la méthode SJM a tendance à retourner de faibles ou fortes valeurs de *m*, c'est-à-dire en dehors de la plage de valeurs préconisée dans la littérature, à savoir entre 1,5 et 2,5. La conséquence de ces estimations est l'obtention de classifications dites « dures » (comparables aux K-Means) ou excessivement floues (perte de l'information). Cette observation peut également être faite concernant la moitié des estimations de *m* par la méthode FCM-RBA. En revanche, l'algorithme FCM-URBA estime des valeurs du degré de flou comprises entre 1,5 et 2,6 pour l'ensemble des 9 jeux de données testés, ce qui est conforme aux recommandations de la littérature.

Conclusion

L'estimation automatique du nombre de classes *C* et du degré de flou *m* est un problème important de la classification floue, et c'est pour cette raison qu'ont été créés les indices de validité. Cependant, ces indices ne permettent qu'une estimation de *C* et sont couramment calculés avec une valeur de *m* choisie par défaut égale à 2. Or, il a été montré que non seulement ce choix pouvait mener à des erreurs de classification, mais également que les valeurs de *C* et *m* étaient intimement liées. Dans cette étude, un nouvel algorithme, appelé FCM-URBA, a été proposé et a montré sa supériorité au regard des résultats obtenus sur 9 jeux de données réels et simulés, en comparaison des résultats obtenus par 5 indices de validité récents et deux autres méthodes estimant à la fois *C* et *m*.

A new method for the simultaneous estimation of the number of clusters and the fuzziness parameter in FCM

Teddy Happillon[1], David Sebiskveradze[1], Valeriu Vrabie[2], Michel Manfait[1], Cyril Gobinet[1,*]

Abstract

An open problem in fuzzy clustering is the estimation of its two parameters, i.e. the number of clusters and the fuzziness parameter. Validity indices have been developed to estimate the number of clusters, while the fuzzy parameter is commonly fixed to 2. Some recent algorithms, named $K-m$ methods in this paper, have been proposed to also assessed the fuzziness parameter. However, these methods are sequential since they estimate first the fuzziness parameter, and then the number of clusters. This article aims at proposing a new method to automatically and simultaneously estimate these two parameters. This method is based on the observation that redundant clusters are estimated when a too high number of clusters is fixed for a given fuzziness parameter. The proposed algorithm is thus dedicated to determine the maximum number of non-redundant clusters for several consecutive fuzziness

*Corresponding author

Email addresses: teddy.happillon@univ-reims.fr (Teddy Happillon), david.sebiskveradze@univ-reims.fr (David Sebiskveradze), valeriu.vrabie@univ-reims.fr (Valeriu Vrabie), michel.manfait@univ-reims.fr (Michel Manfait), cyril.gobinet@univ-reims.fr (Cyril Gobinet)

[1]Unité MEDyC, CNRS FRE 3481, Université de Reims Champagne-Ardenne, 51 rue Cognacq-Jay, 51096 Reims Cedex, France

[2]CReSTIC, Université de Reims Champagne-Ardenne, Moulin de la Housse, BP 1039, 51687 Reims Cedex 2, France

parameter values. The redundancy is here computed by a fuzzy similarity measure. Compared to recently developed validity indices and $K-m$ methods applied on 5 simulated and 4 real datasets, this new algorithm is shown to be the best method for the estimation of the number of clusters. Furthermore, the estimated fuzziness parameter value is in accordance with the range recommended by literature.
Keywords: Fuzziness parameter, fuzzy clustering, number of clusters, redundancy, validity indices

1. Introduction

The development of high-quality acquisition systems, in biomedical imaging for instance, has induced a natural increase of the available amount of acquired data. The automation of the data mining is thus become an important and necessary step for the exploitation of these data. Particularly, unsupervised classification (also called clustering) methods have shown their efficiency to simply and rapidly analyze such data with few parameters managed by the user. This kind of method makes few assumptions about the data, especially on their labelling. Among these techniques, Fuzzy C-Means (FCM) [5] has rapidly gained popularity because of its unquestionable advantages (rapidity, less sensitivity to initialization, possible distribution of an object on all clusters) compared to hard clustering techniques such as Hierarchical Cluster Analysis (HCA) [36] and K-Means [21].

FCM is an iterative partitional method which divides the dataset into groups to simultaneously minimize the intra-group variability and maximize the inter-group variability. Suppose a dataset $X = [x_1, x_2, \ldots, x_N] \in \mathbb{R}^{P \times N}$

which is considered to be decomposed into K clusters, where each $x_i \in \mathbb{R}^P$ is an object (*e.g.* a pixel, a spectrum, ...) and N is the number of objects composing the dataset. The aim of FCM is to minimize the following objective function:

$$J_m(U,V) = \sum_{i=1}^{N}\sum_{j=1}^{K} u_{ij}^m \|x_i - v_j\|^2. \quad (1)$$

In this function, u_{ij} is the membership value of the $i-th$ object x_i to the $j-th$ cluster represented by its prototype $v_j \in \mathbb{R}^P$. The matrices $U \in \mathbb{R}^{N\times K}$ and $V \in \mathbb{R}^{P\times K}$ are respectively constructed from the u_{ij} values and v_j vectors. The following constraints are imposed to FCM for each object: $0 \leqslant u_{ij} \leqslant 1$ and $\sum_{j=1}^{K} u_{ij} = 1$. This last constraint means that each object x_i is split into pieces of different sizes (represented by the membership values u_{ij}) on the K clusters without exceeding a total membership of 1 (representing 100% of the object). This concept of membership value is thus ideal to reveal overlapping between clusters and thus to introduce the notion of uncertainty in object clustering. The following inequalities $0 < \sum_{i=1}^{N} u_{ij} < N$ insure that each cluster is not empty. m refers to the fuzziness parameter which manages the amount of fuzziness in the partitioning, and must be chosen such that $1 < m < +\infty$.

In [5], Bezdek showed that the repetition of the two following FCM update rules converges to a strict local minimum of J_m:

$$u_{ij} = \frac{1}{\sum_{k=1}^{K}\left(\frac{\|x_i - v_j\|}{\|x_i - v_k\|}\right)^{\frac{2}{m-1}}}, 1 \leqslant i \leqslant N, 1 \leqslant j \leqslant K \quad (2)$$

$$\text{and } v_j = \frac{\sum_{i=1}^{N} u_{ij}^m x_i}{\sum_{k=1}^{N} u_{ij}^m}, 1 \leqslant j \leqslant K. \tag{3}$$

If $m \longrightarrow 1$, the FCM result is equivalent to K-Means. If $m \longrightarrow +\infty$, each object is uniformly distributed on each cluster and thus all the clusters are identical.

Although it is qualified as unsupervised, FCM depends on two parameters, K and m, that must be fixed by the user. Traditionally, K is chosen based on *a priori* knowledge of the composition of the data or by successive and time consuming trial-and-error steps. When multidimensional dataset acquired on complex samples are manipulated, the adjustment of this parameter becomes extremely difficult. To resolve this problem, automatic methods named validity indices [33] have been developed to estimate the optimal number of clusters K.

The majority of the validity indices are known to be relatively insensitive to the fuzziness parameter m. This is true in cases where the clusters fulfill the validity index assumptions, *e.g.* compactness of clusters and separation between them. Nevertheless, numerous applications of FCM to real-life datasets [27, 34, 23, 8] assumed that the fuzziness parameter m is equal to 2 as recommended by Pal and Bezdek [26]. However, as soon as the clusters don't fulfill these assumptions, the validity indices can have an unpredictable behavior [11].

The development of numerical tools for the automatic selection of m and K is thus of great interest. Up to our knowledge, only two methods exist to reach this goal. The first one [14] is based on the increase of the m-value, for a fixed predefined K, till empty clusters are estimated by FCM applied

on the randomized dataset. Then, K is selected as the number of clusters which induces empty clusters by FCM applied on the dataset, with m fixed to the previously estimated value. The second method [30] can be seen as an updated and optimized version of the first one where the detection of empty clusters is replaced by the minimum centroid distance. This kind of methods will be referred as K-m methods in the rest of the paper. The main drawback of these methods is that the estimation of m and K is sequential. Thus, an estimation error on m may induce an estimation error on K.

It is thus of great interest to propose new methods able to simultaneously estimate m and K in order to minimize the estimation errors. Several studies have pointed out the generation of redundant clusters by FCM when the number of clusters is too high for a m-value fixed around 2 [11, 27, 35]. In a previous work [31], we made the same observations on infrared spectral images recorded on human skin tumoral biopsies. The redundant clusters had no physical interpretation according to the pathologists associated to this study. We developed a new empirical K-m methodology to automatically and simultaneously estimate m and K best suited to generate FCM results without redundant clusters for a given dataset. The proposed algorithm was successfully, but exclusively, applied on infrared spectral images [31], and revealed time-consuming. We propose in the following of this paper an algorithm based on a new redundant cluster criterion which improves the computational time and the accuracy of the results on real-world and simulated datasets.

The following of this paper is organized as follows. In section 2, recently proposed validity indices and K-m methods are briefly described. In section

3, the proposed algorithm is presented. Section 4 describes the experimental setup used to evaluate the proposed methods. Section 5 deals with the results obtained by the application of the new method, the validity indices and the K-m methods on simulated and real-world datasets. Section 6 proposes a discussion about the comparison of the proposed methods and their strengths and weaknesses. Finally, section 7 concludes the paper.

2. Previous works

2.1. Validity indices

The application principle of the validity indices is simple. First, the FCM partition is estimated for each $2 \leqslant K \leqslant K_{max}$ where K_{max} is the maximum number of clusters authorized by the user based on a priori knowledge of the dataset or on heuristic principle. Then, the validity index value is computed for each of these estimated FCM partitions. Then, the number of clusters $\widehat{K}$, which optimizes the validity index value, is considered as the number of clusters that gives the best FCM partitioning of the data.

The validity indices can be roughly divided into two groups. The first group is composed of validity indices computed exclusively thanks to the membership values. These indices have been criticized because they only relate to the membership values of the clustering without taking into account the structure of the data, because they have a monotonic behavior with K, and because they are reported as sensible to m [33].

To overcome these limitations, validity indices which rely on the FCM results (membership values and prototypes) and also on the data have been developed and actually constitute the second group. These validity indices

share the same construction ideas since they are all based on the maximization of the inter cluster separation and the minimization of the intra cluster compactness.

In the literature, more than 30 different validity indices can be listed. In this work, only the most recent are presented and tested because they have been shown as being better than the older validity indices which have been largely studied. The validity indices presented in the following of this section are constructed on the same model, i.e. the combination of a compactness measure and a separation measure.

2.1.1. Separation and total compactness index (STCI)

This index was developed by Bouguessa *et al* [6] on the basis of a previous work by Geva *et al* [15]. It is defined as the ratio between the fuzzy separation between clusters and the total compactness of the partition:

$$V_{STCI}(K) = \frac{\text{trace}\left(\sum_{i=1}^{N}\sum_{j=1}^{K} u_{ij}^{m}\left(v_j - \frac{1}{N}\sum_{l=1}^{N} x_l\right)\left(v_j - \frac{1}{N}\sum_{l=1}^{N} x_l\right)^{T}\right)}{\sum_{j=1}^{K}\text{trace}\left(\frac{\sum_{i=1}^{N} u_{ij}^{m}(x_i - v_j)(x_i - v_j)^{T}}{\sum_{i=1}^{N} u_{ij}^{m}}\right)} \tag{4}$$

This index has been shown as being free from the monotonic behavior in function of the number of clusters that is a recurrent problem in validity indices. It has also been proved as more robust to the variation of some parameters, such as the orientation or overlapping of clusters, compared to other validity indices.

The optimal value $\widehat{K}$ of K is obtained when V_{STCI} is maximized.

2.1.2. Separation and compactness index (SCI)

This index is based on the sum of a measure of compactness and a measure of separation [28] which are differently mathematically defined compared to STCI:

$$V_{SCI}(K) = \frac{Sep(K)}{\max_{2\leq L\leq K_{max}} Sep(L)} + \frac{Comp(K)}{\max_{2\leq L\leq K_{max}} Comp(L)} \tag{5}$$

where

$$Sep(K) = -\frac{2}{K(K-1)} \sum_{p\neq q=1}^{K} \sum_{i=1}^{N} K \min(u_{ip}, u_{iq}) \sum_{j=1}^{K} u_{ij} \log_a u_{ij} \tag{6}$$

is the separation or overlap measure defined in [19], and

$$Comp(K) = \sum_{i=1}^{N} \sum_{j=1}^{K} u_{ij}^m \|x_i - v_j\|^2 \tag{7}$$

is the total compactness of the fuzzy partition which is a compactness measure traditionally used for validity indices [13, 37].

Thanks to tests on simulated datasets, SCI has been shown as being robust to overlapping of clusters and insensitive to the shape, to the size and to the density of the clusters.

The optimal value $\widehat{K}$ of K is obtained when V_{SCI} is minimized.

2.1.3. Overlap and separation index (OSI)

This validity index is developed on a separation measure and an overlap measure between clusters. The two measures are based on an aggregation operation of membership values [20]. The theoretical definition of this index

is general and requires the choice of a T-norm couple such as standard, algebraic, Lukasiewicz, Hamacher or Dombi. In this article, we selected the standard T-norm couple since it is a non-parametrized T-norm. Furthermore the validity index defined with this T-norm couple has been shown as the best validity index in [20] when applied on simulated and real datasets.

The OSI based on the standard T-norm couple is defined as follows:

$$V_{OSI}(K) = \frac{1}{N}\sum_{i=1}^{N}\frac{u_{i(2)}}{u_{i(1)}} \tag{8}$$

where $u_{i(1)}$ and $u_{i(2)}$ are respectively the first and second largest elements of $u_{i.} = [u_{i1}, u_{i2}, \ldots, u_{iK}]$.

This index has been shown as insensitive to the value of the fuzziness parameter m and robust to the structure of the data.

The optimal value $\widehat{K}$ of K is obtained when V_{OSI} is minimized.

2.1.4. Fuzzy silhouette index (FSI)

In [7], a fuzzy generalization of the well-known Average Silhouette Width Criterion [29], developed for hard clustering, was proposed as:

$$V_{FSI}(K) = \frac{\sum_{i=1}^{N}(u_{i(1)} - u_{i(2)})^{\alpha} s_i}{\sum_{i=1}^{N}(u_{i(1)} - u_{i(2)})^{\alpha}} \tag{9}$$

where $u_{i(1)}$ and $u_{i(2)}$ are respectively the first and second largest elements of $u_{i.}$ such as for V_{OSI}, $\alpha \geq 0$ is a weighting coefficient which control the weight of objects in overlapping areas, and

$$s_i = \frac{b_i - a_i}{\max(a_i, b_i)} \tag{10}$$

being the silhouette of object x_i, with a_i being the mean distance of object x_i to the other objects belonging to the same cluster than x_i (thus the compactness distance), and b_i being the minimum of the mean distance of x_i to all the objects belonging to another cluster than x_i (thus the separation distance).

Following the application setup used in [7], the α parameter is set to 1 is the rest of this article.

This index has been shown to be robust to clusters overlapping, to different sizes and densities of clusters.

The optimal value $\widehat{K}$ of K is obtained when V_{FSI} is maximized.

2.1.5. Generalized intra-inter silhouette index (GIISI)

A generalization of FSI has been recently proposed in [29] and is defined by:

$$V_{GIISI}(K) = \frac{\sum_{i=1}^{N} s_i}{N} \tag{11}$$

where s_i is the silhouette of object x_i defined by equation (10) with:

$$a_i = \min_{1 \leq k \leq K} \frac{\sum_{j=1}^{N} \min(u_{ik}, u_{jk}) \|x_i - x_j\|}{\sum_{j=1}^{N} \min(u_{ik}, u_{jk})} \tag{12}$$

being the compactness distance, and

$$b_i = \min_{1 \leq r < s \leq K} \frac{\sum_{j=1}^{N} \max\Big(\min(u_{ir,u_{js}}), \min(u_{is}, u_{jr})\Big) \|x_i - x_j\|}{\sum_{j=1}^{N} \max\Big(\min(u_{ir,u_{js}}), \min(u_{is}, u_{jr})\Big)} \tag{13}$$

being the separation distance.

This index has been shown to outperform FSI and traditional validity

indices.

The optimal value $\widehat{K}$ of K is obtained when V_{GIISI} is maximized.

2.2. K-m methods

2.2.1. Schwämmle-Jensen method (SJM)

As stated in the introduction, SJM is an updated and optimized version of the previous work by Futschik [14]. That is the reason why only SJM is described in this article.

The objects of a randomized version of a dataset being almost homogeneously distributed, no structure into the data exists. The application of FCM is thus supposed to return highly overlapped clusters. However, if the m-value is too small, then the algorithm will always find a structure into the data, such as the K-Means algorithm. By increasing the m-value, the membership values of the objects will be more distributed between clusters. For a m-value, named m-threshold, specific to the data structure, two clusters with identical prototypes will be estimated, thus revealing the real lack of data structure due to the randomization of the objects. This m-threshold is of importance since it indicates the optimal value of m in order to find not randomly induced structure into a dataset. Furthermore, the independence of this m-threshold to the number of clusters has been experimentally demonstrated in [30]. This m-threshold is defined as the optimal m-value for the considered dataset.

Then, for this optimal m-value, the number of clusters is initialized to 2 and is progressively increased. For a specific number of clusters K_t, the minimal distance between clusters estimated by the FCM algorithm applied to the not randomized dataset for K_t clusters and $K_t + 1$ clusters will dras-

tically decrease, suggesting that additional clusters are useless and that the structure of the data can be completely described by K_t clusters, i.e. the optimal number of clusters is $\widehat{K} = K_t$.

These ideas have been developed in [30] which proposed a rapid and efficient algorithm able to sequentially estimate K and m on simulated and real datasets. First, the FCM algorithm is run for successively increased m-values till the minimal euclidean distance $d_{min}(\sqrt{N})$ between prototype pairs of $\sqrt{N}$ clusters is lower than a predefined threshold T_m, leading to the estimation of the optimal m-value $\widehat{m}$. Second, for $m = \widehat{m}$, for K varying from 2 to $\sqrt{N}$, the minimal euclidean distance $d_{min}(K)$ between prototype pairs of K clusters estimated by FCM is computed. Then, the number of clusters verifying:

$$\widehat{K} = \arg\max(d_{min}(K) - d_{min}(K+1)) \tag{14}$$

is computed ans considered as the optimal number of clusters.

2.2.2. Redundant based algorithm (RBA)

In [31], it has been observed that the application of the FCM algorithm with a too high number of clusters and a fixed m-value leads to the estimation of redundant clusters, i.e. the membership column vectors $u_{.i} = [u_{1i}, u_{2i}, \ldots, u_{Ni}]^T$ and $u_{.j} = [u_{1j}, u_{2j}, \ldots, u_{Nj}]^T$ of pairs of clusters i and j are highly correlated.

The basis of RBA is thus to find, for each m-value, the maximum number of clusters leading to non-redundant membership column vectors between each possible pair of clusters, i.e. leading to a correlation coefficient between

each pair of clusters that is inferior to a correlation threshold T. When m is around 1, FCM is equivalent to hard clustering and thus no redundant clusters are estimated, whatever the number of clusters. The maximum of non-redundant clusters is thus equal to the maximum number of clusters K_{max} authorized by the user. By increasing m, this situation remains till a specific m-value, from which the increase of m will introduces enough fuzziness to create fuzzy redundant clusters. The number of maximum non-redundant clusters drastically decreases till a plateau which can be interpreted as a stabilization of the fuzzy solution. The beginning of this plateau indicates the best number of clusters and the best m-value for the considered correlation threshold. By repeating this procedure for several correlation thresholds, a consensus can be found for the best numbers of clusters and the best m-values. The optimal number of clusters and the optimal m-value for the studied dataset can thus be simultaneously estimated.

This algorithm has been shown to be very efficient when applied on spectral infrared images acquired on human skin tumor samples [31]. However, it is very time-consuming and not efficient when applied on a panel of different datasets as shown in the section 5 of this article.

The limited applicability of this method to spectral infrared images is mainly due to the correlation coefficient, used to measure the redundancy between clusters, which is not adapted for fuzzy clustering. The computation time is mainly due to multiple computations of FCM for each (m, K, T) triplet.

In the next section, we propose an upgraded version of RBA in order to correct these problems by measuring the redundancy with a measure better

adapted to fuzzy clustering and by limiting the number of FCM computations.

3. Upgraded Redundant Based Algorithm (URBA)

3.1. Redundancy measure

Our algorithm being based on the redundancy between clusters, this notion will now be explained. The redundancy between two clusters i and j represents the amount of similar information shared by their respective membership column vectors $u_{.i}$ and $u_{.j}$. The redundancy is here quantified by a similarity measure represented by the following function:

$$\begin{aligned} f \quad : \quad & \mathbb{R}^N \times \mathbb{R}^N \longrightarrow [0,1] \\ & (u_{.i}, u_{.j}) \longmapsto f(u_{.i}, u_{.j}). \end{aligned} \tag{15}$$

If the two clusters represent exactly the same information, then $f(u_{.i}, u_{.j}) = 1$. On the contrary, if the two clusters represent completely different information, then $f(u_{.i}, u_{.j}) = 0$. Two clusters i and j are considered as redundant if their redundancy value $f(u_{.i}, u_{.j})$ is greater than a predefined threshold $T \in [0, 1]$.

The Redundancy Based Algorithm (RBA) proposed in [31] used the following redundancy measure:

$$f(u_{.i}, u_{.j}) = \max(0, r(u_{.i}, u_{.j})), \tag{16}$$

where $r(u_{.i}, u_{.j})$ is the Pearson correlation coefficient between $u_{.i}$ and $u_{.j}$:

$$r(u_{.i}, u_{.j}) = \frac{\sum_{k=1}^{N}(u_{ki} - \overline{u_i})(u_{kj} - \overline{u_j})}{\sqrt{\Big(\sum_{k=1}^{N}(u_{ki} - \overline{u_i})^2\Big)\Big(\sum_{k=1}^{N}(u_{kj} - \overline{u_j})^2\Big)}}, \tag{17}$$

where $\overline{u_i} = \frac{\sum_{k=1}^{N} u_{ki}}{N}$ and $\overline{u_j} = \frac{\sum_{k=1}^{N} u_{kj}}{N}$.

When applied to infrared spectral images of human skin tumor sections, this redundancy measure was proved efficient to estimate a couple $(\widehat{K}, \widehat{m})$ that optimally describes histological areas of the tissue sections by FCM [31].

However, the use of Pearson correlation coefficient leads to an over estimation of the real number of clusters for some datasets as it will be shown in section 5. This is due to a sensitivity of Pearson correlation coefficient to outliers.

The Upgraded Redundant Based Algorithm (URBA), described in this paper, is thus based on the following fuzzy similarity measure function [32] which is better suited for FCM results interpretation:

$$f(u_{.i}, u_{.j}) = \frac{\sum_{k=1}^{N} min(u_{ki}, u_{kj})}{\sum_{k=1}^{N} max(u_{ki}, u_{kj})}. \tag{18}$$

The principal characteristic of this similarity measure is that it is by construction only influenced by shared objects, contrary to the Pearson correlation coefficient based redundancy measure. If no objects are shared betweens clusters i and j, then $min(u_{ki}, u_{kj}) = 0, \forall k \in \{1, \ldots, N\}$. Thus, $f(u_{.i}, u_{.j}) = 0$ for completely different clusters. If two clusters i and j are completely identical, then $min(u_{ki}, u_{kj}) = max(u_{ki}, u_{kj})$, $\forall k \in \{1, \ldots, N\}$. Thus, $f(u_{.i}, u_{.j}) = 1$

for completely identical clusters.

3.2. Description of the algorithm

The proposed algorithm, named Upgraded Redundant Based Algorithm (URBA), is based on the computation of FCM for several (K, m) couples. For a given m, the algorithm will find the maximum number of clusters that leads to FCM results without redundant clusters, the redundancy being measured by the fuzzy similarity measure function of equation (18).

The algorithm is decomposed in the following steps.

(i) First, the threshold T is fixed to a value T_1, m to a value m_1, and K to a value K_{max} higher than the real number of clusters composing the dataset.

(ii) Then, the classical FCM algorithm [5] is run with the parameters (K, m). The redundancy $f(u_{.i}, u_{.j})$, $1 \leqslant i, j \leqslant K$ and $i \neq j$, is computed for all $\frac{K(K-1)}{2}$ cluster pairs. The set of redundancy values can thus be defined as $\mathcal{R} = \{R_1, \ldots, , R_l, \ldots, R_{\frac{K(K-1)}{2}}\} = \{f(u_{.1}, u_{.2}), \ldots, f(u_{.i}, u_{.j}), \ldots, f(u_{.K-1}, u_{.K})\}$. If for any l, $1 \leqslant l \leqslant \frac{K(K-1)}{2}$, a redundancy value is greater than the threshold, $R_l \geqslant T_1$, then the corresponding clusters are redundant and thus one is useless. The potential number of clusters can thus be decreased.

(iii) Step (ii) is repeated for $K = K - 1$, but for the same $m = m_1$ and $T = T_1$, until no redundant clusters are present. The remaining clusters are thus composed of distinct information in respect to the threshold T_1, and the maximum number of non-redundant clusters $K_{nr}(T_1, m_1)$ is thus reached for the couple (T_1, m_1). It has to be noticed that if redundant clusters are estimated whatever is the value of $K \in \{2, \ldots, K_{max}\}$, then there is only one cluster and $K_{nr}(T_1, m_1)$ is set to 1.

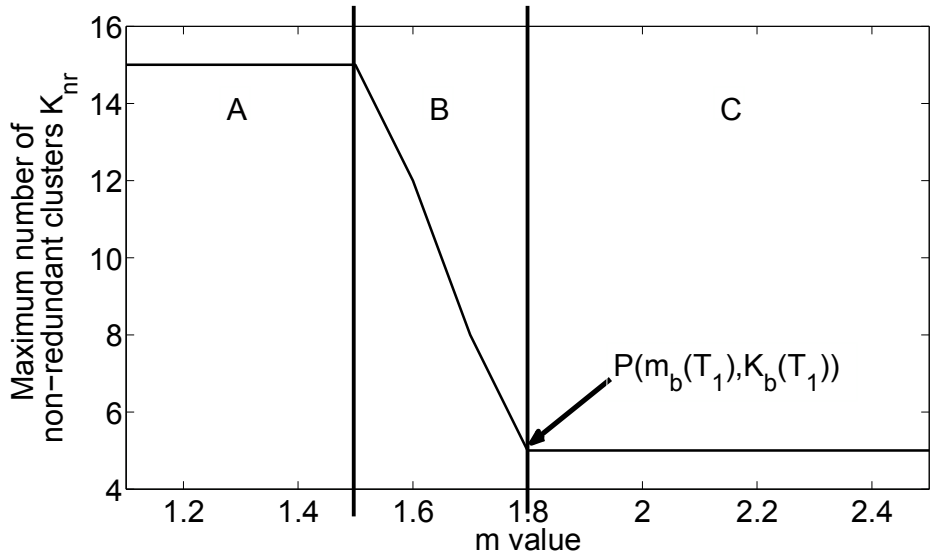


Figure 1: Caricatural shape of the maximum number of non-redundant clusters $K_{nr}(T_1, m)$ in function of the fuzzy parameter m for a specific threshold T_1. The point P represents the break point of coordinates $(m_b(T_1), K_b(T_1))$.

(iv) For the same threshold T_1, steps (i) to (iii) are repeated for each m-value belonging to the set $\mathcal{M} = \{m_1, \ldots, m_p, \ldots, m_M\}$ previously defined by the user. For example, the set can be chosen as the values between 1.1 and 3 with a step of 0.1. However, increasing the m-value means introduction of more fuzziness, and thus the estimation of more similar clusters by FCM. In order to save computational time, the procedure for the next m-value should begin with a number of clusters K initialized to the number of non-redundant clusters estimated for the previous m-value. However, to take into account the random initialization of the FCM algorithm, K is set to the number of non-redundant clusters estimated for the previous m-value plus two, i.e. for $m = m_p$ and $K = K_{nr}(T_1, m_{p-1}) + 2$.

Algorithm 1 Pseudo-code of URBA

$\mathcal{M} \leftarrow \{m_1, \ldots, m_p, \ldots, m_M\}$
$\mathcal{T} \leftarrow \{T_1, \ldots, T_q, \ldots, T_Q\}$
for $q = 1$ to Q **do**
  $K \leftarrow K_{max}$
  $T \leftarrow T_q$
  $BreakFlag \leftarrow 0$
  $p \leftarrow 1$
  while $BreakFlag = 0$ **do**
    $m \leftarrow m_p$
    $f_{max} \leftarrow 1$
    if $p \geqslant 2$ **then**
      $K \leftarrow K_{nr}(q, p-1) + 2$
      $K \leftarrow min(K, K_{max})$
    end if
    while $f_{max} \geqslant T$ **do**
      if $R\{K, m\}$ does not exist **then**
        $U\{K, m\} \leftarrow FCM(K, m)$
        $l \leftarrow 1$
        for $i = 1$ to $K - 1$ **do**
          for $j = i + 1$ to K **do**
            $R\{K, m\}_l \leftarrow f(u\{K, m\}_{.i}, u\{K, m\}_{.j})$
            $l \leftarrow l + 1$
          end for
        end for
      end if
      $f_{max} \leftarrow \max_l(R\{K, m\}_l)$
      if $f_{max} \geqslant T$ **then**
        $K \leftarrow K - 1$
        if $K = 1$ **then**
          $f_{max} \leftarrow T - 1$
        end if
      end if
    end while
    $K_{nr}(q, p) \leftarrow K$
    if $p \geqslant 2$ **then**

if $K_{nr}(q,p) < K_{max}$ **then**
if $K_{nr}(q,p) = K_{nr}(q,p-1)$ **then**
$K_b(q) \leftarrow K_{nr}(q,p)$
$m_b(q) \leftarrow m_{p-1}$
$BreakFlag \leftarrow 1$
end if
end if
end if
$p \leftarrow p+1$
end while
end for
for $k = 1$ to K_{max} **do**
$D(k) \leftarrow 0$
end for
for $q = 1$ to Q **do**
$D(K_b(q)) \leftarrow D(K_b(q)) + 1$
end for
$\widehat{K} \leftarrow \arg\max_k(D_k)$
$indices \leftarrow find(K_b = \widehat{K})$
$\widehat{m} \leftarrow \frac{\sum_{k=1}^{D(\widehat{K})} m_b(indices(k))}{D(\widehat{K})}$

(v) After application of steps (i) to (iv), the set of maximum number of non-redundant clusters $\mathcal{K}_{nr}(T_1, m) = \{K_{nr}(T_1, m_1), K_{nr}(T_1, m_2), \ldots, K_{nr}(T_1, m_M)\}$ is estimated for the threshold T_1. The caricatural shape of $\mathcal{K}_{nr}(T_1, m)$ in function of m is represented on figure 1. Three parts compose this shape. In part A, $K_{nr}(T_1, m)$ is equal to the maximum number of clusters K_{max} chosen by the user since FCM behaves as K-Means when m is around 1.1. There are thus no redundant clusters whatever is the number of clusters. Part B is a sharp decrease of $K_{nr}(T_1, m)$ in function of m. m being sufficiently high, fuzziness begins to appear inside clusters. The maximum number of clusters K_{max} being intentionally high, redundant clusters surely appear. Part C corresponds to a stabilization of $K_{nr}(T_1, m)$ which can be explained by a stabilization of the FCM solution in function of m. For this threshold T_1, the best number of non-redundant clusters $K_b(T_1)$ necessary to describe the studied dataset is defined as the constant number of non-redundant clusters designed by plateau C. The best fuzziness parameter $m_b(T_1)$ can take any value included in plateau C. However, we choose the smallest value of m belonging to plateau C. This procedure enables to limit the complexity of the method since there is no need to continue to compute FCM for the remaining m-values as soon as the break point P is detected. To resume, $K_b(T_1)$ and $m_b(T_1)$ are respectively the ordinate and the abscissa of the break point P on figure 1.

(vi) By repeating steps (i) to (v) for several values of the threshold T belonging to the set $\mathcal{T} = \{T_1, \ldots, T_q, \ldots, T_Q\}$ previously defined by the user, Q sets $\mathcal{K}_{nr}(T_q, m)$, $1 \leqslant q \leqslant Q$, with a shape similar to the one represented on

figure 1 are obtained. Q break points, and thus Q couples $(K_b(T_q), m_b(T_q))$, are computed. For example, this set $\mathcal{T}$ can be fixed to the values between 0.4 and 0.95, representing high redundancy interval, with a step of 0.05. It is important to notice here that in order to reduce the number of computed FCM, for a given threshold T_q, if the FCM results have already been computed for a given couple (K, m) for a previous threshold T_r, $1 \leqslant r \leqslant q-1$, then there is no need to compute the FCM algorithm again. Furthermore, for this given couple (K, m), the set of redundant values $\mathcal{R}$ has already been computed and stored in memory. Hence, the redundant values must only be compared to the current threshold T_q.

(vii) The optimal couple $(\widehat{K}, \widehat{m})$ is then estimated as follows. For $k \in \{1, 2, \ldots, K_{max}\}$, we define $I(k) = \{q, K_b(T_q) = k, 1 \leqslant q \leqslant Q\} \in \mathbb{N}^{D(k)}$, where $D(k)$ represents the size of $I(k)$. $I(k)$ is thus a vector composed of the indices of the thresholds for which the best number of non-redundant clusters is equal to k. The optimal number of clusters is defined as $\widehat{K} = \arg\max_k D(k)$ and represents the most frequent number of clusters among the Q values $K_b(T_q)$, $1 \leqslant q \leqslant Q$. The optimal fuzziness parameter is estimated as $\widehat{m} = \frac{\sum_{q=1}^{D(\widehat{K})} m_b(T_{I(\widehat{K},q)})}{D(\widehat{K})}$ where $I(\widehat{K}, q)$ is the q-th element of $I(\widehat{K})$. $\widehat{m}$ represents the mean value of the abscissa of the break points localized at an ordinate equal to $\widehat{K}$. This procedure to estimate $\widehat{K}$ and $\widehat{m}$ based on the break points has shown its efficiency when applied on infrared spectral images of human skin cancer tissue sections [31].

The pseudo-code of this algorithm is provided in algorithm 1.

4. Experimental setup

In order to assess their performances, validity indices and $K-m$ methods presented in sections 2 and 3 have been applied on simulated and real datasets described in the following paragraphs. Furthermore, the values chosen for the common and particular parameters of the proposed methods are given.

4.1. Datasets

The methods are tested on 9 different datasets traditionally used in literature to assess the performances of validity indices, especially in function of the number of clusters, the separation between clusters and the number of objects. These datasets are divided into two categories. The first one is composed of 5 simulated datasets and the second of 4 real datasets.

4.1.1. Simulated datasets

The simulated datasets are built on the same model. Each cluster is composed of objects generated with a 2-dimensional gaussian or uniform distribution with a diagonal covariance matrix. From one dataset to another, the number of clusters, the number of objects per cluster, the separation between clusters and the cluster compactness are different. These datasets were previously introduced and used to test validity indices in [25, 33].

DataSet_4_2 is composed of 4 clusters, each containing 50 objects. As can be seen on figure 2, the 4 clusters are overlapped and are not compact.

DataSet_5_2 is composed of 5 clusters, each containing 100 objects. As shown on figure 3, the clusters are overlapped but compact.

DataSet_6_2 is composed of 6 clusters, each containing 50 objects. Figure 4 depicts these clusters which are well separated and compact.

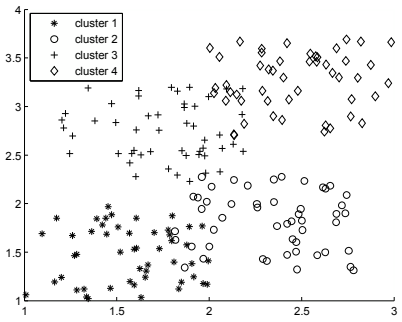


Figure 2: Plot of DataSet_4_2

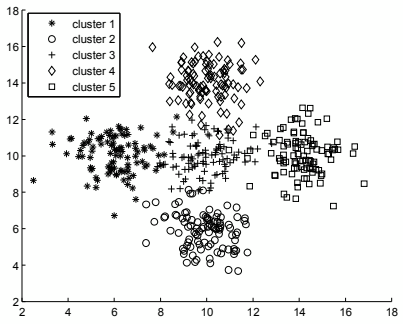


Figure 3: Plot of DataSet_5_2

DataSet_15_2 is composed of 15 clusters, each containing 15 objects. These clusters are well separated and compact as can be observed on figure 5.

The last simulated dataset is called X4 and was described in [28]. It is composed of 9 clusters with different densities and shapes, each containing 200 objects. As can be seen on figure 6, some clusters are small and compact, while others are big. Furthermore, some clusters overlap or are very close. These dataset is particularly complex since clusters are grouped 3 by 3, leading to the possible estimation of a 3 clusters structure by validity indices and $K - m$ methods.

4.1.2. Real datasets

The Ecoli dataset is composed of 336 protein sequences of the *E.coli* bacteria localized in 8 various sites within a cell. Each sequence is composed of 7 features. The real number of clusters is the number of cell sites, i.e. 8 [17, 30]. The clusters are respectively composed of 143, 77, 2, 2, 35, 20, 5 and 52 objects. This dataset is freely available at [1].

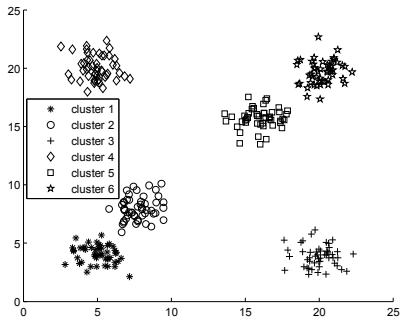


Figure 4: Plot of DataSet_6_2

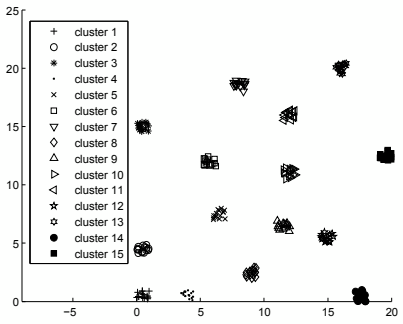


Figure 5: Plot of DataSet_15_2

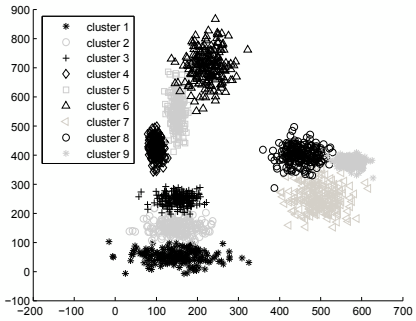


Figure 6: Plot of X4

The Iris dataset [12] is a very well known real dataset composed of 3 clusters, each containing 50 objects in a 4 dimensional space. The three clusters are iris species (Iris setosa, Iris virginica and Iris versicolor), and the four features of each object are the lenght and width of the sepals and petals of flowers in centimeter.

The Serum dataset [18] is composed of 517 gene expression recorded at 13 different time-points. The 10 clusters are respectively composed of 100, 145, 34, 43, 7, 34, 14, 63, 19 and 25 genes. 33 genes were identified as outliers. This dataset can be freely uploaded at [2].

The Wisconsin Breast Cancer (WBC) dataset [22] is composed of a benign cluster and a malignant cluster. The two clusters respectively contain 444 and 239 objects. Each object is described by 9 cellular features. The data are freely available at [3].

4.2. Parameters

To apply all methods in identical conditions, the FCM stopping criteria were identically configured in each case as follows. The maximum number of iterations was fixed to 500. The minimum amount of improvement of the objective function (1) between two consecutive iterations was fixed to 10^{-5}.

The maximum number of clusters K_{max} being dependent on the considered dataset, the convention largely shared in literature is to fix K_{max} to $\sqrt{N}$. Here, we used $K_{max} = [\sqrt{N}]+1$, where [.] represents the floor function.

The presented validity indices are supposed to be robust in function of the fuzziness parameter m. In the literature, these indices are thus applied to the datasets mentioned above using $m = 2$. The same conventions are used in this article.

For SJM, the m-value begins equal to 1.1 and is iteratively increased of 0.1. Two cluster prototypes are considered as identical if their euclidian distance is smaller than a threshold T_m fixed to 0.1 as proposed in [30].

For RBA and URBA, the m-value set is chosen as $\mathcal{M} = \{1.1, 1.2, \ldots, 20\}$, and the set of threshold T is fixed to $\mathcal{T} = \{0.05, 0.1, \ldots, 0.95\}$.

FCM algorithm being dependent on the initialization of the membership matrix, 10 repetitions of FCM for each (K,m) couple were performed. For each (K,m) couple, validity indices and $K-m$ methods were applied on the FCM result providing the smallest value of the objective function (1) such as suggested in [30].

Usually, the number of clusters estimated by the validity indices or $K-m$ methods is directly compared to the real number of clusters of the studied dataset. If the estimated number of clusters is equal to the real one, then the method is said to be efficient. If they are different, then the method is said to fail. However, this interpretation of the results is correct only if the FCM algorithm, applied with the real number of clusters, is able to retrieve the underlying structure of the dataset. If not, validity indices and $K-m$ methods won't be able to estimate the real number of clusters. In this case, the estimated number of clusters must be compared to another reference. In this study, the evaluation methodology proposed in [16] was implemented, i.e. a validity index or a $K-m$ method succeeds if its estimated number of clusters $\widehat{K}$ is equal to the number of clusters K_{VI} which induces the most similar partition to the real partition according to the variation of information

concept, i.e.:

$$K_{VI} = \underset{2 \leq k \leq K_{max}}{\arg\min} \left(H(P_k) + H(P) - 2M_I(P_k, P) \right) \tag{19}$$

with P_k the deffuzified partition composed of k clusters estimated by FCM, P the real ideal partition of the data, $H(P_k)$ and $H(P)$ the entropies of partition P_k and P respectively, and $M_I(P_k, P)$ the mutual information of partitions P_k and P.

5. Results

5.1. Example of application of URBA to DataSet_4_2

In order to illustrate the originality (which is to estimate the m-value at the same time as the number of clusters K) and explain the estimation procedure of the URBA algorithm, it has been firstly applied to DataSet_4_2.

Figure 7 presents the plot of the maximum number of non-redundant clusters K_{nr} in function of m for several values of the threshold $T \in \{0.1, \ldots, 0.9\}$ as estimated by the URBA algorithm when applied on DataSet_4_2. It can be easily seen on this figure that the general shape of the different curves is the same as the one presented on figure 1. The decrease of K_{nr} is extremely rapid in function of m whatever is the value of the threshold T. As described earlier, it can be seen on figure 7 that, for a specific threshold T_q, $1 \leqslant q \leqslant Q$, the research procedure of the best number of clusters $K_b(T_q)$ is stopped as soon as a break point is detected. Computational time is thus shortened since FCM is not computed for the remaining m-values.

From these plots of K_{nr} in function of m for several T_q, the coordinates of the break points can be determined as presented in table

q	1	2	3	4	5	6	7	8	9
T_q	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
$K_b(T_q)$	1	4	4	4	4	4	4	4	7
$m_b(T_q)$	1.6	1.9	2.2	2.5	2.8	3	3.3	3.6	3.1

Table 1: Coordinates $K_b(T_q)$ and $m_b(T_q)$ of the break points of figure 7 for the thresholds $T = \{T_1, \ldots, T_q, \ldots, T_9\} = \{0.1, \ldots, q \times 0.1, \ldots, 0.9\}$.

1. From these results, the sets $I(k)$ can be constructed as explained in section 3.2: $I(1) = \{1\}$, $I(4) = \{2, 3, 4, 5, 6, 7, 8\}$, $I(7) = \{9\}$, and the remaining $I(k)$ are empty. The dimensions of these sets are thus respectively $D(1) = 1$, $D(4) = 7$, $D(7) = 1$, and the remaining $D(k)$ are equal to 0. Thus, $\max_k D(k) = 7$ for $k = 4$, and thus the optimal number of clusters is $\widehat{K} = \arg\max_k D(k) = 4$. The estimation of the optimal fuzziness parameter is obtained by $\widehat{m} = \frac{\sum_{q=1}^{D(4)} m_b(T_{I(4,q)})}{D(4)} = \frac{\sum_{q=1}^{7} m_b(T_{I(4,q)})}{7} = \frac{m_b(T_2)+m_b(T_3)+m_b(T_4)+m_b(T_5)+m_b(T_6)+m_b(T_7)+m_b(T_8)}{7} = \frac{m_b(0.2)+m_b(0.3)+m_b(0.4)+m_b(0.5)+m_b(0.6)+m_b(0.7)+m_b(0.8)}{7} = \frac{1.9+2.2+2.5+2.8+3+3.3+3.6}{7} = 2.76$.

For this experimental setup, the optimal FCM parameters estimated by URBA for DataSet_4_2 are thus $\widehat{K} = 4$ and $\widehat{m} = 2.76$.

5.2. Application of validity indices and $K - m$ methods to the simulated datasets

The numbers of clusters $\widehat{K}$ estimated by the validity indices and $K - m$ methods when applied to the simulated datasets are presented in table 2. For each dataset, the real number of clusters $K^\star$ and the number of clusters K_{VI}, that induces the most similar partition to the real partition, are identical. FCM algorithm is thus perfectly suited to estimate a realistic partition of these datasets.

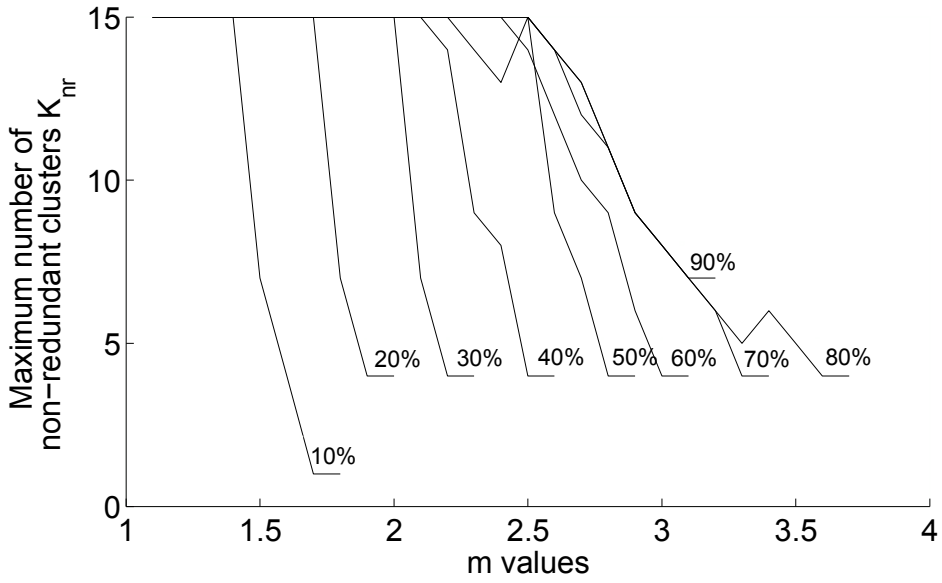


Figure 7: Maximum number of non-redundant clusters K_{nr}, in function of m for $T \in \{0.1, \ldots, 0.9\}$, as estimated by URBA algorithm on DataSet_4_2

For DataSet_4_2, STCI, SCI and RBA estimate a wrong number of clusters due to the large spread of the clusters, their uniform distribution, and their overlapping. These difficulties are however overcome by OSI, FSI, GIISI, SJM and URBA which correctly estimate $K^\star$.

Even if the clusters of DataSet_5_2 slightly overlap, all the algorithms, except RBA, are able to perfectly estimate $K^\star$.

For DataSet_6_2 which is composed of compact and well separated clusters, only STCI, SCI and URBA are able to find the real underlying structure of the data. FSI, GIISI and SJM estimate a wrong number of clusters equal to 4. This mistake can be explained by the hierarchical structure of the dataset, i.e. clusters 1 and 2 of figure 4 can be seen as a unique cluster. The same can be said for clusters 5 and 6 of figure 4.

DataSet_15_2 shares the same characteristics as DataSet_6_2 since it is composed of compact and well separated clusters. The only difference is the number of objects composing the clusters which is smaller for DataSet_15_2 than for DataSet_6_2. STCI and SJM fail while the remaining algorithms succeed to estimate $K^\star$.

X4 is the most difficult dataset to be partitioned since it is a hierarchical dataset composed of clusters with different densities, shapes and overlapping rates. OSI, FSI, GIISI and SJM estimate a wrong number of clusters equal to 3, as expected due to the structure of this dataset. STCI, SCI, RBA and URBA are insensitive to these variable characteristics of the datasets and to the hierarchical structure of the data since they correctly estimate the real number of clusters.

To summarize, only URBA is able to correctly estimate the right number

Dataset	K_{max}	STCI	SCI	OSI	FSI	GIISI	SJM	RBA	URBA	$K^\star$	K_{VI}
DataSet_4_2	15	15	7	4	4	4	4	8	4	4	4
DataSet_5_2	23	5	5	5	5	5	5	6	5	5	5
DataSet_6_2	18	6	6	5	4	4	4	7	6	6	6
DataSet_15_2	16	16	15	15	15	15	4	15	15	15	15
X4	43	9	9	3	3	3	3	9	9	9	9

Table 2: Estimation of the numbers of clusters $\widehat{K}$ for the simulated datasets by validity indices and $K-m$ methods. The first column gives the name of the datasets, the second the maximum number of clusters K_{max} authorized in the algorithms, the third to the seventh $\widehat{K}$ estimated by the validity indices, the eighth to the tenth $\widehat{K}$ estimated by the $K-m$ methods, the eleventh the real number of clusters $K^\star$, and the twelfth the number of clusters K_{VI} recommended by the variation of information concept [16] and computed by equation (19).

of clusters of the five simulated datasets, and thus to adapt to a large range of characteristics of datasets.

5.3. Application of validity indices and $K-m$ methods to the real datasets

The numbers of clusters $\widehat{K}$ estimated by the validity indices and $K-m$ methods when applied to the real datasets are presented in table 3. Contrary to the simulated datasets, Ecoli and Serum datasets present a real number of clusters $K^\star$ that is different from the number of clusters K_{VI} that best partitions the data. This may be due to the complexity of the datasets or to the shape and density of the clusters which may be incompatible with the euclidean distance used in the FCM algorithm applied in this article. Furthermore, some clusters are composed of only few objects and can thus be difficultly be recovered by FCM. K_{VI} is thus here considered as the gold-standard number of clusters necessary to best partition the dataset with the euclidean distance FCM algorithm.

For the Ecoli dataset, the number of clusters $K_{VI}=3$ is correctly estimated by STCI, OSI, FSI, SJM and URBA. No method is able to estimate the real number of clusters $K^\star$.

Dataset	K_{max}	STCI	SCI	OSI	FSI	GIISI	SJM	RBA	URBA	$K^\star$	K_{VI}
Ecoli	19	3	5	3	3	2	3	4	3	8	3
Iris	13	3	3	2	2	2	2	8	3	3	3
Serum	23	3	6	2	2	2	5	5	5	10	5
WBC	27	2	6	2	4	2	2	2	2	2	2

Table 3: Estimation of the numbers of clusters $\widehat{K}$ for the real datasets by validity indices and $K-m$ methods. The first column gives the name of the datasets, the second the maximum number of clusters K_{max} authorized in the algorithms, the third to the seventh $\widehat{K}$ estimated by the validity indices, the eighth to the tenth $\widehat{K}$ estimated by the $K-m$ methods, the eleventh the real number of clusters $K^\star$, and the twelfth the number of clusters K_{VI} recommended by the variation of information concept [16] and computed by equation (19).

The number of clusters of Iris is found by only STCI, SCI and URBA. The small number of succeeded methods for this dataset can be explained by the fact that two clusters of the Iris dataset are highly overlapped. Thus, the majority of the algorithms, i.e. OSI, FSI, GIISI and SJM estimates two clusters instead of three.

Like the Ecoli dataset, the Serum dataset is a dataset difficult to partition as illustrated by the results of the validity indices which are unable to correctly estimate the K_{VI} number of clusters. Only the $K-m$ methods succeed.

The structure of the WBC dataset is correctly determined by the $K-m$ methods and the STCI, OSI and GIISI validity indices. SCI and FSI overestimated the real number of clusters.

Only URBA demonstrates the ability to correctly estimate the number of clusters that best partition the data whatever the real dataset.

When considering both real and simulated datasets, the best validity index is STCI with a correct estimation of the gold-standard number of clusters for 6 out of 9 datasets. However, the best algorithm is the URBA K-m method with 9 correct estimation out of 9.

Dataset	SJM	RBA	URBA
DataSet_4_2	2.8	2.7	2.4
DataSet_5_2	2.7	2.5	2.2
DataSet_6_2	1.1	2.6	2.2
DataSet_15_2	3.1	2.6	2.6
X4	6.6	2.2	2.1
Ecoli	1.4	2	1.7
Iris	3.1	2	2.1
Serum	1.2	1.9	1.7
WBC	1.3	10.3	1.5

Table 4: Estimation of the optimal fuzziness parameter $\widehat{m}$ by the $K-m$ methods for the simulated and real datasets

5.4. Estimation of the fuzziness parameter

Contrary to the validity indices, $K-m$ methods have the property to estimate the optimal fuzziness parameter $\widehat{m}$ to highlight the structure of the considered dataset simultaneously to the optimal number of clusters.

The optimal m-values estimated by the three $K-m$ methods applied to the real and simulated datasets are summarized in table 4. SJM has a tendency to estimate small (less than 1.5) or high (more than 2.5) m-values. Thus, too hard or to fuzzy results will be estimated by FCM when applied with the parameters assessed by SJM. The same observations can be done for RBA, which overestimates m-values. URBA exhibits an appealing behavior since it estimates m-values belonging to a compact range, i.e. the $[1.5; 2.6]$ range.

6. Discussion

When applied to simulated datasets, validity indices are better than SJM $K-m$ method for two main reasons. The first one is that the real clusters are well separated and compact for the majority of the proposed simulated datasets. The conceptual assumptions on which are based the mathematical

definitions of the proposed validity indices are thus completely fulfilled. The second reason is the weakness of SJM when applied on datasets presenting a hierarchical structure due to equation (14) which can clearly underestimate the number of clusters in this special case. Our proposed URBA is not sensible to hierarchical structures since FCM parameters are estimated from the first stabilization plateau of the FCM results by successively decreasing the number of clusters. The number of clusters corresponding to the higher structure of the data (thus corresponding to a lower number of clusters) will thus correspond to the apparition of a second plateau for higher m-values.

When applied to real datasets, the validity indices are shown to be less efficient than the $K - m$ methods in the previously described experiments. We think that this is due to two main reasons. First, the presented validity indices are defined on compactness and separation concepts which are not fulfilled for the proposed real datasets. Second, the validity indices are applied following the conventions which fix the m-value to 2 whatever is the data structure, contrary to the $K - m$ methods which are able to provide an estimation of m in function of the data structure. This assumption is in accordance with several studies [10, 38, 24], which report that the m-value must be optimized for each dataset since it is directly linked to the data structure. Some studies have also reported that the validity indices can reach to a failure of FCM to extract a meaningful clustering structure if $m = 2$ [11]. The best choice of m has also been shown to be directly linked to the degree of overlapping or mixture between clusters [10]. The choice of m is thus very important [39] especially for real datasets.

Some methods have been developed in order to estimate only the m-value

[11, 9, 24, 38]. However, the majority of these methods requires the *a priori* knowledge of the number of clusters included into the dataset. Furthermore, the validity indices tested in this article are known to be robust in m when compactness and separation of clusters are fulfilled. There is thus *a priori* no need to use such methods for the estimation of the m-value in addition to the validity indices. We thus ignored this kind of techniques in this paper.

The gold-standard for the choice of the m-values has been experimentally established by Pal and Bezdek [26] who recommended to select a m-value in the $[1.5; 2.5]$ range. This range is considered in this paper as the reference range for the selection of the m-values. Thus, among the $K-m$ methods, only URBA is able to estimate relevant m-values according to the recommendation done by Pal and Bezdek, except for DataSet_15_2. For this dataset, URBA estimated a m-value equal to 2.6 which is very near to the upper bound of 2.5 suggested by Pal and Bezdek.

The RBA algorithm developed in [31] shows a tendency to overestimate the number of clusters as presented in tables 2 and 3. This behavior is directly linked to the Pearson correlation coefficient used as a measure of redundancy in RBA. In fact, Pearson correlation coefficient is known to be sensitive to outliers [4]. In our case, this property will induce a huge decrease of the correlation coefficient and thus will lead to the estimation of non redundant clusters by RBA for high number of clusters. To illustrate this sensitivity to outliers, the membership matrix U was estimated by FCM with $K = 9$ and $m = 2.7$ applied on DataSet_4_2. The highest correlation coefficient is equal to 0.30. The corresponding clusters are thus relatively not correlated. When observing the membership values, only 10 points have a

very high membership value for one of the two considered clusters. If these 10 points are removed, then the correlation coefficient is equal to 0.53, indicating the existence of correlation between clusters. The removal of outliers thus induces a high increase of the value of the correlation coefficient. For the same clusters, the fuzzy similarity measure is equal to 0.52 when computed on the entire dataset, and to 0.59 after removal of the 10 outliers. The fuzzy similarity measure is thus relatively insensitive to outliers.

Q curves $\mathcal{K}_{nr}(T_q, m)$, $1 \leqslant q \leqslant Q$, are used by RBA and URBA to estimate the optimal number of clusters and fuzziness parameter, contrary to the other presented methods which use only one curve $V(K)$, $2 \leqslant K \leqslant K_{max}$. RBA and URBA are thus more robust than the other presented methods to variability of FCM results since the estimation of these FCM parameters results from a consensus.

7. Conclusion

The estimation of the number of clusters is an important problem of fuzzy clustering. Validity indices have been developed in literature to bring a response to this problem. However, validity indices are used with a fuzziness parameter m equal to 2. Some studies have reported that this predefined choice of m can lead to mistakes. In this article, an alternative solution is proposed by simultaneously estimate these two parameters, i.e. the number of clusters and the fuzziness parameter. Based on a fuzziness similarity measure, the proposed URBA algorithm is searching for the maximum number of non-redundant clusters for each m-value. When the FCM algorithm finds a stable solution, the maximum number of non-redundant clusters remains

constant. The number of non-redundant clusters and the fuzziness parameter of this break point are chosen as the optimal parameter values of the FCM algorithm. Compared to 5 well-known and recent validity indices and 2 other $K-m$ methods on 9 different simulated and real datasets, our method reveals the best. Our algorithm proposes some interesting characteristics. First, the fuzziness parameter is estimated at the same time as the number of clusters. Second, it is insensitive to hierarchical structures. This algorithm is very efficient and should help to easily and automatically process huge and complex real life datasets such as those generated in microarray and vibrational spectroscopy for instance.

Acknowledgements

The authors would like to thank CNRS Projets Exploratoires Pluridisciplinaires for financial support.

References

[1] http://archive.ics.uci.edu/ml/datasets/Ecoli.

[2] http://genome-www.stanford.edu/serum/data.html.

[3] http://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+%28Original%29.

[4] M.B. Abdullah. On a robust correlation coefficient. *The Statistician*, 39(4):455–460, 1990.

[5] J. C. Bezdek. *Pattern recognition with fuzzy objective function algorithms*. Plenum Press, New York, 1981.

[6] M. Bouguessa, S. Wang, and H. Sun. An objective approach to cluster validation. *Pattern Recognition Letters*, 27:1419–1430, 2006.

[7] R.J.G.B. Campello and E.R. Hruschka. A fuzzy extension of the silhouette width criterion for cluster analysis. *Fuzzy Sets and Systems*, 157(21):2858–2875, 2006.

[8] W. Chen, M. L. Giger, and U. Bick. A Fuzzy C-Means (FCM)-based approach for computerized segmentation of breast lesions in dynamic contrast-enhanced MR images. *Academic Radiology*, 13(1):63–72, 2006.

[9] H. Choe and J. B. Jordan. On the optimal choice of parameters in a Fuzzy C-Means algorithm. In *Proceedings of the IEEE International Conference on Fuzzy Systems*, pages 349–354, San Diego, CA , USA, March 1992.

[10] P. J. Deer and P. Eklund. A study of parameter values for a Mahalanobis distance fuzzy classifier. *Fuzzy Sets and Systems*, 137(2):191–213, 2003.

[11] D. Dembélé and P. Kastner. Fuzzy C-Means method for clustering microarray data. *Bioinformatics*, 19(8):973–980, 2003.

[12] R.A. Fisher. The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, 7(2):179–188, 1936.

[13] Y. Fukuyama and M. Sugeno. A new method of choosing the number of clusters for the Fuzzy C-Means method. In *Proceedings of the 5th Fuzzy Systems Symposium*, pages 247–250, Japan, 1989.

644 [14] M.E. Futschik and B. Carlisle. Noise-robust soft clustering of gene ex-
645 pression time-course data. *Journal of Bioinformatics and Computational*
646 *Biology*, 3(4):965–988, 2005.

647 [15] A.B. Geva, Y. Steinberg, S. Bruckmair, and G. Nahum. A comparison of
648 cluster validity criteria for a mixture of normal distributed data. *Pattern*
649 *Recognition Letters*, 21(6–7):511–529, 2000.

650 [16] I. Gurrutxaga, J. Muguerza, O. Arbelaitz, J.M. Pérez, and J.I. Martín.
651 Towards a standard methodology to evaluate internal cluster validity
652 indices. *Pattern Recognition Letters*, 32:505–515, 2011.

653 [17] P. Horton and K. Nakai. A probablistic classification system for pre-
654 dicting the cellular localization sites of proteins. In *Proceedings of the*
655 *Fourth International Conference on Intelligent Systems in Molecular Bi-*
656 *ology*, pages 109–115, St Louis, MO, USA, June 1996.

657 [18] V.R. Iyer, M.B. Eisen, D.T. Ross, G. Schuler, T. Moore, J.C. Lee, J.M.
658 Trent, L.M. Staudt, J.Jr. Hudson, M.S. Boguski, D. Lashkari, D. Shalon,
659 D. Botstein, and P.O. Brown. The transcriptional program in the re-
660 sponse of human fibroblasts to serum. *Science*, 283(5398):83–87, 1999.

661 [19] Y. I. Kim, D.-W. Kim, K. Lee, and K. H. Lee. A cluster validation index
662 for GK cluster analysis based on relative degree of sharing. *Information*
663 *Sciences*, 168:225–242, 2004.

664 [20] H. Le Capitaine and C. Frélicot. A cluster-validity index combining an
665 overlap measure and a separation measure based on fuzzy-aggregation
666 operators. *IEEE Transactions on Fuzzy Systems*, 19(3):580–588, 2011.

667 [21] J. B. MacQueen. Some methods for classification and analysis of mul-
668 tivariate observations. In *Proceedings of 5th Berkeley Symposium on*
669 *Mathematical Statistics and Probability*, pages 281–297. University of
670 California Press, 1967.

671 [22] O.L. Mangasarian and W.H. Wolberg. Cancer diagnosis via linear pro-
672 gramming. *SIAM News*, 23(5):1–18, 1990.

673 [23] A. Marec, J.-H. Thomas, and R. El Guerjouma. Damage characteri-
674 zation of polymer-based composite materials: Multivariate analysis and
675 wavelet transform for clustering acoustic emission data. *Mechanical Sys-*
676 *tems and Signal Processing*, 22:1441–1464, 2008.

677 [24] F. Okeke and A. Karnieli. Linear mixture model approach for selecting
678 fuzzy exponent value in Fuzzy C-Means algorithm. *Ecological Informat-*
679 *ics*, 1(1):117–124, 2006.

680 [25] M. K. Pakhira, S. Bandyopadhyay, and U. Maulik. Validity index for
681 crisp and fuzzy clusters. *Pattern Recognition*, 37:487–501, 2004.

682 [26] N. R. Pal and J. C. Bezdek. On cluster validity for the Fuzzy C-Means
683 model. *IEEE Transactions on Fuzzy Systems*, 3(3):370–379, 1995.

684 [27] R. Parthasarathy, G. Thiagarajan, X. Yao, Y.-P. Wang, P. Spencer, and
685 Y. Wang. Application of multivariate spectral analysis in micro-Raman
686 imaging to unveil structural/chemical features of the adhesive/dentin
687 interface. *Journal of Biomedical Optics*, 13(1):014020, 2008.

688 [28] B. Rezaee. A cluster validity index for fuzzy clustering. *Fuzzy Sets and*
689 *Systems*, 161(23):3014–3025, 2010.

690 [29] P.J. Rousseeuw. Silhouettes: A graphical aid to the interpretation and
691 validation of cluster analysis. *Computational and Applied Mathematics*,
692 20:53–65, 1987.

693 [30] V. Schwämmle and O.N. Jensen. A simple and fast method to deter-
694 mine the parameters for fuzzy c-means cluster analysis. *Bioinformatics*,
695 26(22):2841–2848, 2010.

696 [31] D. Sebiskveradze, V. Vrabie, C. Gobinet, A. Durlach, P. Bernard, E. Ly,
697 M. Manfait, P. Jeannesson, and O. Piot. Automation of an algorithm
698 based on fuzzy clustering for analyzing tumoral heterogeneity in human
699 skin carcinoma tissue sections. *Laboratory Investigation*, 91(5):799–811,
700 2011.

701 [32] M. Setnes, R. Babuska, U. Kaymak, and H. R. van Nauta Lemke. Simi-
702 larity measures in fuzzy rule base simplification. *IEEE Transactions on*
703 *Systems, Man, and Cybernetics - Part B: Cybernetics*, 28(3):376–386,
704 1998.

705 [33] W. Wang and Y. Zhang. On fuzzy cluster validity indices. *Fuzzy Sets*
706 *and Systems*, 158:2095–2117, 2007.

707 [34] X.-Y. Wang, J. M. Garibaldi, B. Bird, and M. W. George. A novel
708 fuzzy clustering algorithm for the analysis of axillary lymph node tissue
709 sections. *Applied Intelligence*, 27:237–248, 2007.

710 [35] Y. Wang, X. Yao, and R. Parthasarathy. Characterization of interfa-
711 cial chemistry of adhesive/dentin bond using FTIR chemical imaging

712 with univariate and multivariate data processing. *Journal of Biomedical*
713 *Materials Research - Part A*, 91(1):251–262, 2009.

714 [36] J. H. Ward. Hierarchical grouping to optimize an objective function.
715 *Journal of the American Statistical Association*, 58(301):236–244, 1963.

716 [37] X. L. Xie and G. Beni. A validity measure for fuzzy clustering. *IEEE*
717 *Transactions on Pattern Analysis and Machine Intelligence*, 13(8):841–
718 847, 1991.

719 [38] J. Yu, Q. Cheng, and H. Huang. Analysis of the weighting exponent in
720 the FCM. *IEEE Transactions on Systems, Man, and Cybernetics–Part*
721 *B: Cybernetics*, 34(1):634–639, 2004.

722 [39] J. Yu, H. Huang, and S. Tian. An efficient optimality test for the Fuzzy
723 C-Means algorithm. In *Proceedings of the IEEE International Confer-*
724 *ence on Fuzzy Systems*, volume 1, pages 98–103, Honolulu, HI, USA,
725 May 2002.

Article #III

Publication n°3

A review of parameter estimation methods for Fuzzy C-Means

T. Happillon, D. Sebiskveradze, V. Vrabie, M. Manfait, C. Gobinet.

En cours de rédaction

Présentation de la publication n°3

Contexte

Comme déjà décrit dans les préambules des publications n°1 et n°2, l'algorithme des Fuzzy C-Means est géré par deux paramètres, le nombre de classes voulues *C*, et le degré de flou *m*. Les indices de validité sont des méthodes bien connues pour estimer le nombre de classes optimales d'un jeu de données pour une valeur fixée de *m*. L'estimation du paramètre de flou *m* est souvent occultée puisque la grande majorité des études considère que le choix de *m* n'est pas important. Cependant, des études ont montré que ce choix de *m* est primordial et lié au choix de *C*. C'est pourquoi des méthodes appelées règles en *m*, ou *m*-rules, ont été développées afin d'estimer la valeur optimale de *m* pour un jeu de données considéré. Or ces deux catégories de méthodes ont été développées indépendamment l'une de l'autre. Ainsi, pour appliquer les indices de validité, *m* doit être connu afin de déterminer *C.* Inversement la majorité des règles en *m* supposent *C* connu. D'autres outils ont été développés afin d'estimer à la fois *C* et *m*, tels la méthode de Schwämmle-Jensen (SJM), l'algorithme FCM-RBA ou encore sa version améliorée FCM-URBA.

Objectif

Face au manque de bibliographie sur les règles en *m*, et sur les méthodes *K-m* telles SJM, FCM-RBA et FCM-URBA, et donc sur l'importance du choix de *m*, l'objectif de ce papier est de proposer une revue de ces méthodes et des indices de validité, et de comparer leurs performances sur des jeux de données simulés ou réels couramment utilisés dans la littérature.

Matériels et Méthodes

Dans cette revue, 28 indices de validité, 5 règles en *m* (FDO, DK, OK, αR, βR), et trois méthodes *K-m* (SJM, FCM-RBA, FCM-URBA) ont été appliquées à 27 jeux de données différents (17 simulés et 6 réels). De même que pour la publication n°2, la comparaison des résultats est faite de façon objective en utilisant la méthode de variation de l'information. Dans cette étude, trois des cinq règles en *m* ont estimé des valeurs de *m* exploitables (*i.e.* différentes de $+\infty$). Ainsi, pour chaque jeu de données testé, ces trois valeurs ont été

considérées pour estimer le nombre de classe optimal. La valeur standard de *m* fixée à 2 dans la littérature est également testée.

Résultats

Lorsque la valeur optimale de *m* pour chaque jeu de données est déterminée par la règle en *m* FDO, la méthode estimant le nombre de classe optimal pour un maximum de jeux de données est l'indice de validité V_{ASI} avec un score de 34 jeux correctement partitionnés sur un total de 49. Vient ensuite l'indice V_{EASI} avec un score de 31/49, puis URBA avec un score de 29/49.

En considérant les valeurs de *m* déterminées par la méthode DK, l'indice V_P et URBA apparaissent les meilleurs avec un score de 36/49, puis les indices V_{MPC} et V_{ASI} avec un score de 35/49 et enfin les indices V_{GIISI} et V_{OSI} avec un score de 34/49.

Les valeurs de *m* optimales estimées pour chaque jeu de données par la méthode OK permettent à URBA, V_P, V_{ASI} et V_{MPC} d'obtenir un score de 34/49. Suivis de V_{FHV} et V_{GIISI} avec un score de 32/49, puis V_{EASI} avec un score de 31/49.

Enfin, le plus grand nombre de partitions optimales est obtenu avec une valeur de *m* fixée à sa valeur standard dans la littérature (*m*=2) et est obtenu par les indices V_P et V_{ASI} avec un score de 37/49, URBA atteint un score de 36/49 et les trois indices V_{MPC}, V_{os} et V_{OSI} obtiennent un score de 35/49.

Conclusion

La détermination du paramétrage optimal de l'algorithme FCM est connue dans la littérature pour être un réel problème puisque les résultats obtenus en dépendent fortement. Les indices de validité, les règles en *m* et les méthodes *K-m* ont donc été proposées dans le but de répondre à ce problème. Cependant, on notera que de par sa nature et son fonctionnement, l'algorithme FCM n'est pas capable de segmenter idéalement tout type de jeu de données, et c'est la raison pour laquelle le nombre de classes optimal considéré pour chaque jeu de donnée ne peut pas être son nombre de classes réel mais le nombre de classes générant la partition la plus en accord avec le partitionnement original (estimé par la méthode de variation de l'information). Dans cette revue, l'indice de validité V_{ASI} ainsi que la méthode URBA apparaissent comme les plus performants quelque soit la méthode utilisée

pour estimer la valeur optimale de *m*. Les indices V_P et V_{MPC} montrent également une bonne capacité à estimer le nombre optimal de classes pour les jeux de données réels et simulés testés. Finalement, peu de méthodes parmi l'ensemble testé apparaissent robustes face à différentes valeurs du degré de flou et lorsqu'elles sont appliquées à des jeux de données variant de par leur nature (simulé ou réel), leur dimension et leur taille.

A review of parameter estimation methods for Fuzzy C-Means

Teddy Happillon, David Sebiskveradze, Valeriu Vrabie, Michel Manfait, and Cyril Gobinet

Abstract

An open problem in fuzzy clustering is the estimation of its two parameters, i.e. the number of clusters C, and the fuzziness parameter m. Validity indices have been developed to estimate the optimal C value, while m-rules assess the best m value for each dataset. However, these two kinds of methods have been independently developed and consequently, one of the two parameters has to be known to determine the other. Thus, this work aims at proposing an overview of 29 well known validity indices, 5 m-rules and 3 methods estimating simultaneously C and m (denoted $C-m$ methods), applied on 49 simulated and real datasets commonly used in the literature. Among the 5 m-rules, 3 returned feasible results which were tested on each dataset and each validity index. The standard value of m, i.e. m=2, is also considered. In this study, the optimal number of clusters obtained by validity indices and $C-m$ methods are not compared to the real number of clusters composing each datasets, but to a number of clusters which provide a correct partition of these datasets [1]. The obtained results show that the best validity indices return a correct partition for 37 datasets among the 49 tested. Depending on the optimal value of m considered, other validity indices and $C-m$ methods provide such comparable results.

Index Terms

Fuzziness parameter, fuzzy clustering, m-rules, number of clusters, redundancy, validity indices.

I. Introduction

Unsupervised classification methods (also called clustering methods) have shown their efficiency to simply and rapidly analyze simulated or real datasets, and for most of them, only few parameters need to be managed. This kind of method makes no assumption about the data, especially on their labelling. Among these techniques, Fuzzy C-Means (FCM) [2] has rapidly gained popularity because of its unquestionable advantages (rapidity, less sensitivity to initialization, possible distribution of an object to all clusters) compared to hard clustering such as Hierarchical Cluster Analysis (HCA) [3] and K-Means [4] algorithms. However, an open problem remain about the automatic and unsupervised estimation of their parameters. As regards the FCM algorithm, two parameters have to be set,

This work was supported in part by the CNRS Projets Exploratoires Pluridisciplinaires and the Institut National du Cancer.

T. Happillon (e-mail: teddy.happillon@gmail.com), D. Sebiskveradze (e-mail: david.sebiskveradze@univ-reims.fr), M. Manfait (e-mail: michel.manfait@univ-reims.fr), and C. Gobinet (corresponding author; phone: +33(0)326918125; fax: +33(0)326913550; e-mail: cyril.gobinet@univ-reims.fr) are with the Unité MEDyC, CNRS UMR 7369, Université de Reims Champagne-Ardenne, Reims, France.

V. Vrabie (e-mail: valeriu.vrabie@univ-reims.fr) is with the CReSTIC, Université de Reims Champagne-Ardenne, Reims, France.

$i.e.$ the number of clusters C, and the fuzziness parameter m. In order to automatically determined the optimal values for both of them, some tools have been proposed in the literature as for example cluster validity indices (CVI), estimating the optimal C value, or m-rules assessing the best fuzziness estimation for a given dataset. Nevertheless, a general limit for such techniques has to be put forward. Indeed, validity indices and m-rules have been developed independently from the other and consequently, to obtain a meaningful FCM partition, C must be estimated knowing m (validity indices), and m must be estimated knowing C (m-rules). It is clear that, both m and C being intimately linked, it is difficult to estimate one parameter independently from the other. The development of numerical tools for the automatic and simultaneous selection of these two parameters values is thus of great interest. In the following of this paper, this kind of method will be called, C-m methods. In this review, FCM classifications are realized on 49 simulated and real datasets commonly used in the literature, and 29 well known validity indices, 5 m-rules and 3 C-m methods are tested to optimize the FCM settings. However, it has to be mentioned that optimal numbers of clusters, obtained either by CVI or C-m methods, are not compared with the real number of clusters composing each dataset, but with the number of clusters of the partition the most similar to the real one. Indeed, this kind of classification tools are not always adapted for all kinds of existing datasets and for many of them, they can only estimate an approximative but relevant clustering of the data [1].

The following of this paper is organized as follow. In section II, the FCM algorithm is detailed and the different validity indices, m-rules and C-m methods are listed. Section III deals with the results obtained by their applications on a large number of real-world and simulated datasets. Section IV opens the discussion on the comparison between these different tools, and finally, section V concludes the paper.

II. Material and Methods

A. Fuzzy C-Means algorithm

FCM is an iterative partitioning method which divides the dataset into C clusters by minimizing an objective function, based on both a minimization and a maximization of intra-group and inter-group variabilities, respectively. When the final partition is returned, each object is distributed to every C clusters unequally, thanks to different membership values.

Suppose a dataset $X = [x_1, x_2, \ldots, x_N] \in \mathbb{R}^{P \times N}$ composed of N objects $x_i \in \mathbb{R}^P$ and divided into C clusters, each represented by its own prototype vector $v_j \in \mathbb{R}^P$. The objective function is thus :

$$J_m(U, V) = \sum_{i=1}^{N} \sum_{j=1}^{C} u_{ij}^m \|x_i - v_j\|^2. \tag{1}$$

Where u_{ij} is the membership value of the $i-th$ object x_i to the $j-th$ prototype $v_j \in \mathbb{R}^P$. $U \in \mathbb{R}^{N \times K}$ and $V \in \mathbb{R}^{P \times K}$ are matrices respectively constructed with the different u_{ij} values and v_j vectors.

u_{ij} is computed by:

$$u_{ij} = \frac{1}{\sum_{c=1}^{C} \left(\frac{\|x_i - v_j\|}{\|x_i - v_c\|} \right)^{\frac{2}{m-1}}}, 1 \leqslant i \leqslant N, 1 \leqslant j \leqslant C \tag{2}$$

and v_j according to:

$$\text{and } v_j = \frac{\sum_{i=1}^{N} u_{ij}^m x_i}{\sum_{c=1}^{N} u_{ij}^m}, 1 \leqslant j \leqslant C. \tag{3}$$

Some constraints are imposed in a FCM classification. As regards the computation of u_{ij}, the values of m should be chosen to be greater than 1, $i.e. 1 < m < +\infty$. When $m \longrightarrow 1$, the obtained result is highly comparable with hard clustering results such as K-Means ones [2]. On the contrary, if $m \longrightarrow +\infty$, each object is uniformly distributed on each cluster and thus all appear identical.

Two other constraints deals with the membership values, which must be for each x_i object $0 \leqslant u_{ij} \leqslant 1$ and $\sum_{j=1}^{C} u_{ij} = 1$ (i.e. full and unequally distribution of an object to every C clusters). This concept of membership values is here of great interest since overlapped clusters can easily be identified and uncertainty in object attributions can also be highlighted. Moreover, the following inequalities $0 < \sum_{i=1}^{N} u_{ij} < N$ insure that each cluster is not empty.

B. Validity indices

Although it is qualified as unsupervised, FCM depends on two parameters, C and m, which must be fixed by the user. Traditionally, C is chosen based on *a priori* knowledge of the composition of the data or by successive and time consuming trial-and-error steps. When multidimensional dataset acquired on complex samples is manipulated, the adjustment of this parameter becomes more difficult. To resolve this problem, automatic methods named cluster validity indices [5] have been developed to estimate the optimal number of clusters. The principle of use of these CVI is simple. First, FCM partitions are computed for several values of C, $2 \leqslant C \leqslant C_{max}$. Then, a validity index value is computed for each partition, and the smallest (or biggest depending on indices) corresponds to the optimal number of clusters, denoted C^*.

A listing of 29 most widely used is provided here. The interested reader can refer to [5] to have a complete review and interpretation of CVI developed since 30 years.

The validity indices can be roughly divided into two groups. The first one is composed with those only considering membership values of a classification. The partition coefficient (PC) [2] computes the average of object sharing between clusters. The partition entropy (PE) [2] indicates the fuzziness of the partition. The modified partition coefficient (MPC) [6] is a linear transformation of PC which reduces its dependence on the number of clusters. The Windham proportion exponent (WPE) [7] informs on the relative overlapping between fuzzy clusters. The Kim-Kim-Lee-Lee coefficient (KYI) [8] is based on a mathematical definition of the relative sharing between pairs of clusters. And finally, the fuzzy partition measure (P) [9] which results from the sum of compactness and separation terms. Several criticisms have been done on these indices. First, concerning their monotonic behavior toward C.

Then, the fact that they only relate to the membership values of the partition without taking into account the structure of the data itself. And finally, they are sensible to m [5].

To overcome these limitations, CVI which rely on both data and FCM results (membership values and prototypes) have been developed on separation and compactness concepts. They constitute the second group. Among these indices, the most popular are the Fukuyama-Sugeno index (FS) [10], the Xie-Beni index (XB) [11], the Kwon index (K) [12] which is an extension of the XB index, the Tang index (T) [13] which can be seen as another extension of XB index, the partition coefficient and exponential separation (PCAES) [14], the validity index of Tsekouras and Sarimveis (VI) [15], the compose within and between scattering (CWB) [16], the Wang-Sun-Jiang index (WSJ) [17] which is a modified version of CWB, the fuzzy version of Pakhira-Bandyopadhyay-Maulik index (PBMF) [18], the separation and total compactness index (STC) [19], the fuzzy separation and compactness index (FSC) [20]. All these validity indices share the same construction idea, which is based on the maximization of the inter-cluster separation and the minimization of the intra-cluster compactness.

Others validity indices of this group are based on other concepts such as the fuzzy hypervolume (FHV) [21], the partition density (PD) [21], the average partition density (APD) [21], the separation and compactness index proposed by Bouguessa and co-workers (SCBWS) [19], the granularity-dissimilarity measure (GD) [22], the optimality test (OT) [23], the ratio of the overlap degree to the separation (OS) [24], the average silhouette index (ASI) [25] and its extensions (EASI) [26] but also the generalized intra-inter sihouette index (GIISI) [27], the pattern distances ratio, differences ranking (PDRDR) [28], and finally the recent overlap and separation index (OSI) [29].

The validity indices cited above are listed in appendices B to E which provide for each a reference, the name, the acronym used in this paper, the mathematical formulation and the criteria used to estimate the optimal number of clusters.

C. m-rules

The majority of the previously presented validity indices are known to be relatively insensitive to the fuzziness parameter m. For these reason, numerous applications of FCM to real-life datasets [30], [31], [32], [33] assumed that m is equal to 2. Indeed, this assertion is true when the clusters fulfill the validity index assumptions, *e.g.* compactness of clusters and separation between them. Otherwise, the validity indices can have an unpredictable behavior and can reach to an impossibility for FCM to extract a meaningful clustering structure [34]. The choice of m is thus very important [23]. As stated in several studies [35], [36], [37], the m value must be optimized for each dataset since it is directly linked to the data structure, especially concerning the degree of overlapping or mixture between clusters [35].

Pal and Bezdek [38] recommended a choice of m in the range $[1.5; 2.5]$. To refine this estimation, different kinds of methods have been developed, the m-rules. In [39], the optimal value of m, denoted m^*, is determined as the argument that maximizes the membership function of a fuzzy decision, defined as the intersection between a fuzzy goal (i.e. the number of objects in each cluster) and a fuzzy constraint (i.e. the sum of squared error within each cluster). This method is named the fuzzy decision optimization (FDO). In [34], an empirical procedure, named

the Dembélé-Kastner (DK) rule which is based on the computation of the coefficient of variation, is proposed to estimate an upper bound value for m, which is named m_{ub}. Another m-rule presents the advantage to compute the optimal value m^* in function of m_{ub} and only thanks to the dataset, without any FCM result. However, this optimal m^* value is restricted to be less than or equal to 2. Two theoretical rules, named α-rule (αR) and β-rule (βR), have been recently developed in [36] and also compute an upper bound of m. Finally, in [37], the authors proposed an objective function named Okeke-Karnieli (OK) rule which has to be minimized. This function is defined as the error norm between the original dataset and the dataset reconstructed following a linear mixture model, estimated by FCM. Even based on a simple concept, this method results in a reliable estimation of m on real-life dataset. As validity indices, these 5 m-rules are detailed in appendix F.

D. $C - m$ methods

Several studies have pointed out the generation of redundant clusters into a FCM partition when this latest is obtained with a number of clusters higher than the real one, and when the m value is fixed around 2 [34], [30], [40]. In previous work, we made the same observations on infrared spectral images recorded on human skin tumoral biopsies. Afterward, the redundant clusters have been interpreted by pathologists, who concluded that no physical interpretation could be made on it. Thus, we developed a new empirical methodology to automatically and simultaneously estimate m and C values in order to avoid redundant clusters into the returned FCM partition. The proposed algorithm, Redundant Based Algorithm (RBA) was successfully but exclusively applied on infrared spectral images [41], and reveals time-consuming. As a result, we developed an improvement of this first algorithm, Upgraded Redundant Based Algorithm (URBA) based on a new redundant cluster criterion, and which improves both computational time and results accuracies when applied on a large number of real-world and simulated datasets (publication in progress).

A third C-m method, proposed by Schwämmle and Jensen [42] and called SJM, has been investigated. This method aims at first estimating an optimal m value, denoted $m\_threshold$, and from which an optimal number of clusters can then be found. With a criterion based on a minimum distance between prototypes of a FCM partition, the SJM method is able to determine an upper value of m, throughout an increase of its value, and for which a clustering solution will present at least 2 prototypes that coincide, $i.e.$ there is no need to go through higher values of m to determine the optimal one for the considered dataset. Moreover, this determination has been experimentally demonstrated to be independent of the number of clusters. Then, by considering this optimal value of fuzziness, FCM classifications are realized with a number of clusters starting from 2 and increasing with a unit step until a number of clusters equals to the square root of the number of observations. Among these classifications, this which presents a minimum distance between prototypes is pointed out, and reflects that all structures of the dataset are completely described by the current number of clusters.

E. Datasets

All these methods are tested on 49 different datasets traditionally used in literature to evaluate classification algorithms behavior, especially according to the number of clusters, the separation between them and the number of objects considered. These datasets are divided into two categories. 38 and 11 datasets for simulated and real ones respectively.

The characteristics of these datasets are listed in table I. From left to right, columns represent: the reference article where each dataset is described; the traditional name used; the real number of clusters C^* that must be estimated; the dimension of each object; the number of objects and their repartition in each cluster and finally, their probability distribution.

F. Parameters

To apply all methods in identical conditions, the FCM stopping criteria were constant, $i.e.$ the maximum number of iterations was fixed to 500, and the minimum value between two consecutive computations of the objective function (1) was fixed to 10^{-5}. Finally, FCM partition being sensitive to the random initialization of the membership matrix, 20 FCM classifications were realized for each (C,m) setting. The partition presenting the lower value of the final objective function, among the 20 replicates, is selected for the different applications of CVI, m-rules and C-m methods.

As regards the validity indices, they are supposed to be robust regardless the fuzziness parameter value, m. Thus in the literature, they are applied on partitions obtained with $m = 2$ as a standard value. In order to be consistent with these studies, $m = 2$ was considered in this paper. However, when compactness or separation of clusters are not fulfilled criteria, CVI can substantially depend on the choice of m. To analyze this behavior, we considered three other m values such as estimated by FDO, DK and OK m-rules. Some of them assume that the number of clusters is already known or estimated by validity indices, as well as a set of m values is needed. To simplify the study, we decided to fix the number of clusters C to the real number of clusters of each dataset as listed in table I, and we considered a set values for m starting from 1 to 20 with a step of 0.1.

Moreover, the FDO rule is dependent on 3 constants named A, B and C (see appendix E). The value range of these constants being not discussed in [39] or in any other article up to our knowledge, we empirically chose $A = 0.012$, $B = 0.00008$ and $C = 1$ in order to retrieve results similar to those presented in [39].

For SJM, two prototypes are considered as identical when their euclidean distance is lower than a threshold fixed to 0.1 as proposed in [42]. The setting of m starts from 1.1 and is iteratively incremented with a step value of 0.1.

For both RBA and URBA, thresholds T values are fixed to $\mathcal{T} = \{0.05, 0.1, \ldots, 0.95\}$, m is set in the same manner as for SJM.

More generally for all C-m methods and CVI, the maximum number of clusters C_{max} being dependent on the considered dataset, and the convention largely shared in literature being to fix C_{max} to $\sqrt{N}$, we used $C_{max} = [\sqrt{N}] + 1$ as upper bound for C, where $[.]$ represents the floor function.

	Reference	Name	C^*	Dimension	Number of objects	Distribution
SIMULATED DATA	[24]	Kim3	3	2	60: 20 each	Random
	[17]	Sun2	5	3	250: 50 each	Gaussian
	[17]	Sun3	10	2	500: 50 each	Gaussian
	[1]	3x3	9	2	225: 25 each	Uniform random
	[28]	4	5	3	715: 195, 4*130	Gaussian random
	[28]	5	5	3	1444: 195, 4*130 + 729 noisy objects	Gaussian random
	[28]	6	5	3	2159: 390, 4*260 + 729 noisy objects	Gaussian random
	[28]	7	8	3	400: 50 each	Gaussian random
	[28]	8	7	3	350: 50 each	Gaussian random
	[1]	Arcs	4	2	104: 26 each	Sinusoidal
	[43]	Bensaid	3	2	49: 6, 40, 3	N/A
	[44]	Butterfly	2	2	19: 9,10	Fixed
	[2]	Butterfly 2	2	2	15: 8, 7	Fixed
	[5]	Dataset_3_3	3	3	150: 50 each	Gaussian
	[5]	Dataset_4_2	4	2	200: 50 each	Gaussian
	[18]	Dataset_4_3	4	3	400: 100 each	Gaussian
	[5]	Dataset_4_noise	4	2	35: 8 each + 3 noisy objects	Gaussian
	[18]	Dataset_5_2	5	2	500: 100 each	Gaussian
	[18]	Dataset_6_2	6	2	300: 50 each	Gaussian
	[18]	Dataset_10_2	10	2	500: 50 each	Gaussian
	[5]	Dataset_15_2	15	2	225: 15 each	Gaussian
	[1]	Density	4	2	200: 50 each	Uniform random
	[12]	Example_1	3	2	16: 5, 6, 5	N/A
	[12]	Example_2	4	2	16: 4 each	N/A
	[?]	Example_2_Silouhette	5	2	1000: 200 each	Gaussian random
	[1]	Noisy	4	2	225: 4*50, 25	Uniform random
	[16]	Normal_4	4	4	800: 200 each	Gaussian
	[1]	Size	4	2	250: 2*100, 2*25	Uniform random
	[1]	Spheres	4	2	200: 50 each	Uniform random
	[1]	TnU	2	2	300: 150 each	Uniform random
	[45]	X1	4	2	300: 75 each	Gaussian random
	[45]	X2	5	2	300: 60 each	Gaussian random
	[45]	X3	6	2	300: 50 each	Gaussian random
	[45]	X4	9	2	1800: 200 each	Gaussian random
	[45]	X5	5	2	325: 25, 3*50, 150	Gaussian random
	[45]	X6	10	2	500: 50 each	Gaussian random
	[45]	X7	5	3	250: 50 each	Gaussian random
	[46]	X30	3	2	30: 10 each	N/A
REAL DATA	[47]	Bupa Liver Disorder	2	6	345: 200, 145	/
	[42]	Ecoli	8	7	336: 143, 77, 2, 2, 35, 20, 5, 52	/
	[14]	Glass	6	9	214: 70,76,17,13,9,29	/
	[42]	Ionosphere	2	34	351: 126, 225	/
	[48]	Iris	3	4	150: 50 each	/
	[?]	MMDS	2	5	830: 427, 403	/
	[42]	Serum	10	13	517: 100, 145, 34, 43, 7, 34, 14, 63, 19, 25, 33	/
	[24]	Starfield	9	2	66: 5, 5, 3, 7, 10, 8, 6, 15, 7	/
	[47]	WBC	2	9	693: 444, 239	/
	[47]	WDBC	2	30	569: 357, 212	/
	[47]	Wine	3	13	178: 59, 48	/

TABLE I: Characteristics of the simulated and real datasets used to compare validity indices, m-rules and $C-m$ methods

Usually, the optimal number of clusters $\hat{C}$ returned by either CVI or C-m methods is compared with the real number of clusters C^* composing a dataset. If these two values match, the method is considered as efficient, and unsuitable otherwise. However, this validation can only be made when the FCM algorithm, considering C^*, is able to retrieve the underlying structures of the dataset. If it is not, CVI and C-m methods will not be able to determine an optimal value equals to C^*. That is the reason why in such conditions, the estimated number of clusters must be compared to another reference. Such an evaluation methodology is proposed in [1] and computes the number of clusters, denoted C_{VI}, inducing the most similar partition to the real one. In this study, a validity index or a C-m

method will thus be considered as efficient if its estimated number of clusters $\hat{C}$ match with C_{VI}.

C_{VI} is computed according to the variation information concept, described as follow :

$$C_{VI} = \underset{2 \leqslant C \leqslant C_{max}}{\arg\min} \left(H(P_C) + H(P) - 2M_I(P_C, P) \right) \tag{4}$$

With P_C the defuzzified partition obtained by FCM and composed of C clusters, P the real partition of the data, $H(P_C)$ and $H(P)$ the entropies of partitions P_C and P respectively, and $M_I(P_C, P)$ the mutual information of partitions P_C and P. For more details and explanations, the interested reader could refers to [1]. However, it has to be noticed that C_{VI} depends on a FCM partition P_C. This partition is computed with a number of cluster C but also with a fuzziness parameter m, from which the partition can depend. Consequently, the C_{VI} references estimated for each dataset can differ in regard of the four m values considered in this study.

III. Results

The performances of validity indices, m-rules and C-m methods described in section II are presented here.

Validity indices, SJM, RBA and URBA were applied on the datasets presented in table I. The results using the 4 different fuzziness values are presented in tables III, IV, V and VI where the optimal numbers of clusters $\hat{C}$ according to each method are reported. The greyed out cells correspond to cases where the estimated number of clusters coincides with C_{VI}. The "summarized results" indicate the number of well estimated cases out of the total number of datasets available for each category (simulated and real). The last line reports the global results, $i.e.$ the amount of well estimations for both simulated and real datasets.

A. Estimation of optimal fuzziness values with the real number of clusters

The optimal fuzziness values m^*, estimated for the 49 datasets by the five m-rules and the three C-m methods are presented in table II.

No cells are greyed out since there is no gold-standard method to provide an expected m value. The description of the results is thus very subjective. The only m-rule based on theoretical proofs being βR (αR m-rule is an approximation of βR according to [36]), we considered its result has an upper-bound reference such as defined in the literature [36].

B. Estimation of $\hat{C}$ with $m = 2$

By analyzing table III, few validity indices stand out. For simulated datasets, V_P appears the best validity index since it correctly estimates C_{VI} for 30 datasets out of 38. V_{ASI} and $URBA$ are very close with 29 on 38. For real datasets, 6 CVI estimate correctly 8 datasets partitions out of 11, among which we can find $V_{WPE}, V_K, V_{GD}, V_{OS}, V_{ASI}$ and V_{OSI}. SJM, URBA and 7 CVI (namely, $V_{MPC}, V_P, V_{XB}, V_T, V_{PCAES}, V_{VI}, V_{SCBWS}$) reached the second best results with 7 efficient estimations on 11 datasets. Overall, V_P, V_{ASI} and $URBA$ appear the best techniques to estimate C_{VI} when they are applied on 49 datasets with a m value equals to 2. The two CVI present a score of 37 on 49 while URBA correctly estimated the number of clusters for 36 datasets on 49.

	Dataset\m-rule	FDO	DK	αR	βR	OK	SJM	RBA	URBA
SIMULATED DATA	**Kim3**	2.6	2	$+\infty$	$+\infty$	1.6	3.41	2.825	2.6
	Sun2	3.5	2	3	7.09	1.4	2.93	2.267	2.24
	Sun3	4.1	2	$+\infty$	$+\infty$	1.5	2.72	2.23	2.169
	3x3	1.1	2	$+\infty$	$+\infty$	1.8	4.99	3.238	2.509
	4	1.1	2	3	$+\infty$	1.6	1.5	2.233	1.92
	5	1.9	2	3	25.917	1.3	1.63	1.544	2.56
	6	1.6	2	3	$+\infty$	1.2	1.53	2.096	2.014
	7	3.1	2	3	4.797	1.5	1.76	2.242	1.61
	8	3.4	2	3	7.973	1.5	1.67	2.327	1.6
	Arcs	5.2	2	$+\infty$	$+\infty$	1.6	6.97	3.801	2.466
	Bensaid	18.1	2	$+\infty$	$+\infty$	1.1	6.19	4.808	4.17
	Butterfly	1.1	2	$+\infty$	$+\infty$	1.4	4.42	5.526	5.562
	Butterfly 2	1.1	2	$+\infty$	$+\infty$	1.2	4.97	3.735	3.538
	Dataset_3_3	11	2	3	$+\infty$	1.5	3.12	2.775	1.824
	Dataset_4_2	16.8	2	$+\infty$	$+\infty$	1.5	2.81	2.689	2.443
	Dataset_4_3	1.1	2	3	$+\infty$	1.8	1.99	5.605	1.715
	Dataset_4noise	2.1	2	$+\infty$	$+\infty$	1.8	3.4	3.11	4.194
	Dataset_5_2	5.4	2	$+\infty$	$+\infty$	1.6	2.67	2.543	2.168
	Dataset_6_2	1.1	2	$+\infty$	$+\infty$	1.8	1.1	2.63	2.166
	Dataset_10_2	1.1	2	$+\infty$	$+\infty$	1.7	2.4	2.672	2.278
	Dataset_15_2	1.1	2	$+\infty$	$+\infty$	1.9	3.12	2.586	2.557
	Density	1.1	2	$+\infty$	$+\infty$	1.9	5.12	9.857	2.393
	Example_1	1.1	2	$+\infty$	40	1.7	5.83	4.353	5.338
	Example_2	1.1	2	$+\infty$	$+\infty$	1.8	4.3	4.397	5.301
	Exemple_2_Silhouette	1.3	2	$+\infty$	$+\infty$	1.6	3.33	2.806	2.235
	Noisy	1.1	2	$+\infty$	$+\infty$	1.5	5.82	8.947	2.754
	Normal_4	15.3	1.948	2	3.259	1.5	3	2.326	1.993
	Size	17.1	2	$+\infty$	$+\infty$	1.8	4.49	2.427	1.993
	Spheres	1.1	2	$+\infty$	$+\infty$	1.8	6.27	3.507	2.925
	TnU	2.8	2	$+\infty$	$+\infty$	1.5	6.48	2.685	2.717
	X1	1.1	2	$+\infty$	$+\infty$	1.7	5.56	10.64	2.03
	X2	1.1	2	$+\infty$	$+\infty$	1.6	6.31	2.759	1.891
	X3	2.7	2	$+\infty$	$+\infty$	1.7	5.75	2.071	2.378
	X4	2.1	2	$+\infty$	$+\infty$	1.6	6.56	2.194	2.109
	X5	17.4	2	$+\infty$	$+\infty$	1.6	4.1	3.21	2.522
	X6	3.7	2	$+\infty$	$+\infty$	1.6	3.27	2.679	2.329
	X7	2.1	2	3	$+\infty$	1.3	2.39	2.456	2.117
	X30	1.1	2	$+\infty$	$+\infty$	1.8	1.1	2.568	4.861
REAL DATA	**Bupa Liver Disorder**	19.1	1.663	1.5	2.9251	1.2	1.57	561	1.797
	Ecoli	2	1.534	1.4	$+\infty$	1.3	1.44	2.002	1.72
	Glass	3.2	1.680	1.286	$+\infty$	1.3	1.48	2.084	2.114
	Ionosphere	13.5	1.161	1.063	7.439	1.2	1.12	1.252	1.352
	Iris	1.1	2	2	$+\infty$	1.8	3.14	1.956	2.146
	MMDS	2.6	2	1.667	$+\infty$	1.5	2.13	2.197	2.263
	Serum	1.7	1.249	1.182	4.065	1.2	1.24	1.867	1.675
	Starfield	1.1	2	$+\infty$	$+\infty$	1.6	3.01	2.453	3.71
	WBC	1.8	1.71	1.29	$+\infty$	1.5	1.36	10.32	1.478
	WDBC	1.2	1.213	1.071	5.169	1.3	1.13	1.928	1.562
	Wine	2.6	1.25	1.18	3.52	1.3	1.27	2.586	1.621

TABLE II: Fuzzy parameter estimated by classical m-rules (with a number of clusters C fixed to the real value), SJM, RBA and URBA (with C and m simultaneously estimated without any *a priori* information) for simulated and real datasets.

C. Estimation of $\hat{C}$ with m values assessed by m-rules

For one dataset, a m-rule gives most often a different m value than the others. In order to compare their performances, we applied each validity index to datasets with the m-values estimated by the FDO, DK and OK m-rules. The number of estimated clusters are respectively summarized in tables IV, V and VI. Hence, this study provides results on the robustness of the validity indices to m. It is important to notice here that SJM, RBA and URBA results are identical throughout these tables since these algorithms estimate themselves their optimal m value.

For m values estimated by FDO, the best validity index for the simulated datasets is V_{ASI} with 26 out of 38 which have their number of clusters correctly estimated, V_{EASI} succeeds for 25 over 38 and URBA for 22 out of 38. 8 real datasets over 11 were correctly partitioned with V_{ASI} again, V_{PCAES} and URBA are close with 7 on 11. Globally, V_{ASI} appears the best validity index when m values are computed by FDO, with a good estimation of C_{VI} for 34 datasets on 49. The second one is V_{EASI} with 31 on 49 and the third is URBA with only 29 good estimations on 49 tested.

The m values estimated by the DK m-rule being always equal to 2 for the simulated datasets, the results of table V are identical to those of table III for this category. However, slight differences exist for the real datasets. Indeed, V_{KYI} and V_{PCAES} well estimate 8 datasets partitions over 11. Then, URBA and 6 CVI (namely, $V_{PC}, V_{WPE}, V_{MPC}, V_{SCBWS}, V_{GIISI}$ and V_{OSI}) obtained a score of 7 on 11. The third rank is reached with 6 datasets on 11 and was obtained by the SJM method and 6 CVI (namely, $V_{PE}, V_P, V_K, V_T, V_{OS}$ and V_{ASI}). The global results show that V_P and URBA are the best techniques for m-values determined by DK with 36 good estimations on 49. The second ones are V_{MPC} and V_{ASI} with 35 on 49, and finally V_{GIISI} and V_{OSI} reached to a score of 34 on 49.

Finally, m values estimated by the OK m-rule obtained a highest score of 27 datasets on 38 for simulated ones. This result concerns URBA and 4 CVI (namely, V_{MPC}, V_P, V_{FHV} and V_{ASI}). Then V_{GIISI} got a score of 26 on 38 and both V_{VI} and V_{EASI} gave 25 on 38. For real datasets, V_{OSI} obtained 8 on 11, followed by SJM, URBA and 6 CVI (namely, $V_{WPE}, V_{MPC}, V_P, V_{XB}, V_{OS}$ and V_{ASI}) with 7 on 11. Overall, URBA, V_P, V_{ASI} and V_{MPC} obtained 34 good estimations on 49 datasets, followed by V_{FHV} and V_{GIISI} (with 32 on 49) and finally V_{EASI} got 31 on 49.

IV. Discussion

The majority of the presented validity indices are said to be robust to m. To assess this property, they were applied to the same m range as used for $C - m$ methods, *i.e.* $m \in \{1.1, 1.2, \ldots, 20\}$. The optimal number of clusters was badly estimated by the majority of the validity indices and for both simulated and real datasets (results not shown). These results are in accordance with a previous study [34] which showed that m value also need to be adaptively estimated according to the considered dataset. Moreover, in [34], FCM with $m = 2$ failed to extract real clusters from microarray datasets since redundant clusters were estimated. Furthermore, the m range for which validity indices are robust is narrow and does not include $m = 2$.

	Dataset \ Validity indices	PC	PE	WPE	MPC	KYI	P	FS	XB	K	T	STC	FHV	APD	PD	PCAES	VI	CWB	WSJ	PBMF	SCBWS	GD	OT	OS	FSC	ASI	EASI	GHSI	PDRDR	OSI	SJM	RBA	URBA
SIMULATED DATA	Kim 3 (3,3)	2	2	8	3	2	3	8	3	3	3	3	2	2	2	3	8	3	3	8	8	3	3	3	7	3	3	2	4	2	2	6	3
	Sun 2 (5,5)	2	2	5	5	2	5	16	5	5	5	2	2	5	5	5	5	5	5	16	5	5	4	5	5	5	5	5	10	5	5	7	5
	Sun3 (10,10)	2	2	3	10	2	10	10	5	5	5	10	2	2	2	4	10	5	7	8	10	10	2	10	10	10	10	10	22	10	10	10	10
	3x3 (9,9)	9	9	4	9	9	9	9	9	9	9	11	9	9	9	9	9	9	9	9	9	9	9	9	9	9	9	9	9	9	6	9	9
	4 (5,5)	2	2	2	4	2	4	4	4	4	4	5	4	5	5	3	5	4	5	5	4	4	4	2	5	4	4	4	6	4	2	6	6
	5 (5,2)	2	2	2	4	2	4	9	4	4	4	2	2	2	2	6	9	7	7	33	4	4	4	2	7	2	7	2	11	4	6	10	2
	6 (5,6)	2	2	2	4	2	4	6	4	4	4	2	2	4	4	3	4	6	6	46	4	4	4	2	6	2	4	2	11	2	4	6	6
	7 (8,8)	2	2	4	7	2	7	9	8	8	8	2	2	7	4	4	8	8	8	19	8	8	4	6	8	8	8	7	13	8	8	3	8
	8 (7,7)	2	2	4	7	7	7	8	7	7	7	2	2	7	5	4	7	7	7	16	7	7	3	6	7	7	7	7	18	7	7	3	7
	Ares (4,2)	2	2	7	2	2	2	11	8	8	2	2	2	2	2	6	11	2	2	10	11	11	2	2	11	2	11	2	6	2	2	8	2
	Bensaid (3,4)	2	2	8	5	4	4	8	7	5	5	5	2	7	7	6	7	5	5	8	8	6	2	4	5	5	5	6	3	6	3	7	7
	Butterfly (2,2)	2	2	5	4	2	4	4	5	4	4	2	2	4	4	5	4	4	4	5	4	4	2	4	4	2	4	2	4	2	2	4	4
	Butterfly 2 (2,2)	2	2	3	2	2	2	4	2	2	2	2	2	2	2	2	4	2	2	3	2	2	2	2	3	2	2	2	3	2	2	3	3
	DataSet_3_3 (3,3)	2	2	9	3	3	3	3	3	3	3	3	2	3	3	3	3	3	3	8	3	3	3	3	3	3	3	3	5	3	3	4	3
	DataSet_4_2 (4,4)	2	2	4	4	2	4	14	4	4	4	4	2	2	2	4	15	4	4	11	15	4	2	4	7	4	4	4	8	4	4	8	4
	DataSet_4_3 (4,4)	4	2	2	4	4	4	4	4	4	4	4	4	4	4	3	4	4	4	4	4	4	4	4	4	4	4	4	10	4	4	4	4
	DataSet_4_noise (4,6)	3	3	7	5	6	6	7	4	4	4	6	3	5	3	3	6	4	4	7	7	6	3	4	6	4	4	6	5	6	3	6	4
	DataSet_5_2 (5,5)	4	2	3	5	5	5	5	4	4	4	5	2	5	5	4	4	4	5	14	5	4	4	5	5	5	5	5	13	5	5	6	5
	DataSet_6_2 (6,6)	6	6	3	6	6	6	6	4	4	4	6	6	6	6	4	6	4	4	12	6	4	6	6	6	4	4	4	8	5	4	7	6
	DataSet_10_2 (10,10)	2	2	3	10	10	10	10	10	10	10	11	2	10	10	5	10	5	6	11	10	10	2	10	10	10	10	10	10	10	2	10	10
	DataSet_15_2 (15,15)	15	15	4	15	15	15	15	15	15	15	15	15	15	15	15	15	8	13	15	16	15	14	15	15	15	15	15	13	15	4	15	15
	Density (4,4)	4	4	3	4	4	4	5	4	4	4	15	4	10	6	4	4	4	4	15	4	4	4	4	4	4	4	4	8	4	4	4	4
	Example_1 (3,3)	3	3	5	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	5	3	3	3	3	3	3	3	3	3	3	3	4	4
	Example_2 (4,4)	4	4	5	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	3	4	4	4	4
	Example_2_silhouette (5,5)	5	2	2	5	5	5	5	3	3	3	5	5	5	5	5	5	5	5	14	5	3	5	5	5	5	5	5	24	5	3	7	5
	Noisy (4,4)	4	4	3	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	13	4	4	4	4	4	4	4	4	4	4	4	4	4
	Normal 4 (4,4)	4	2	3	4	4	4	5	4	4	4	4	2	14	7	4	4	4	4	8	4	4	8	4	4	4	4	4	20	4	4	4	4
	Size (4,4)	4	4	2	4	4	4	4	4	4	4	14	6	4	6	2	4	4	4	12	4	4	2	4	4	4	4	4	4	4	4	8	4
	Spheres (4,4)	4	4	3	4	4	4	4	4	4	4	15	4	4	4	4	4	4	4	8	4	4	4	4	4	4	4	4	4	4	4	5	4
	TnU (2,2)	2	2	3	9	2	11	9	9	9	9	9	2	2	2	5	11	5	5	8	18	11	2	8	12	11	11	11	13	11	2	11	10
	X1 (4,4)	2	2	2	2	2	2	2	2	2	2	5	2	4	4	4	4	2	4	9	4	2	3	2	5	2	2	2	15	2	2	4	4
	X2 (5,5)	2	2	3	5	2	5	7	5	5	5	5	2	5	5	4	5	4	5	11	5	5	5	5	5	5	5	5	8	5	2	6	5
	X3 (6,6)	2	2	3	6	2	6	6	6	6	6	6	2	6	6	4	6	4	6	14	6	6	2	2	6	6	6	2	6	6	2	7	6
	X4 (9,9)	3	2	2	3	3	3	8	3	3	3	11	3	3	3	3	8	4	8	10	9	3	3	3	9	3	3	3	38	3	3	9	9
	X5 (5,5)	2	2	3	5	2	5	5	3	3	3	6	2	6	7	2	6	5	6	7	7	6	4	2	6	6	6	2	18	6	3	7	7
	X6 (10,10)	2	2	3	10	2	10	10	5	5	5	5	2	5	3	5	10	5	8	7	10	10	2	10	10	10	10	10	17	10	10	10	10
	X7 (5,2)	2	2	8	2	2	2	15	11	11	11	2	2	2	2	6	15	8	7	16	2	2	2	2	6	2	2	2	11	2	2	4	2
	X30 (3,3)	3	3	6	3	3	3	5	3	2	2	6	3	6	3	3	3	3	3	6	5	3	3	3	3	3	3	3	4	3	2	4	5
	Summarized results	19/38	16/38	3/38	28/38	25/38	30/38	18/38	21/38	20/38	21/38	22/38	17/38	24/38	23/38	16/38	25/38	21/38	25/38	5/38	25/38	25/38	18/38	27/38	27/38	29/38	26/38	27/38	7/38	27/38	21/38	14/38	29/38
REAL DATA	Bupa Liver Disorder (2,2)	2	2	2	2	2	2	19	2	2	2	2	2	2	2	2	2	4	9	11	18	2	2	2	5	2	2	2	9	2	2	2	2
	Ecoli (8,3)	2	2	6	2	2	3	5	3	3	3	2	2	3	2	3	3	3	3	18	3	3	3	3	5	3	3	2	16	3	3	4	3
	Glass (6,3)	2	2	3	3	2	3	3	3	3	3	3	2	2	15	2	6	3	4	14	4	3	5	3	4	3	3	2	7	3	3	10	4
	Ionosphere (2,2)	2	2	14	2	2	2	19	2	2	2	2	2	2	2	2	2	12	13	19	2	2	2	2	6	2	3	2	5	2	4	8	8
	Iris (3,3)	2	2	3	2	2	2	2	2	2	2	5	2	12	2	2	3	2	3	8	3	2	5	2	3	2	2	2	6	2	2	8	3
	MMDS (2,2)	2	2	2	2	2	5	6	4	4	4	7	2	15	15	2	8	4	6	12	8	2	15	2	7	2	2	2	8	2	2	9	5
	Serum (10,5)	2	2	6	2	2	2	7	2	2	2	2	4	15	4	4	5	17	17	19	3	3	3	2	6	2	2	2	12	2	5	5	5
	Starfield (9,8)	2	2	8	9	9	9	9	9	8	5	9	2	9	9	4	9	5	5	9	9	9	2	9	9	9	9	9	6	9	9	9	9
	WBC (2,2)	2	2	2	2	2	2	4	2	2	2	2	2	4	2	2	2	24	24	8	2	2	2	2	6	2	4	2	5	2	2	2	2
	WDBC (2,2)	2	2	2	2	2	2	24	2	2	2	2	2	2	2	2	2	20	23	25	2	2	2	2	6	2	2	2	8	2	3	3	2
	Wine (3,3)	2	2	3	3	2	3	14	3	3	3	2	2	2	2	3	3	13	13	4	3	3	3	3	6	3	4	2	8	3	3	3	3
	Summarized results	5/11	5/11	8/11	7/11	5/11	7/11	1/11	7/11	8/11	7/11	5/11	5/11	4/11	4/11	7/11	7/11	2/11	2/11	0/11	7/11	8/11	6/11	8/11	1/11	8/11	5/11	5/11	0/11	8/11	7/11	4/11	7/11
	Global results	24/49	21/49	11/49	35/49	30/49	37/49	19/49	28/49	28/49	28/49	27/49	22/49	28/49	27/49	23/49	32/49	23/49	27/49	5/49	32/49	33/49	24/49	35/49	28/49	37/49	31/49	32/49	7/49	35/49	28/49	18/49	36/49

TABLE III: Number of clusters estimated by validity indices with a fuzzy parameter m equal to 2. "Summarized results" show the amount of good estimations per validity index for each data category (i.e. "Simulated" and "Real"). "Global results" provides the full results considering both "Simulated" and " Real" datasets. SJM, RBA and URBA results are also shown. The greyed out cells highlight indices estimations equals to C_{VI} ones. C^* and C_{VI} values are respectively indicated in parenthesis next to datasets names.

	Dataset\ Validity indices	PC	PE	WPE	MPC	KTI	P	FS	XB	K	T	STC	FHV	APD	PD	PCAES	VI	CWB	WSJ	PBMF	SCBWS	GD	OT	OS	FSC	ASI	EASI	GHSI	PDRDR	OSI	SJM	RBA	URBA
SIMULATED DATA	**Kim 3 (3,3)**	2	2	8	2	2	3	6	3	3	3	7	2	2	2	3	8	3	3	8	8	3	3	2	4	2	3	2	3	3	2	6	3
	Sun 2 (5,6)	2	2	16	15	2	5	16	8	5	5	16	2	2	2	5	15	8	8	16	15	5	5	15	5	5	5	5	5	5	5	7	5
	Sun3 (10,11)	2	2	15	4	2	6	5	8	6	6	23	2	2	2	5	20	8	11	23	23	4	3	6	4	11	7	2	5	8	10	10	10
	3x3 (9,9)	9	9	2	9	9	9	14	9	9	9	11	16	15	16	9	9	9	9	9	10	9	9	3	10	9	9	9	9	9	6	9	9
	4 (5,5)	2	2	2	2	2	2	27	4	4	4	4	5	5	5	3	4	4	4	4	5	4	5	2	27	4	4	4	7	2	2	6	6
	5 (5,2)	2	2	2	4	2	4	10	4	4	4	2	2	2	4	6	8	8	8	22	4	8	4	2	8	2	8	2	13	2	6	10	2
	6 (5,3)	2	2	2	4	2	3	9	3	3	4	2	2	8	4	6	4	8	8	4	4	9	5	9	9	3	2	44	2	4	6	6	6
	7 (8,4)	2	2	16	4	2	4	8	7	7	7	8	2	4	4	5	20	7	8	21	4	5	4	12	6	7	7	9	10	7	8	3	8
	8 (7,7)	2	2	18	5	2	5	7	5	4	4	19	2	3	3	5	19	5	7	19	19	5	4	9	5	7	7	5	10	7	7	3	7
	Ares (4,2)	2	2	11	11	2	2	11	8	2	2	11	2	2	2	3	11	5	7	11	11	2	2	11	3	2	2	2	3	2	2	8	2
	Bensaid (3,4)	2	2	2	4	2	6	4	8	2	2	8	2	2	2	3	8	3	3	8	2	2	2	6	3	6	6	2	3	6	3	7	7
	Butterfly (2,2)	5	5	2	5	5	5	5	4	4	4	5	5	4	4	4	4	2	2	4	5	4	4	3	5	2	2	2	3	5	2	4	4
	Butterfly 2 (2,2)	3	3	2	4	4	4	4	2	2	2	4	4	4	4	2	4	2	2	2	4	2	2	2	4	2	2	2	3	3	2	3	3
	DataSet_3_3 (3,3)	2	2	2	3	2	3	13	13	3	3	13	2	2	2	9	13	3	4	13	2	3	3	5	3	3	3	3	9	3	3	4	3
	DataSet_4_2 (4,4)	2	2	2	2	2	2	7	15	2	2	15	2	2	2	13	15	7	7	15	2	2	2	11	3	2	2	2	3	2	4	8	4
	DataSet_4_3 (4,4)	4	4	2	4	4	4	12	4	4	4	4	4	4	4	4	5	3	4	4	4	4	4	2	4	4	4	4	20	4	4	4	4
	DataSet_4_noise (4,5)	3	2	7	5	5	5	6	3	3	3	7	3	5	4	4	5	3	3	7	7	5	3	5	5	5	5	5	3	5	3	6	4
	DataSet_5_2 (5,5)	2	2	16	4	2	4	4	5	4	4	16	2	2	2	4	16	14	14	23	4	4	4	4	5	5	5	4	9	5	5	6	5
	DataSet_6_2 (6,6)	4	4	2	4	4	4	13	4	4	4	6	6	8	8	6	6	4	4	6	6	4	4	2	5	4	4	4	17	4	7	7	6
	DataSet_10_2 (10,10)	7	8	2	8	7	7	14	10	10	10	12	10	10	10	10	10	5	5	10	23	10	2	6	21	10	10	10	16	7	2	10	10
	DataSet_15_2 (15,16)	14	14	2	14	14	14	16	12	12	12	16	16	16	16	4	16	6	11	16	16	4	2	16	16	16	16	5	14	4	15	15	
	Density (4,4)	4	4	2	4	4	4	15	4	4	4	12	15	4	12	4	4	4	4	4	15	4	2	4	4	4	4	4	11	4	4	4	4
	Example_1 (3,3)	4	4	2	4	4	4	5	3	3	3	5	5	4	4	3	3	3	3	3	4	3	2	5	3	3	3	3	3	4	3	4	4
	Example_2 (4,4)	4	4	3	4	4	4	5	4	4	4	5	5	5	5	4	4	4	4	4	5	4	4	3	5	4	4	4	3	4	4	4	4
	Example_2_silhouette (5,5)	5	5	2	5	5	5	19	3	3	3	6	5	5	5	5	5	3	5	5	8	3	5	5	31	5	5	5	27	5	3	7	5
	Noisy (4,4)	4	4	2	4	7	4	14	4	4	4	4	15	13	4	4	4	4	4	4	4	4	4	2	7	4	4	4	13	4	4	4	4
	Normal 4 (4,8)	2	2	2	9	2	9	29	28	10	4	29	2	5	5	29	29	11	21	29	2	9	2	7	3	6	6	9	25	9	4	4	4
	Size (4,3)	2	2	2	2	2	2	2	14	2	2	16	2	2	2	3	16	14	14	16	2	2	2	10	3	3	3	2	3	2	4	8	4
	Spheres (4,4)	4	4	2	4	4	4	11	4	4	4	4	15	11	11	4	4	4	4	4	15	4	4	2	4	4	4	4	12	4	4	5	4
	TnU (2,2)	2	2	7	2	2	2	7	10	8	8	12	2	2	2	5	18	8	10	18	12	5	2	8	10	10	2	15	10	2	11	10	
	X1 (4,4)	2	2	2	2	2	2	18	2	2	2	5	4	18	4	4	4	2	3	4	4	2	4	2	4	2	2	6	2	2	4	4	
	X2 (5,5)	5	5	2	5	18	6	10	5	5	5	4	5	5	5	5	5	3	4	5	5	5	5	3	18	5	5	15	6	2	6	5	
	X3 (6,6)	2	2	4	4	2	2	5	6	6	6	17	2	2	2	4	6	6	6	18	6	2	2	2	6	6	6	6	6	6	2	7	6
	X4 (9,9)	3	2	2	3	3	3	8	3	3	3	12	3	3	3	3	8	8	8	10	9	3	3	3	9	3	3	3	17	3	3	9	9
	X5 (5,6)	2	2	2	2	2	2	4	18	2	2	19	2	2	2	6	18	14	14	19	2	2	2	8	3	2	2	2	7	2	3	7	7
	X6 (10,10)	2	2	11	2	2	2	5	10	5	5	23	2	2	2	5	19	10	15	23	23	2	2	8	5	10	10	2	8	10	10	10	10
	X7 (5,2)	2	2	9	2	2	2	16	13	2	2	2	2	2	2	6	14	7	7	16	2	2	2	2	6	2	2	2	13	2	2	4	2
	X30 (3,3)	3	3	2	3	3	3	5	3	2	2	6	6	6	6	3	3	2	3	4	6	3	3	2	5	3	3	3	4	3	2	4	5
	Summarized results	13/38	13/38	3/38	15/38	12/38	16/38	3/38	16/38	16/38	15/38	7/38	12/38	12/38	12/38	17/38	16/38	12/38	15/38	14/38	11/38	17/38	20/38	6/38	10/38	26/38	25/38	19/38	5/38	18/38	16/38	9/38	22/38
REAL DATA	**Bupa Liver Disorder (2,2)**	2	2	2	10	2	8	19	19	7	7	19	2	2	2	15	19	7	7	19	2	7	2	12	3	2	2	8	3	8	2	2	2
	Ecoli (8,3)	2	2	6	2	2	3	5	3	3	3	2	2	3	2	3	3	3	3	18	3	3	3	3	5	3	3	2	16	3	3	4	3
	Glass (6,2)	2	2	15	2	2	2	4	2	2	2	4	2	2	2	2	15	6	6	15	2	2	3	2	4	2	2	2	7	2	3	10	4
	Ionosphere (2,2)	2	2	2	6	2	6	19	17	11	11	19	2	2	2	10	19	11	11	19	2	11	14	6	3	2	2	14	12	14	4	8	8
	Iris (3,3)	2	2	2	2	2	2	13	2	2	2	3	13	11	13	3	2	2	2	3	13	2	3	2	11	2	2	2	4	2	2	8	3
	MMDS (2,2)	2	2	2	2	2	2	5	4	2	2	29	2	2	2	2	27	5	8	29	5	2	3	2	5	2	2	2	5	2	2	9	5
	Serum (10,5)	2	2	6	2	2	2	7	2	2	2	2	4	15	4	4	5	10	11	3	3	2	2	2	5	2	2	2	15	2	5	5	5
	Starfield (9,8)	10	10	2	10	10	10	10	9	9	5	10	10	10	10	4	10	3	5	9	10	10	2	8	10	10	10	10	3	10	9	9	9
	WBC (2,2)	2	2	2	2	2	2	4	2	2	2	2	2	8	2	2	2	12	18	2	2	2	2	2	5	2	4	2	11	2	2	2	2
	WDBC (2,2)	2	2	2	2	2	2	19	2	2	2	2	24	24	24	2	10	3	3	2	2	2	2	3	24	2	2	2	14	2	3	3	2
	Wine (3,3)	2	2	2	2	2	3	14	3	3	3	2	2	2	2	3	5	8	8	14	2	2	2	14	4	3	5	2	3	2	3	3	3
	Summarized results	6/11	6/11	5/11	4/11	6/11	6/11	0/11	5/11	6/11	6/11	3/11	5/11	5/11	5/11	7/11	3/11	1/11	1/11	3/11	6/11	5/11	5/11	5/11	1/11	8/11	6/11	4/11	1/11	5/11	6/11	4/11	7/11
	Global results	19/49	19/49	8/49	19/49	18/49	22/49	3/49	21/49	22/49	21/49	10/49	17/49	17/49	17/49	24/49	19/49	13/49	16/49	17/49	17/49	22/49	25/49	11/49	11/49	34/49	31/49	23/49	6/49	23/49	22/49	13/49	29/49

TABLE IV: Number of clusters estimated by validity indices with a fuzzy parameter m estimated by FDO m-rule. "Summarized results" show the amount of good estimations per validity index for each data category (i.e. "Simulated" and "Real"). "Global results" provides the full results considering both "Simulated" and " Real" datasets. SJM, RBA and URBA results are also shown. The greyed out cells highlight indices estimations equals to C_{VI} ones. C^* and C_{VI} values are respectively indicated in parenthesis next to datasets names.

	Dataset \ Validity indices	PC	PE	WPE	MPC	KYI	P	FS	XB	K	T	STC	FHV	APD	PD	PCAES	VI	CWB	WSJ	PBMF	SCBWS	GD	OT	OS	FSC	ASI	EASI	GISI	PDRDR	OSI	SJM	RBA	URBA
SIMULATED DATA	**Kim 3 (3,3)**	2	2	8	3	2	3	8	3	3	3	3	2	2	2	3	8	3	3	8	8	3	3	3	7	3	3	2	4	2	2	6	3
	Sun 2 (5,5)	2	2	5	5	2	5	16	5	5	5	2	2	2	5	5	5	5	5	16	5	5	4	5	5	5	5	5	10	5	5	7	5
	Sun3 (10,10)	2	2	3	10	2	10	10	5	5	5	10	2	2	2	4	10	5	7	8	10	10	2	10	10	10	10	10	22	10	10	10	10
	3x3 (9,9)	9	9	4	9	9	9	9	9	9	9	11	9	9	9	9	9	9	9	9	9	9	9	9	9	9	9	9	9	9	6	9	9
	4 (5,5)	2	2	2	4	2	4	4	4	4	4	5	4	5	5	3	5	4	5	5	4	4	4	2	5	4	4	4	6	4	2	6	6
	5 (5,2)	2	2	2	4	2	4	9	4	4	4	2	2	2	2	6	9	7	7	33	4	4	4	2	7	2	7	2	11	4	6	10	2
	6 (5,6)	2	2	2	4	2	4	6	4	4	4	2	2	4	4	3	4	6	6	46	4	8	4	4	2	6	2	4	2	11	4	6	6
	7 (8,8)	2	2	4	7	2	7	9	8	8	8	2	2	7	4	4	8	8	8	19	8	8	4	6	8	8	8	7	13	8	8	3	8
	8 (7,7)	2	2	4	7	2	7	8	7	7	7	2	2	7	5	4	7	7	7	16	7	7	3	6	7	7	7	7	18	7	7	3	7
	Ares (4,2)	2	2	7	2	2	2	11	8	8	2	2	2	2	2	6	11	2	2	10	11	11	2	2	11	2	11	2	6	2	2	8	2
	Bensaid (3,4)	2	2	8	5	4	4	8	7	5	5	5	2	7	7	6	7	5	5	8	8	6	2	4	5	5	5	6	3	6	3	7	7
	Butterfly (2,2)	2	2	5	4	2	4	4	5	4	4	2	2	4	4	5	4	4	4	5	4	4	2	4	4	2	4	2	4	2	2	4	4
	Butterfly 2 (2,2)	2	2	3	2	2	2	4	2	2	2	2	2	2	2	2	4	2	2	3	2	2	2	2	3	2	2	2	3	2	2	3	3
	DataSet_3_3 (3,3)	2	2	9	3	3	3	3	3	3	3	3	2	3	3	3	3	3	3	8	3	3	3	3	3	3	3	3	5	3	3	4	3
	DataSet_4_2 (4,4)	2	2	4	4	2	4	14	4	4	4	4	2	2	2	4	15	4	4	11	15	4	2	4	7	4	4	4	8	4	4	8	4
	DataSet_4_3 (4,4)	4	2	2	4	2	4	4	4	4	4	4	4	4	4	3	4	4	4	4	4	4	4	4	4	4	4	4	10	4	4	4	4
	DataSet_4_noise (4,6)	3	3	7	5	6	6	7	4	4	4	6	3	5	3	3	6	4	4	7	7	6	3	4	6	4	4	6	5	6	3	6	4
	DataSet_5_2 (5,5)	4	2	3	5	5	5	5	4	4	4	5	2	5	5	4	4	4	5	14	5	4	4	5	5	5	5	5	13	5	5	6	5
	DataSet_6_2 (6,6)	6	6	3	6	6	6	6	4	4	4	6	6	6	6	4	6	4	4	12	6	6	6	6	6	4	4	4	8	5	4	7	6
	DataSet_10_2 (10,10)	2	2	3	10	10	10	10	10	10	10	11	2	10	10	5	10	5	6	11	10	10	2	10	10	10	10	10	10	10	2	10	10
	DataSet_15_2 (15,15)	15	15	4	15	15	15	15	15	15	15	15	15	15	15	15	15	8	13	15	16	15	14	15	15	15	15	15	13	15	4	15	15
	Density (4,4)	4	4	3	4	4	4	5	4	4	4	15	4	10	6	4	4	4	4	15	4	4	4	4	4	4	4	4	8	4	4	4	4
	Example_1 (3,3)	3	3	5	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	5	3	3	3	3	3	3	3	3	3	3	3	4	4
	Example_2 (4,4)	4	4	5	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	5	4	4	4	4	4	4	4	3	4	4	4	4
	Example_2_silhouette (5,5)	5	2	2	5	5	5	5	3	3	3	5	5	5	5	5	5	5	5	14	5	3	5	5	5	5	5	5	24	5	3	7	5
	Noisy (4,4)	4	4	3	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	13	4	4	4	4	4	4	4	4	4	4	4	4	4
	Normal 4 (4,4)	4	2	3	4	4	4	26	4	4	4	4	4	4	4	4	4	4	4	8	4	4	8	4	4	4	4	4	23	4	4	4	4
	Size (4,4)	4	4	2	4	4	4	4	4	4	4	14	6	4	6	2	4	4	4	12	4	4	2	4	4	4	4	4	4	8	4	8	4
	Spheres (4,4)	4	4	3	4	4	4	4	4	4	4	15	4	4	4	4	4	4	4	8	4	4	4	4	4	4	4	4	4	5	4	5	4
	TnU (2,2)	2	2	3	9	2	11	9	9	9	9	9	2	2	2	5	11	5	5	8	18	11	2	8	12	11	11	11	13	11	2	11	10
	X1 (4,4)	2	2	2	2	2	2	2	2	2	2	5	2	4	4	4	4	2	4	9	4	2	3	2	5	2	2	2	15	2	2	4	4
	X2 (5,5)	2	2	3	5	2	5	7	5	5	5	5	2	5	5	4	5	4	5	11	5	5	5	5	5	5	5	5	8	5	2	6	5
	X3 (6,6)	2	2	3	6	2	6	6	6	6	6	6	2	6	6	4	6	4	6	14	6	6	2	2	6	6	6	2	6	6	2	7	6
	X4 (9,9)	3	2	2	3	3	3	8	3	3	3	11	3	3	3	3	8	4	8	10	9	3	3	3	9	3	3	3	38	3	3	9	9
	X5 (5,5)	2	2	3	5	2	5	5	3	3	3	6	2	6	7	2	6	5	6	7	7	6	4	2	6	6	6	2	18	6	3	7	7
	X6 (10,10)	2	2	3	10	2	10	10	5	5	5	5	2	5	3	5	10	5	8	7	10	10	2	10	10	10	10	10	17	10	10	10	10
	X7 (5,2)	2	2	8	2	2	2	15	11	11	11	2	2	2	2	6	15	8	7	16	2	2	2	2	6	2	2	2	11	2	2	4	2
	X30 (3,3)	3	3	6	3	3	3	5	3	2	2	6	3	6	3	3	3	3	3	6	5	3	3	3	3	3	3	3	4	3	2	4	5
	Summarized results	19/38	16/38	3/38	28/38	25/38	25/38	30/38	18/38	21/38	20/38	21/38	22/38	18/38	24/38	23/38	16/38	25/38	21/38	25/38	25/38	25/38	18/38	27/38	27/38	29/38	26/38	27/38	7/38	27/38	21/38	14/38	29/38
REAL DATA	**Bupa Liver Disorder (2,2)**	2	2	6	2	2	2	10	2	2	2	2	2	2	2	5	2	10	4	4	8	2	10	5	2	6	2	2	2	8	2	2	2
	Ecoli (8,3)	2	2	2	3	3	3	18	2	3	3	2	2	2	14	5	3	3	3	3	2	3	3	5	3	3	3	2	3	3	3	4	3
	Glass (6,2)	2	2	3	3	2	3	5	3	3	3	2	10	5	10	2	5	3	5	6	4	3	2	3	5	3	3	2	4	3	3	10	4
	Ionosphere (2,2)	2	2	2	2	2	2	18	2	4	2	2	2	2	2	2	4	4	4	2	2	8	2	2	18	4	4	2	7	2	4	8	8
	Iris (3,3)	2	2	3	2	2	2	2	2	2	2	5	2	12	2	2	3	2	3	8	3	2	5	2	3	2	2	2	6	2	2	8	3
	MMDS (2,2)	2	2	2	2	5	6	4	4	4	7	2	15	15	2	8	4	6	12	8	2	15	2	7	2	2	2	2	8	2	2	9	5
	Serum (10,5)	2	2	2	3	4	3	23	2	2	2	2	4	15	4	6	4	5	5	2	3	3	6	3	5	2	2	2	14	2	5	5	5
	Starfield (9,8)	2	2	8	9	9	9	9	9	8	5	9	2	9	9	4	9	5	5	9	9	9	2	9	9	9	9	9	6	9	9	9	9
	WBC (2,2)	2	2	2	2	2	2	4	2	2	2	2	2	15	2	2	2	19	21	2	2	2	2	2	5	2	4	2	7	2	2	2	2
	WDBC (2,2)	2	2	2	2	2	2	19	2	2	2	2	24	24	24	2	10	3	3	2	2	2	2	3	24	2	2	2	14	2	3	3	2
	Wine (3,3)	3	2	3	3	3	3	14	3	3	3	2	14	2	2	3	3	3	3	3	3	3	3	3	14	3	3	3	4	3	3	3	3
	Summarized results	7/11	6/11	7/11	7/11	8/11	6/11	0/11	5/11	6/11	6/11	5/11	4/11	2/11	2/11	8/11	4/11	3/11	4/11	4/11	7/11	5/11	5/11	6/11	3/11	6/11	5/11	7/11	1/11	7/11	6/11	4/11	7/11
	Global results	26/49	22/49	10/49	35/49	33/49	36/49	18/49	26/49	26/49	27/49	27/49	22/49	26/49	25/49	24/49	29/49	24/49	29/49	9/49	32/49	30/49	23/49	33/49	30/49	35/49	31/49	34/49	8/49	34/49	27/49	18/49	36/49

TABLE V: Number of clusters estimated by validity indices with a fuzzy parameter m estimated by DK m-rule. "Summarized results" show the amount of good estimations per validity index for each data category (i.e. "Simulated" and "Real"). "Global results" provides the full results considering both "Simulated" and " Real" datasets. SJM, RBA and URBA results are also shown. The greyed out cells highlight indices estimations equals to C_{VI} ones. C^* and C_{VI} values are respectively indicated in parenthesis next to datasets names.

	Dataset\ Validity indices	PC	PE	WPE	MPC	KYI	P	FS	XB	K	T	STC	FHV	APD	PD	PCAES	VI	CWB	WSJ	PBMF	SCBWS	GD	OT	OS	FSC	ASI	EASI	GHSI	PDRDR	OSI	SJM	RBA	URBA
SIMULATED DATA	**Kim 3 (3,3)**	2	2	7	3	3	3	8	3	3	3	3	3	6	6	3	4	3	3	7	8	3	3	3	8	3	3	2	7	2	2	6	3
	Sun 2 (5,5)	2	2	3	5	15	5	16	4	4	4	2	13	12	12	5	5	5	5	4	5	5	8	5	16	5	5	5	5	6	5	7	5
	Sun3 (10,10)	10	2	2	10	23	10	16	10	10	10	2	10	10	10	4	10	5	6	8	21	10	2	10	23	10	10	10	16	10	10	10	10
	3x3 (9,9)	9	9	4	9	9	9	11	9	9	9	15	9	9	9	9	9	9	9	9	9	9	9	9	9	9	9	9	9	6	9	9	9
	4 (5,5)	2	2	2	4	2	4	7	4	4	4	4	5	5	5	3	4	4	4	4	5	4	5	2	5	4	4	4	20	2	2	6	6
	5 (5,2)	2	2	2	8	34	6	36	8	8	8	2	2	9	4	5	8	7	8	2	8	8	4	6	34	7	7	7	23	2	6	10	2
	6 (5,3)	2	2	2	6	45	6	45	6	6	6	2	4	7	4	7	4	6	6	2	4	9	4	6	45	6	6	6	46	2	4	6	6
	7 (8,8)	8	2	3	8	8	8	10	8	8	8	2	8	9	9	4	8	8	8	4	8	8	4	8	8	8	8	8	19	4	8	3	8
	8 (7,7)	7	7	2	7	7	7	9	7	7	7	4	7	9	8	4	7	7	7	7	7	7	3	7	8	7	7	7	16	4	7	3	7
	Ares (4,3)	2	2	5	2	11	11	11	7	7	6	2	2	11	11	6	11	2	2	3	11	11	2	2	11	2	11	2	6	2	2	8	2
	Bensaid (3,4)	7	7	2	7	7	7	8	7	5	2	4	8	3	7	2	5	3	3	5	7	6	2	4	8	8	8	8	3	7	3	7	7
	Butterfly (2,2)	2	2	5	5	5	5	5	4	4	4	5	4	2	2	4	4	4	4	4	5	4	4	5	5	2	4	4	3	2	2	4	4
	Butterfly 2 (2,2)	3	3	4	3	4	3	4	2	2	2	4	4	4	2	2	4	2	2	2	4	2	2	3	4	2	2	2	3	3	2	3	3
	DataSet_3_3 (3,3)	3	3	3	3	3	3	13	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	4	3	3	4	3
	DataSet_4_2 (4,4)	4	2	3	4	15	4	15	4	4	4	4	4	15	15	4	6	4	4	4	15	4	2	4	15	4	4	4	7	4	4	8	4
	DataSet_4_3 (4,4)	4	4	2	4	4	4	6	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	10	4	4	4	4
	DataSet_4_noise (4,6)	4	3	7	6	6	6	7	4	4	4	7	5	6	5	3	6	4	4	5	7	6	3	4	6	4	4	6	5	3	6	4	4
	DataSet_5_2 (5,5)	5	5	2	5	5	5	22	4	4	4	5	5	5	5	5	5	4	5	5	5	5	4	5	23	5	5	5	17	5	5	6	5
	DataSet_6_2 (6,6)	6	6	2	6	6	6	6	4	4	4	8	6	6	6	4	6	4	4	6	6	6	4	6	6	4	4	4	6	4	4	7	6
	DataSet_10_2 (10,10)	10	2	2	10	10	10	13	10	10	10	11	10	10	10	5	10	5	7	10	10	10	2	10	20	10	10	10	10	2	10	10	10
	DataSet_15_2 (15,15)	15	15	4	15	15	15	15	15	15	15	15	15	15	15	15	15	10	13	15	16	15	15	15	15	15	15	15	15	4	15	15	15
	Density (4,4)	4	4	3	4	4	4	5	4	4	4	15	4	10	6	4	4	4	4	6	4	4	4	4	4	4	4	4	10	4	4	4	4
	Example_1 (3,3)	3	3	4	3	4	3	5	3	3	3	3	3	5	5	3	3	3	3	3	3	3	3	4	4	3	3	3	3	3	4	4	4
	Example_2 (4,4)	4	4	5	4	4	4	5	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	3	4	4	4	4
	Example_2_silhouette (5,5)	5	5	2	5	5	5	6	3	3	3	5	5	5	5	5	5	3	5	5	5	3	5	4	5	5	5	5	14	5	3	7	5
	Noisy (4,4)	4	4	3	4	4	4	11	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	8	4	4	4	4
	Normal 4 (4,4)	4	4	2	4	29	4	29	4	4	4	4	26	14	26	4	4	4	4	4	4	4	4	4	29	4	4	4	17	4	4	4	4
	Size (4,4)	4	4	2	4	4	4	4	4	4	4	15	6	6	8	2	4	4	4	6	4	4	2	4	4	4	4	4	4	4	4	8	4
	Spheres (4,4)	4	4	3	4	4	4	5	4	4	4	15	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	13	4	4	5	4
	TnU (2,2)	11	2	2	11	18	11	12	9	7	7	2	18	18	18	5	11	5	5	8	18	11	2	12	18	11	11	11	12	10	2	11	10
	X1 (4,4)	2	2	2	2	2	2	7	2	2	2	5	4	4	4	4	4	2	4	4	4	2	4	2	4	2	2	2	12	2	2	4	4
	X2 (5,5)	2	2	2	5	5	5	7	5	5	5	5	5	5	5	5	5	3	5	5	5	5	5	5	18	5	5	5	15	5	2	6	5
	X3 (6,6)	2	2	2	6	2	6	8	6	6	6	6	6	6	6	6	6	4	4	6	6	6	2	6	6	6	6	6	6	2	2	7	6
	X4 (9,9)	3	3	2	3	9	3	12	3	3	3	9	9	9	9	3	8	4	8	10	9	3	7	3	12	3	3	3	17	3	3	9	9
	X5 (5,5)	4	2	2	4	4	4	4	4	4	4	6	6	7	7	2	6	4	5	7	7	6	2	4	18	4	4	6	13	4	3	7	7
	X6 (10,10)	2	2	2	10	21	10	17	5	5	5	2	10	10	10	5	10	5	5	5	10	9	2	10	21	10	9	10	19	10	10	10	10
	X7 (5,2)	2	2	2	2	16	2	16	2	10	2	2	2	2	2	8	4	4	4	2	2	13	2	2	16	2	2	2	5	2	2	4	2
	X30 (3,3)	3	3	6	3	3	3	5	3	2	2	6	3	6	4	3	3	3	3	4	6	3	3	2	3	3	3	3	4	3	2	4	5
	Summarized results	23/38	20/38	4/38	27/38	20/38	27/38	3/38	22/38	20/38	21/38	18/38	27/38	21/38	21/38	20/38	25/38	17/38	22/38	22/38	24/38	24/38	21/38	23/38	17/38	27/38	25/38	26/38	9/38	22/38	20/38	13/38	27/38
REAL DATA	**Bupa Liver Disorder (2,2)**	2	2	2	2	18	2	11	2	2	2	2	19	19	18	2	5	4	4	5	4	19	4	2	19	2	2	2	7	2	2	2	2
	Ecoli (8,3)	2	2	2	3	3	3	19	2	2	2	2	19	15	15	3	3	3	3	2	15	3	3	3	15	3	3	2	5	3	3	4	3
	Glass (6,3)	2	2	2	3	8	3	6	3	3	3	2	15	15	15	2	11	3	3	6	15	15	2	3	10	3	3	2	9	3	3	10	4
	Ionosphere (2,2)	2	2	2	2	2	2	19	2	2	2	2	2	2	2	2	4	4	8	2	2	8	2	2	4	4	4	2	7	2	4	8	8
	Iris (3,3)	2	2	3	2	2	2	2	2	2	2	5	3	12	3	2	3	2	3	3	3	2	5	2	3	2	2	2	6	2	2	8	3
	MMDS (2,2)	2	2	2	5	8	5	15	2	4	4	7	2	27	16	3	6	4	5	7	10	2	7	27	2	2	2	2	25	2	2	9	5
	Serum (10,5)	2	2	2	3	5	3	23	2	2	2	2	5	9	9	6	4	5	5	2	3	3	6	3	20	2	2	2	12	2	5	5	5
	Starfield (9,8)	10	10	6	10	10	10	10	9	5	5	10	10	10	10	4	10	3	5	9	10	10	2	10	10	10	10	10	3	10	9	9	9
	WBC (2,2)	2	2	2	2	2	2	10	2	2	2	2	2	27	2	2	2	2	4	2	2	2	2	2	3	2	4	2	18	2	2	2	2
	WDBC (2,2)	2	2	2	2	2	2	22	2	2	2	2	23	24	23	2	2	3	9	2	2	2	2	2	24	2	2	2	16	2	3	3	2
	Wine (3,3)	3	2	3	3	3	3	12	3	3	2	2	12	2	2	3	3	3	3	3	3	3	3	3	12	3	3	3	3	3	3	3	3
	Summarized results	6/11	5/11	7/11	7/11	6/11	7/11	0/11	7/11	6/11	6/11	4/11	5/11	1/11	3/11	6/11	5/11	5/11	5/11	5/11	6/11	5/11	6/11	7/11	1/11	7/11	6/11	6/11	1/11	8/11	7/11	4/11	7/11
	Global results	29/49	25/49	11/49	34/49	26/49	34/49	3/49	29/49	26/49	27/49	22/49	32/49	22/49	24/49	26/49	30/49	22/49	27/49	27/49	30/49	29/49	27/49	30/49	18/49	34/49	31/49	32/49	10/49	30/49	27/49	17/49	34/49

TABLE VI: Number of clusters estimated by validity indices with a fuzzy parameter m estimated by OK m-rule. "Summarized results" show the amount of good estimations per validity index for each data category (i.e. "Simulated" and "Real"). "Global results" provides the full results considering both "Simulated" and " Real" datasets. SJM, RBA and URBA results are also shown. The greyed out cells highlight indices estimations equals to C_{VI} ones. C^* and C_{VI} values are respectively indicated in parenthesis next to datasets names.

m-rules\CVI	PC	PE	WPE	MPC	KYI	P	FS	XB	K	T	STC	FHV	APD	PD	PCAES	VI	CWB	WSJ	PBMF	SCBWS	GD	OT	OS	FSC	ASI	EASI	GIISI	PDRDR	OSI
m = 2	24	21	11	35	30	37	19	28	28	28	27	22	28	27	23	32	23	27	5	32	33	24	35	28	37	31	32	7	35
m = FDO	19	19	8	19	18	22	3	21	22	21	10	17	17	17	24	19	13	16	17	17	22	25	11	11	34	31	23	6	23
m = DK	26	22	10	35	33	36	18	26	26	27	27	22	26	25	24	29	24	29	9	32	30	23	33	30	35	31	34	8	34
m = OK	29	25	11	34	26	34	3	29	26	27	22	32	22	24	26	30	22	27	27	30	29	27	30	18	34	31	32	10	30
STD	4.2	2.5	1.41	7.85	6.5	6.95	8.96	3.56	2.52	3.2	8.02	6.29	4.86	4.35	1.26	5.8	5.07	5.91	9.71	7.23	4.65	1.71	11.03	8.88	1.41	0	4.92	1.71	5.45

TABLE VII: Standard deviation (STD) computed on global results obtained for every m values and validity indices. The greyed out cells highlight standard deviation lower than 3.

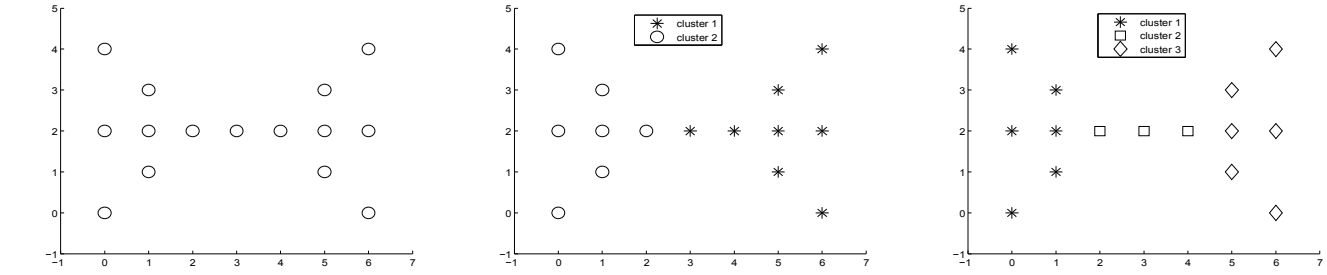
To assist in the choice of m, few m-rules have been developed (see appendix E). Table II presents the values estimated by such techniques when they are applied on the proposed datasets. It mainly shows that all are theoretically consistent since for each dataset, they estimate an optimal m value lower than the one obtained by βR, except in three cases (FDO for Normal_4, Bupa Liver Disorder and Ionosphere datasets). However, this result is of minor interest since we can see in table II that βR estimates an upper bound equal to $+\infty$ for 38 datasets out of 49. The same remark can be done for the αR. Thus, only FDO, DK and OK m-rules are considered meaningful in this study.

The combination of validity indices and these three m-rules leads to interesting results as presented in section III and summarized in table VII. Depending on m, the performances of validity indices change. For example, V_P correctly estimates the number of clusters for 37 datasets out of 49 when m is prefixed to 2, while this result decreases to 22 when m is estimated by FDO. This behavior can also be observed for other indices such as V_{OS}, V_{STC}, V_{MPC} or V_{FS} for which the number of good estimations highly decreases considering a FDO m value instead of $m = 2$. On the contrary, others validity indices show better results when m is estimated by a m-rule technique. Among these indices, V_{PBMF} presents the highest variation with a score of 5 on 49 for $m = 2$, which reaches 27 with a m value estimated by the OK m-rule. In the same conditions, V_{FHV} shows a lower improvement with an increase of 10 in its score, as well as V_{PC} and V_{PE} which increase their good estimations of 5 and 4 datasets respectively.

Only 8 validity indices over 29 (namely, $V_{PE}, V_{WPE}, V_K, V_{PCAES}, V_{OT}, V_{ASI}, V_{EASI}$ and V_{PDRDR}) are relatively not affected by the variation of m, with a standard deviation on their global results lower than 3 (table VII). However, these relatively stable indices do not all present the best results. Only V_{ASI} show the highest ones with at least 34 good estimations over 49 datasets. Overall, V_{ASI} gave the highest results among all indices for 3 m values (namely, $m = 2$, FDO and OK), and rise 2^{nd} (behind V_P and URBA) when the DK m-rule is considered.

All these observations prove that most of validity indices considered in this study are not robust in m, whether they are applied on simulated or real datasets. However, a stability in the number of datasets well estimated through different m values does not imply a stability of the same datasets in each case. Indeed, as it can be seen in table VII, V_{EASI} always performs a score of 31 on 49, but the corresponding datasets are not the same according to the m value considered.

In order to summarize, whether or not validity indices are stable, they are sensitive to m value. In addition, some m-rules, such as FDO and OK, *a priori* require the knowledge of the number of clusters. Of course, a solution



(a) (b) (c)

Fig. 1: (a) Butterfly dataset. (b) FCM clustering using OS parameters. (c) FCM clustering using URBA parameters.

CVI	PC	PE	WPE	MPC	KYI	P	FS	XB	K	T	FSC	FHV	APD	PD	PCAES	VI	CWB	WSJ	PBMF	STC	GD	OT	OS	RBA	URBA
MCT	19	19	134	19	19	19	27	23	23	23	24	22	23	23	20	25	23	23	26	22	24	28	141	52	8

TABLE VIII: Mean computational time (MCT) in seconds for 20 repetitions of cluster validity indices (CVI) on the ranges $C \in \{2, 3, \ldots, 13\}$ and $m \in \{1.1, 1.2, \ldots, 20\}$ for the Iris dataset

could be first to apply a m-rule not depending on C such as αR, βR or DK, and then apply a validity index with the estimated m value.

With αR and βR, this scheme is not applicable since these two m-rules almost always compute a m-value that tends to $+\infty$. By considering DK, we can observe that m is estimated equal to 2 for all simulated datasets (except for Normal_4 where $m = 1.948$).

For a specific dataset, the user is thus still faced to the problem of the interconnected choices of the m-rule and CVI. The legitimacy of a method constructed for the simultaneous estimation of m and C is thus completely justified.

Among the three C-m methods tested, the RBA algorithm shows the lowest results in almost every cases, since only 18 out of 49 datasets have their number of clusters correctly estimated (for its best results). Indeed, the RBA algorithm tends to overestimate the number of clusters. This is due to the fact that RBA is based on the computation of correlation coefficients which are very sensitive to outliers for small datasets. Originally, the RBA algorithm was developed to process infrared spectroscopic images composed of thousands of very similar objects (spectra) of very high dimensions. The weight of the outliers was thus insignificant. RBA tends also to estimate high m-values (> 6) which means a lot of fuzziness in the clustering and thus overlapped estimated clusters. The m-rules (FDO, DK and OK) and the URBA algorithm estimate m-values most often in the range $[1.1; 6]$. All these results show that RBA method is not largely efficient and is exclusively adapted for specific infrared spectroscopy data such as those presented in [41].

On the contrary, URBA method seems to be efficient to estimate both the number of clusters and the fuzziness parameter. It is adapted to a large variety of datasets including infrared spectroscopy data (results not shown) and small datasets such as those studied in this article, because it is based on a fuzzy similarity measure which is

insensitive to outliers. An interesting characteristic of this algorithm is that it is not dependent, by construction, on the m range since break points are expected all along the m values tested. However, the selected m range must be large enough to ensure the apparition of these break points. If no break point appears, the user will thus only need to increase the upper bound of the m range. Considering the estimated number of clusters, the URBA algorithm is not the best method since the V_P and V_{ASI} validity indices have a better estimation percentage if $m = 2$ or if m is estimated by FDO.

A reason to explain the misclustering of URBA can be done on the Butterfly2 dataset (plotted on figure 1(a)). This dataset is composed of symmetrically distributed objects defining 2 clusters. Because the central object of this dataset is belonging for a half to cluster 1 and for a half to cluster 2, we decided to not show the assignment of the objects on figure 1(a). As can be seen on figure 1(b), the two clusters estimated by FCM applied with OS parameters (2 clusters and $m = 2$) are identical to the real clusters except the shared central point that is assigned to the first estimated cluster due to the defuzzification step. The clusters estimated by FCM with URBA parameters (3 clusters and $m = 3.5$) are represented on figure 1(c). Visually, this clustering seems logical since a new cluster is formed from the shared object which also attracts its two neighbors in this cluster. For this dataset, FCM results obtained with URBA parameters are physically interpretable.

In order to compare the computational speed, we calculated the mean computational time for 10 repetitions of validity indices and URBA method applied on the Iris dataset. All the algorithms were applied in identical conditions, *i.e.* for $C \in \{2, 3, \ldots, 13\}$ and $m \in \{1.1, 1.2, \ldots, 20\}$. The results are listed in table VIII. The URBA method has the advantage to be very fast compared to RBA and validity indices. Its computational speed is mainly due to a limited number of runs of the FCM algorithm. Validity indices require the computation of FCM for each possible couple (C, m), while URBA algorithm exploits the fact that the maximal number of non-redundant clusters decreases with m increasing. Indeed, when applied to the Iris dataset for only one repetition, the validity indices requires the computation of FCM 2280 times, while the URBA algorithm runs FCM 395 times.

V. Conclusion

The estimation of the number of clusters is an important problem of unsupervised classification. Validity indices have been developed in literature to bring a response to this problem. However, validity indices are used with a fuzziness parameter m equal to 2. Some studies have reported that this predefined choice of m can lead to mistakes. Methods named m-rules have thus been constructed in order to estimate the fuzziness parameter m in function of the dataset. However, validity indices and m-rules are developed independently one from the other. Furthermore, validity indices require the knowledge of the m-value, while the majority of m-rules needs a predefined number of clusters. C and m being linked, methods estimating both of them appeared more relevant. Three algorithm presenting these characteristics have thus been tested in this article, SJM, RBA and URBA. Compared to 29 well-known validity indices on 49 different simulated and real datasets, the URBA algorithm reveals more efficient than RBA and SJM algorithms and as performant as the best validity indices, V_P and V_{ASI}. Even for some controversially datasets such as Butterfly for which it is not able to estimate the exact number of clusters, it leads to interesting and physically

interpretable solutions. URBA proposes some interesting characteristics. First, the fuzziness parameter is estimated at the same time as the number of clusters. Second, it is independent from the selected m-range. Third, it is very fast compared to validity indices. This algorithm is efficient and should help to easily and automatically process huge and complex real life datasets such as those generated in microarray and vibrational spectroscopy for instance.

REFERENCES

[1] I. Gurrutxaga, J. Muguerza, O. Arbelaitz, J. Prez, and J. I. Martin, "Towards a standard methodology to evaluate internal cluster validity indices," *Pattern Recognition Letters*, vol. 32, no. 3, pp. 505–515, 2011.

[2] J. C. Bezdek, *Pattern recognition with fuzzy objective function algorithms*. New York: Plenum Press, 1981.

[3] J. H. Ward, "Hierarchical grouping to optimize an objective function," *Journal of the American Statistical Association*, vol. 58, no. 301, pp. 236–244, 1963.

[4] J. B. MacQueen, "Some methods for classification and analysis of multivariate observations," in *Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability*. University of California Press, 1967, pp. 281–297.

[5] W. Wang and Y. Zhang, "On fuzzy cluster validity indices," *Fuzzy Sets and Systems*, vol. 158, pp. 2095–2117, 2007.

[6] R. N. Dave, "Validating fuzzy partition obtained through c-shells clustering," *Pattern Recognition Letters*, vol. 17, pp. 613–623, 1996.

[7] M. P. Windham, "Cluster validity for fuzzy clustering algorithms," *Fuzzy Sets and Systems*, vol. 5, pp. 177–185, 1981.

[8] Y. I. Kim, D.-W. Kim, K. Lee, and K. H. Lee, "A cluster validation index for GK cluster analysis based on relative degree of sharing," *Information Sciences*, vol. 168, pp. 225–242, 2004.

[9] M. Y. Chen and D. A. Linkens, "Rule-based self-generation and simplification for data-driven fuzzy models," *Fuzzy Sets and Systems*, vol. 142, pp. 243–265, 2004.

[10] Y. Fukuyama and M. Sugeno, "A new method of choosing the number of clusters for the Fuzzy C-Means method," in *Proceedings of the 5th Fuzzy Systems Symposium*, Japan, 1989, pp. 247–250.

[11] X. L. Xie and G. Beni, "A validity measure for fuzzy clustering," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 13, no. 8, pp. 841–847, 1991.

[12] S. H. Kwon, "Cluster validity index for fuzzy clustering," *Electronics Letters*, vol. 34, no. 22, pp. 2176–2177, 1998.

[13] Y. Tang, F. Sun, and Z. Sun, "Improved validation index for fuzzy clustering," in *Proceedings of the American Control Conference*, vol. 2, Portland, OR, USA, June 2005, pp. 1120–1125.

[14] K.-L. Wu and M.-S. Yang, "A cluster validity index for fuzzy clustering," *Pattern Recognition Letters*, vol. 26, pp. 1275–1291, 2005.

[15] G. E. Tsekouras and H. Sarimveis, "A new approach for measuring the validity of the Fuzzy C-Means algorithm," *Advances in Engineering Software*, vol. 35, pp. 567–575, 2004.

[16] M. R. Rezaee, B. P. F. Lelieveldt, and J. H. C. Reiber, "A new cluster validity index for the Fuzzy C-Mean," *Pattern Recognition Letters*, vol. 19, pp. 237–246, 1998.

[17] H. Sun, S. Wang, and Q. Jiang, "FCM-based model selection algorithms for determining the number of clusters," *Pattern Recognition*, vol. 37, pp. 2027–2037, 2004.

[18] M. K. Pakhira, S. Bandyopadhyay, and U. Maulik, "Validity index for crisp and fuzzy clusters," *Pattern Recognitino*, vol. 37, pp. 487–501, 2004.

[19] M. Bouguessa, S. Wang, and H. Sun, "An objective approach to cluster validation," *Pattern Recognition Letters*, vol. 27, pp. 1419–1430, 2006.

[20] N. Zahid, M. Limouri, and A. Essaid, "A new cluster-validity for fuzzy clustering," *Pattern Recognition*, vol. 32, pp. 1089–1097, 1999.

[21] I. Gath and A. B. Geva, "Unsupervised optimal fuzzy clustering," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 11, no. 7, pp. 773–781, 1989.

[22] Y. Xie, V. V. Raghavan, P. Dhatric, and X. Zhao, "A new fuzzy clustering algorithm for optimally finding granular prototypes," *International Journal of Approximate Reasoning*, vol. 40, pp. 109–124, 2005.

[23] J. Yu, H. Huang, and S. Tian, "An efficient optimality test for the Fuzzy C-Means algorithm," in *Proceedings of the IEEE International Conference on Fuzzy Systems*, vol. 1, Honolulu, HI, USA, May 2002, pp. 98–103.

[24] D.-W. Kim, K. H. Lee, and D. Lee, "On cluster validity index for estimation of the optimal number of fuzzy clusters," *Pattern Recognition*, vol. 37, pp. 2009–2025, 2004.

[25] P. J. Rousseeuw, "Silhouettes: a graphical aid to the interpretation and validation of cluster analysis," *Journal of computational and applied mathematics*, vol. 20, pp. 53–65, 1987.

[26] R. J. G. B. Campello and E. R. Hruschka, "A fuzzy extension of the silhouette width criterion for cluster analysis," *Fuzzy sets and systems*, vol. 157, no. 21, pp. 2858–2875, 2006.

[27] M. Rawashded and A. Ralescu, "Fuzzy cluster validity with generalized silhouettes," in *Proceedings of the 23rd Annual Midwest Artificial Intelligence and Cognitive Science Conference, Cincinnati, Ohio, April*. Citeseer, 2012, pp. 11–18.

[28] S. Brodowski, *A validity criterion for fuzzy clustering*. Springer, 2011.

[29] H. L. Capitaine and C. Frlicot, "A cluster-validity index combining an overlap measure and a separation measure based on fuzzy-aggregation operators," *IEEE Transactions on Fuzzy Systems*, vol. 19, no. 3, pp. 580–588, 2011.

[30] R. Parthasarathy, G. Thiagarajan, X. Yao, Y.-P. Wang, P. Spencer, and Y. Wang, "Application of multivariate spectral analysis in micro-Raman imaging to unveil structural/chemical features of the adhesive/dentin interface," *Journal of Biomedical Optics*, vol. 13, no. 1, p. 014020, 2008.

[31] X.-Y. 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*, vol. 27, pp. 237–248, 2007.

[32] A. Marec, J.-H. Thomas, and R. E. Guerjouma, "Damage characterization of polymer-based composite materials: Multivariate analysis and wavelet transform for clustering acoustic emission data," *Mechanical Systems and Signal Processing*, vol. 22, pp. 1441–1464, 2008.

[33] W. Chen, M. L. Giger, and U. Bick, "A Fuzzy C-Means (FCM)-based approach for computerized segmentation of breast lesions in dynamic contrast-enhanced MR images," *Academic Radiology*, vol. 13, no. 1, pp. 63–72, 2006.

[34] D. Dembélé and P. Kastner, "Fuzzy C-Means method for clustering microarray data," *Bioinformatics*, vol. 19, no. 8, pp. 973–980, 2003.

[35] P. J. Deer and P. Eklund, "A study of parameter values for a Mahalanobis distance fuzzy classifier," *Fuzzy Sets and Systems*, vol. 137, no. 2, pp. 191–213, 2003.

[36] J. Yu, Q. Cheng, and H. Huang, "Analysis of the weighting exponent in the FCM," *IEEE Transactions on Systems, Man, and Cybernetics–Part B: Cybernetics*, vol. 34, no. 1, pp. 634–639, 2004.

[37] F. Okeke and A. Karnieli, "Linear mixture model approach for selecting fuzzy exponent value in Fuzzy C-Means algorithm," *Ecological Informatics*, vol. 1, no. 1, pp. 117–124, 2006.

[38] N. R. Pal and J. C. Bezdek, "On cluster validity for the Fuzzy C-Means model," *IEEE Transactions on Fuzzy Systems*, vol. 3, no. 3, pp. 370–379, 1995.

[39] H. Choe and J. B. Jordan, "On the optimal choice of parameters in a Fuzzy C-Means algorithm," in *Proceedings of the IEEE International Conference on Fuzzy Systems*, San Diego, CA , USA, March 1992, pp. 349–354.

[40] Y. Wang, X. Yao, and R. Parthasarathy, "Characterization of interfacial chemistry of adhesive/dentin bond using FTIR chemical imaging with univariate and multivariate data processing," *Journal of Biomedical Materials Research - Part A*, vol. 91, no. 1, pp. 251–262, 2009.

[41] 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*, vol. 91, no. 5, pp. 799–811, 2011.

[42] V. Schwmmle and O. N. Jensen, "A simple and fast method to determine the parameters for fuzzy cmeans cluster analysis," *Bioinformatics*, vol. 26, no. 22, pp. 2841–2848, 2010.

[43] A. Bensaid, L. Hall, J. Bezdek, L. Clarke, M. Silbiger, and J. Arrington, "Validity-guided (re)clustering with applications to image segmentation," *IEEE Transactions on Fuzzy Systems*, vol. 4, no. 2, pp. 112–123, 1996.

[44] C. Döring, M.-J. Lesot, and R. Kruse, "Data analysis with fuzzy clustering methods," *Computational Statistics & Data Analysis*, vol. 51, pp. 192–214, 2006.

[45] B. Rezaee, "A cluster validity index for fuzzy clustering," *Fuzzy sets and systems*, vol. 161, no. 23, pp. 3014–3025, 2010.

[46] J. Bezdek and N. Pal, "Some new indexes of cluster validity," *IEEE Transactions on Systems, Man, and Cybernetics–Part B: Cybernetics*, vol. 28, no. 3, pp. 301–315, 1998.

[47] A. Frank and A. Asuncion, "UCI machine learning repository," 2010. [Online]. Available: http://archive.ics.uci.edu/ml

[48] R. Fisher, "The use of multiple measurements in taxonomic problems," *Annals of Eugenics*, vol. 7, no. 2, pp. 179–188, 1936.



Teddy Happillon received the M.S. degree in informatic and the PhD degree in biophysics from the University of Reims Champagne-Ardenne (URCA), Reims, France, respectively in 2009 and 2013.

He is currently an Assistant Professor (ATER) from September 2013 to August 2014 in the Institut Universitaire et Technologie (IUT) of Châlons-en-Champagne, France. His interests include clustering, validity indices, classification, and their applications to vibrational spectroscopy.



David Sebiskveradze received the M.S. degree in physics from the University of Paris XII, France, in 2007, and the PhD degree in biophysics from the University of Reims Champagne-Ardenne (URCA), Reims, France, in 2011.

His research involves chemometrics, blind source separation and fuzzy clustering, and their applications to infrared spectroscopy.



Valeriu Vrabie received the B.S. degree in electronics and telecommunications from the University Politehnica of Bucharest (UPB), Bucharest, Romania, in 1998, the M.S. degree in signal processing from the Grenoble Institute of Technology (INP Grenoble), Grenoble, France, in 2000, and the Ph.D. degree in signal processing jointly from the INPG and UPB in 2003.

He is currently Associate Professor (Maître de Conférences) at the Centre de Recherche en STIC (CReSTIC), University of Reims, France. His research involves statistical signal processing and source separation methods with applications in geophysics, biomedical engineering, and biometry.



Michel Manfait received the Ph.D. degree in physics from the University of Reims Champagne-Ardenne (URCA), Reims, France, in 1981.

He is currently a Professor at the Extracellular Matrix and Cell Dynamics (MEDyC CNRS FRE 3481) unit where he heads the MéDIAN laboratory. His role is to coordinate research projects in biophysics. His research interests include the cellular and tissular engineering, the study of intracellular dynamics of anti-tumour drugs and of their cellular targets, and the modulation of structural and functional organisation of the nucleus induced by drugs.

Prof. Manfait is a member of AACR.



Cyril Gobinet received the M.S. and Ph.D. degrees in signal processing from the University of Reims Champagne-Ardenne (URCA), Reims, France, respectively in 2002 and 2006.

He was a post-doctoral fellow from 2006 to 2007 and an Assistant Professor (ATER) from 2007 to 2008 at the Extracellular Matrix and Cell Dynamics (MEDyC CNRS FRE 3481) unit, URCA, France. He is currently an Associate Professor (Maître de Conférences) at the MEDyC unit. His research fields include chemometrics, signal processing, source separation, clustering and classification methods with applications in vibrational spectroscopy.

APPENDIX A
MATHEMATICAL NOTATIONS

The following mathematical notations are used throughout this paper:

- $\|.\|$ denotes the Euclidean norm;
- $.^T$ refers to the transpose operator;
- $\mathrm{diag}(B)$ is the column vector composed of the diagonal elements of the matrix B;
- $u_{i.}$ represents the line vector composed of the values u_{ij}, $1 \leqslant j \leqslant K$, and $u_{.j}$ represents the column vector composed of the values u_{ij}, $1 \leqslant i \leqslant N$;
- I_P is the $P \times P$ identity matrix;
- $min(a, b)$ gives the minimum value between a and b;
- $max(a, b)$ gives the maximum value between a and b.

APPENDIX B
FIRST GROUP OF VALIDITY INDICES COMPUTED EXCLUSIVELY WITH MEMBERSHIP VALUES

Reference	Validity indices	Acronym	Formula	Criteria
[2]	Partition coefficient	PC	$V_{PC} = \frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{C} u_{ij}^2$	$C^* = \underset{2 \leqslant C \leqslant C_{max}}{\arg\max}\ V_{PC}$
[2]	Partition entropy	PE	$V_{PE} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{C} u_{ij}\log_a u_{ij}$	$C^* = \underset{2 \leqslant C \leqslant C_{max}}{\arg\min}\ V_{PE}$
[6]	Modified partition coefficient	MPC	$V_{MPC} = 1 - \frac{C}{C-1}(1 - V_{PC})$	$C^* = \underset{2 \leqslant C \leqslant C_{max}}{\arg\max}\ V_{MPC}$
[7]	Windham proportion exponent	WPE	$V_{WPE} = -\log_e\Big[\prod_{i=1}^{N}\Big[\sum_{j=1}^{[(\max_c u_{ic})^{-1}]}(-1)^{j+1}\binom{C}{j}(1 - j\max_c u_{ic})^{C-1}\Big]\Big]$	$C^* = \underset{2 \leqslant C \leqslant C_{max}}{\arg\max}\ V_{WPE}$
[8]	Kim-Kim-Lee-Lee	KYI	$V_{KYI} = \frac{2}{C(C-1)}\sum_{j=1}^{C-1}\sum_{c=j+1}^{C}\sum_{i=1}^{N}\Big(C\min(u_{ij}, u_{ic})\Big(-\sum_{l=1}^{C} u_{il}\log_a u_{il}\Big)\Big)$	$C^* = \underset{2 \leqslant C \leqslant C_{max}}{\arg\min}\ V_{KYI}$
[9]	Fuzzy partition	P	$V_P = \frac{1}{N}\sum_{i=1}^{N}\max_j u_{ij} - \frac{1}{\sum_{j=1}^{C-1} j}\sum_{j=1}^{C-1}\sum_{c=j+1}^{C}\Big[\frac{1}{N}\sum_{i=1}^{N}\min(u_{ij}, u_{ic})\Big]$	$C^* = \underset{2 \leqslant C \leqslant C_{max}}{\arg\max}\ V_P$

APPENDIX C

SECOND GROUP OF VALIDITY INDICES COMPUTED WITH BOTH DATA AND FCM RESULTS (MEMBERSHIP VALUES AND PROTOTYPES), AND BASED ON SEPARATION AND COMPACTNESS CONCEPTS

Reference	Validity indices	Acronym	Formula	Criteria
[10]	Fukuyama-Sugeno	FS	$V_{FS} = \sum_{i=1}^{N}\sum_{j=1}^{C} u_{ij}^m \|x_i - v_j\|^2 - \sum_{i=1}^{N}\sum_{j=1}^{C} u_{ij}^m \left\| v_j - \left(\frac{\sum_{c=1}^{C} v_c}{C}\right)\right\|^2$	$C^* = \underset{2\leqslant C\leqslant C_{max}}{\arg\min}\ V_{FS}$
[11]	Xie-Beni	XB	$V_{XB} = \frac{\sum_{i=1}^{N}\sum_{j=1}^{C} u_{ij}^m \|x_i - v_j\|^2}{N \min_{j,c,j\neq c} \|v_j - v_c\|^2}$	$C^* = \underset{2\leqslant C\leqslant C_{max}}{\arg\min}\ V_{XB}$
[12]	Kwon	K	$V_K = \frac{\sum_{i=1}^{N}\sum_{j=1}^{C} u_{ij}^2 \|x_i - v_j\|^2 + \frac{1}{C}\sum_{j=1}^{C}\left\|v_j - \left(\frac{\sum_{c=1}^{C} v_c}{C}\right)\right\|^2}{\min_{j,c,j\neq c} \|v_j - v_c\|^2}$	$C^* = \underset{2\leqslant C\leqslant C_{max}}{\arg\min}\ V_C$
[13]	Tang	T	$V_T = \frac{\sum_{i=1}^{N}\sum_{j=1}^{C} u_{ij}^2 \|x_i - v_j\|^2 + \frac{1}{C(C-1)}\sum_{j=1}^{C}\sum_{c=1,c\neq j}^{C} \|v_j - v_c\|^2}{\min_{j,c,j\neq c} \|v_j - v_c\|^2 + \frac{1}{C}}$	$C^* = \underset{2\leqslant C\leqslant C_{max}}{\arg\min}\ V_T$
[14]	Partition coefficient and exponential separation	PCAES	$V_{PCAES} = \sum_{i=1}^{N}\sum_{j=1}^{C} \frac{u_{ij}^2}{\max_c (\sum_{l=1}^{N} u_{lc}^2)} - \sum_{j=1}^{C} \exp\left(-C \min_{c\neq j}\left(\frac{\|v_j - v_c\|^2}{\sum_{l=1}^{C}\|v_l - \frac{\sum_{i=1}^{N} x_i}{N}\|^2}\right)\right)$	$C^* = \underset{2\leqslant C\leqslant C_{max}}{\arg\max}\ V_{PCAES}$
[15]	Validity index	VI	$V_{VI} = \frac{\sum_{j=1}^{C}\frac{\sum_{i=1}^{N} u_{ij}^m \|x_i - v_j\|^2}{\sum_{i=1}^{N} u_{ij}}}{\sum_{j=1}^{C+1}\sum_{c=1,k\neq j}^{C+1}\left(\sum_{l=1,l\neq c}^{C+1}\left(\frac{\|z_c - z_j\|}{\|z_c - z_l\|}\right)^w\right)^{\frac{w}{w+2}}\|z_c - z_j\|^2}$ with $[z_1, z_2, \ldots, z_C, z_{C+1}] = \left[v_1, v_2, \ldots, v_C, \frac{\sum_{i=1}^{N} x_i}{N}\right]$, and $w = 2.5$	$C^* = \underset{2\leqslant C\leqslant C_{max}}{\arg\min}\ V_{VI}$
[16]	Compose within and between scattering	CWB	$V_{CWB} = D(C_{max})\frac{\frac{1}{C}\sum_{j=1}^{C}\left\|\frac{1}{N}\sum_{i=1}^{N} u_{ij}\,\mathrm{diag}\left((x_i - v_j)(x_i - v_j)^T\right)\right\|}{\left\|\frac{1}{N}\sum_{i=1}^{N}\left(x_i - \frac{1}{N}\sum_{l=1}^{N} x_l\right)\right\|} + D(C)$ with $D(C) = \frac{d_{max}(C)}{d_{min}(C)}\sum_{j=1}^{C}\left(\sum_{c=1}^{C}\|v_j - v_c\|\right)^{-1}$, $d_{max}(C) = \max_{1\leqslant j,c\leqslant C, j\neq c}\|v_j - v_c\|$, $d_{min}(C) = \min_{1\leqslant j,c\leqslant C, j\neq c}\|v_j - v_c\|$	$C^* = \underset{2\leqslant C\leqslant C_{max}}{\arg\min}\ V_{CWB}$
[17]	Wang-Sun-Jiang	WSJ	$V_{WSJ} = \frac{\frac{1}{C}\sum_{j=1}^{C}\left\|\frac{1}{N}\sum_{i=1}^{N} u_{ij}\mathrm{diag}\left((x_i - v_j)(x_i - v_j)^T\right)\right\|}{\left\|\frac{1}{N}\sum_{i=1}^{N}\left(x_i - \frac{1}{N}\sum_{l=1}^{N} x_l\right)\right\|} + \frac{S(C)}{S(C_{max})}$ with $S(C) = \frac{d_{max}^2(C)}{d_{min}^2(C)}\sum_{j=1}^{C}\left(\sum_{c=1}^{C}\|v_j - v_c\|^2\right)^{-1}$ $d_{max}(C) = \max_{1\leqslant j,c\leqslant C, j\neq c}\|v_j - v_c\|$, $d_{min}(C) = \min_{1\leqslant j,c\leqslant C, j\neq c}\|v_j - v_c\|$	$C^* = \underset{2\leqslant C\leqslant C_{max}}{\arg\min}\ V_{WSJ}$
[18]	Pakhira-Bandyopadhyay-Maulik fuzzy	PBMF	$V_{PBMF} = \left(\frac{(\max_{1\leqslant j,c\leqslant C}\|v_j - v_c\|)\left(\sum_{i=1}^{N}\|x_i - \frac{1}{N}\sum_{l=1}^{N} x_l\|\right)}{C\sum_{i=1}^{N}\sum_{j=1}^{C} u_{ij}^m\|x_i - v_j\|}\right)^2$	$C^* = \underset{2\leqslant C\leqslant C_{max}}{\arg\max}\ V_{PBMF}$
[20]	Separation and total compactness	STC	$V_{STC} = \frac{\sum_{j=1}^{C}\left\|v_j - \frac{1}{N}\sum_{i=1}^{N} x_i\right\|^2}{C\sum_{j=1}^{C}\left(\frac{\sum_{i=1}^{N} u_{ij}^m\|x_i - v_j\|^2}{\sum_{i=1}^{N} u_{ij}}\right)} - \frac{\sum_{j=1}^{C-1}\sum_{c=j+1}^{C}\left(\frac{\sum_{i=1}^{N}(\min(u_{ij},u_{ic}))^2}{\sum_{i=1}^{N}(\max_{1\leqslant j\leqslant C} u_{ij})^2}\right)}{\frac{Sep(C)}{Comp(C)}}$	$C^* = \underset{2\leqslant C\leqslant C_{max}}{\arg\max}\ V_{STC}$
[45]	Fuzzy separation and compactness	FSC	$V_{FSC} = \frac{}{max_{2\leq L\leq C_{max}} Sep(L)} + \frac{}{max_{2\leq L\leq C_{max}} Comp(L)}$ Where $Sep(C) = -\frac{2}{C(C-1)}\sum_{p\neq q=1}^{C}\sum_{i=1}^{N} Cmin(u_{ip}, u_{iq})\sum_{j=1}^{C} u_{ij} log_a u_{ij}$ And $Comp(C) = \sum_{i=1}^{N}\sum_{j=1}^{C} u_{ij}^m\|x_i - v_j\|^2$	$C^* = \underset{2\leqslant C\leqslant C_{max}}{\arg\min}\ V_{FSC}$

APPENDIX D

SECOND GROUP OF VALIDITY INDICES COMPUTED THANKS TO FCM RESULTS (MEMBERSHIP VALUES AND PROTOTYPES) AND DATA, AND BASED ON OTHER CONCEPTS THAN SEPARATION AND COMPACTNESS

Reference	Validity index	Acronym	Formula	Criteria
[21]	Fuzzy hypervolume	FHV	$V_{FHV} = \sum_{j=1}^{C} \left[\det \left(\frac{\sum_{i=1}^{N} u_{ij}^{m} (x_i - v_j)(x_i - v_j)^T}{\sum_{i=1}^{N} u_{ij}^{m}} \right) \right]^{\frac{1}{2}}$	$C^* = \underset{2 \leqslant C \leqslant C_{max}}{\arg\min} \; V_{FHV}$
[21]	Partition density	PD	$V_{PD} = \frac{\sum_{j=1}^{C} \sum_{i \in I_j} u_{ij}}{V_{FHV}}$ with $I_j = \{l : (x_l - v_j)^T F_j^{-1} (x_l - v_j) < 1\}$ and $F_j = \left(\frac{\sum_{i=1}^{N} u_{ij}^{m} (x_i - v_j)(x_i - v_j)^T}{\sum_{i=1}^{N} u_{ij}^{m}} \right)$	$C^* = \underset{2 \leqslant C \leqslant C_{max}}{\arg\max} \; V_{PD}$
[21]	Average partition density	APD	$V_{APD} = \frac{1}{C} \sum_{j=1}^{C} \frac{\sum_{i \in I_j} u_{ij}}{[\det(F_j)]^{\frac{1}{2}}}$ with $I_j = \{l : (x_l - v_j)^T F_j^{-1} (x_l - v_j) < 1\}$ and $F_j = \left(\frac{\sum_{i=1}^{N} u_{ij}^{m} (x_i - v_j)(x_i - v_j)^T}{\sum_{i=1}^{N} u_{ij}^{m}} \right)$	$C^* = \underset{2 \leqslant C \leqslant C_{max}}{\arg\max} \; V_{APD}$
[19]	Separation and Compactness	SCBWS	$V_{SCBWS} = \frac{\text{trace}\left(\sum_{i=1}^{N} \sum_{j=1}^{C} u_{ij}^{m} (v_j - \frac{1}{N} \sum_{l=1}^{N} x_l)(v_j - \frac{1}{N} \sum_{l=1}^{N} x_l)^T \right)}{\sum_{j=1}^{C} \text{trace}\left(\frac{\sum_{i=1}^{N} u_{ij}^{m} (x_i - v_j)(x_i - v_j)^T}{\sum_{i=1}^{N} u_{ij}^{m}} \right)}$	$C^* = \underset{2 \leqslant C \leqslant C_{max}}{\arg\max} \; V_{SCBWS}$
[22]	Granularity-dissimilarity measure	GD	$V_{GD} = \frac{\left(\sum_{j=1}^{C} min_{1 \leqslant c \leqslant C, j \neq c} \|v_j - v_c\| \right)^2}{C \sum_{j=1}^{C} \left(\frac{\sum_{i=1}^{N} u_{ij}^{2} \|x_i - v_j\|^2}{\sum_{i=1}^{N} u_{ij}^{2}} \right)}$	$C^* = \underset{2 \leqslant C \leqslant C_{max}}{\arg\max} \; V_{GD}$
[23], [5]	Optimality test	OT	$V_{OT} = \frac{\lambda_{min}}{\lambda_{max}}$ with λ_{min} and λ_{max} the minimum and maximum eigenvalues of: $H = \begin{pmatrix} H_{11} & \cdots & H_{1C} \\ \vdots & \ddots & \vdots \\ H_{C1} & \cdots & H_{CC} \end{pmatrix}$ with $H_{jc} = A_{jc} + B_{jc}$, $A_{jc} = \frac{4m}{m-1} \sum_{i=1}^{N} u_{ij}^{\frac{m+1}{2}} u_{ic}^{\frac{m+1}{2}} \frac{(x_i - v_c)(x_i - v_j)^T}{\|x_i - v_c\| \|x_i - v_j\|}$, $B_{jc} = 2\delta_{jc} \left[\left(\sum_{i=1}^{N} u_{ij}^{m} \right) I_P - \frac{2m}{m-1} \sum_{i=1}^{N} u_{ij}^{m} \frac{(x_i - v_j)(x_i - v_j)^T}{\|x_i - v_j\|^2} \right]$ and $\delta_{jc} = \begin{cases} 1 & \text{if } j = c \\ 0 & \text{otherwise} \end{cases}$	$C^* = \underset{2 \leqslant C \leqslant C_{max}}{\arg\max} \; V_{OT}$
[29]	Overlap and Separation Index	OSI	$V_{OSI} = \frac{1}{N} \sum_{i=1}^{N} \frac{u_{i(2)}}{u_{i(1)}}$ $u_{i(1)}$ and $u_{i(2)}$ are respectively the first and second largest membership of x_i	$C^* = \underset{2 \leqslant C \leqslant C_{max}}{\arg\min} \; V_{OSI}$

APPENDIX E
SECOND GROUP OF VALIDITY INDICES COMPUTED THANKS TO FCM RESULTS (MEMBERSHIP VALUES AND PROTOTYPES) AND DATA, AND BASED ON OTHER CONCEPTS THAN SEPARATION AND COMPACTNESS

Reference	Validity index	Acronym	Formula	Criteria
[24]	Overlap to separation ratio	OS	$V_{OS} = \frac{O(C)}{\max_{2 \leqslant C \leqslant C_{max}} O(C)} \frac{\max_{2 \leqslant C \leqslant C_{max}} S(C)}{S(C)}$ with $O(C) = \frac{2}{C(C-1)} \sum_{j=1}^{C-1} \sum_{c=j+1}^{C} [\sum_{\mu} \sum_{i=1}^{N} \delta(u_{ij}, u_{ic}, \mu) \omega(u_{i.})]$, $\delta(u_{ij}, u_{ic}, \mu) = \begin{cases} 1 & \text{if } u_{ij} \geqslant \mu \text{ and } u_{ic} \geqslant \mu \\ 0 & \text{otherwise} \end{cases}$, $\omega(u_{i.}) = \begin{cases} 0.1 & \text{if } \max_{1 \leqslant l \leqslant C} u_{il} \geqslant 0.8 \\ 0.4 & \text{if } 0.7 \leqslant \max_{1 \leqslant l \leqslant C} u_{il} < 0.8 \\ 0.7 & \text{if } 0.6 \leqslant \max_{1 \leqslant l \leqslant C} u_{il} < 0.7 \\ 1 & \text{otherwise} \end{cases}$ and $S(C) = \max_{1 \leqslant i \leqslant N} (\min_{1 \leqslant j \leqslant C} (u_{ij}))$	$C^* = \arg\min_{2 \leqslant C \leqslant C_{max}} V_{OS}$
[25]	Average Silhouette Index	ASI	Let $MeanDist(x, C)$ be the mean distance between x_i and all other points belonging to C $compactness(x_k) = MeanDist(x_k, C_k), x_k \in C_k$ $separation(x_k) = min(MeanDist(x_k, C_i), x_k \in C_k, \forall C_i \neq C_k$ $silhouette(x_k) = \frac{separation(x_k) - compactness(x_k)}{max(compactness(x_k), separation(x_k))}$ $V_{ASI} = \frac{\sum_{i=1}^{n} silhouette(x_i)}{n}$	$C^* = \arg\max_{2 \leqslant C \leqslant C_{max}} V_{ASI}$
[26]	Extended Average Silhouette Index	EASI	Let u_{ij} and u_{ik} be the two highest membership values of x_i to C_j and C_k respectively $V_{EASI} = \frac{\sum_{i=1}^{n} (u_{ij} - u_{ik}).silhouette(x_i)}{\sum_{i=1}^{n} (u_{ij} - u_{ik})}$	$C^* = \arg\max_{2 \leqslant C \leqslant C_{max}} V_{EASI}$
[27]	Generalized Intra-Inter Silhouette Index	GIISI	$IntraDist_k(i,j) = (u_{ki} \wedge u_{kj}), 1 \leq k \leq C, \forall i \neq j$ $InterDist_{rs}(i,j) = (u_{ri} \wedge u_{sj}) \vee (u_{si} \wedge u_{rj}), 1 \leq r \leq s \leq C, \forall i \neq j$ $compactness(x_i) = min\left\{ \frac{\sum_{j=1}^{n} IntraDist_k(i,j).d_{ij}}{\sum_{j=1}^{n} IntraDist_k(i,j)} \middle\| 1 \leq k \leq C \right\}$ $separation(x_i) = min\left\{ \frac{\sum_{j=1}^{n} InterDist_{rs}(i,j).d_{ij}}{\sum_{j=1}^{n} InterDist_{rs}(i,j)} \middle\| 1 \leq r \leq s \leq C \right\}$ Then, V_{GIISI} is computed like V_{ASI}	$C^* = \arg\max_{2 \leqslant C \leqslant C_{max}} V_{GIISI}$
[28]	Pattern Distance Ratio Differences Ranking	PDRDR	$V_{PDR} = \frac{1}{n} \sum_{i=1}^{n} \frac{\sum_{j=1}^{C} u_{ij}.d(x_i, C_j)}{\sum_{j=1}^{C} (1 - u_{ij}).d(x_i, C_j)}.(C-1)$ $diff_C = V_{PDR}(C) - V_{PDR}(C-1), for C = 3, ..., C_{max} + 1$ Let r_C be the rank of $diff_C$ after sorting it in increase order $V_{PDRDR} = r_C - r_{C+1} for C = 3, ..., C_{max}$	$C^* = \arg\min_{2 \leqslant C \leqslant C_{max}} V_{PDRDR}$

Reference	m-rule	Acronym	Formula	Criteria
[39]	Fuzzy decision optimization	FDO	$\mu_D(m) = \min(\mu_{G_1}(m), \ldots, \mu_{G_C}(m), \mu_{K_1}(m), \ldots, \mu_{K_C}(m))$ with $\mu_{G_i}(m) = 1 - \exp(-A.nC_i)$ $With\mu_{K_i}(m) = \frac{1}{1 + \left(\frac{\sum_{j=1}^{nC_i} \|x_j - v_i\|^2}{C} B\right)}$ With A, B and C constants, nC_i is the cardinality of the i^{th} cluster	
[34]	Dembélé-Kastner	DK	$m_{ub} = \arg\max_m \left\| \frac{\sigma_{Y_m}}{\overline{Y}_m} - 0.03P \right\|$ with $y(m) = \left\{ y_k(m); k = \left\{1, 2, \ldots, \frac{N(N-1)}{2}\right\} \right\} \in \mathbb{R}^{\frac{N(N-1)}{2}}$, $y_k(m) = \left((x_i - x_j)^T (x_i - x_j)\right)^{\frac{1}{m-1}}, j \neq i = \{1, 2, \ldots, N\}$, $\overline{Y}_m = \frac{2}{N(N-1)} \sum_{k=1}^{\frac{N(N-1)}{2}} y_k(m)$, $\sigma_{Y_m} = \sqrt{\frac{1}{N-1} \sum_{k=1}^{\frac{N(N-1)}{2}} \left(y_k(m) - \overline{Y}_m\right)^2}$	$m^* = 1 + \min\left(1, \frac{m_{ub}}{10}\right)$
[36]	α-rule	αR	$m_{ub} = \begin{cases} \frac{\min(P, N-1)}{\min(P, N-1) - 2} & \text{if } \min(P, N-1) \geqslant 3 \\ +\infty & \text{otherwise} \end{cases}$	$m^* \leqslant m_{ub}$
[36]	β-rule	βR	$m_{ub} = \begin{cases} \frac{1}{1 - 2\lambda_{max}(F_U)} & \text{if } \lambda_{max}(F_U) < 0.5 \\ +\infty & \text{otherwise} \end{cases}$, with $\lambda_{max}(F_U)$ the maximum eigenvalue of F_U, $F_U = \frac{H^T H}{N}$, $H = \left[\frac{(x_1 - \overline{x})}{\|x_1 - \overline{x}\|}, \frac{(x_2 - \overline{x})}{\|x_2 - \overline{x}\|}, \ldots, \frac{(x_N - \overline{x})}{\|x_N - \overline{x}\|}\right]$	$m^* \leqslant m_{ub}$
[37]	Okeke-Karnieli	OK	$R_\epsilon(m) = \|X - VU^T\|$	$m^* = \arg\min_m R_\epsilon(m)$

Article #IV

Publication n°4

La biophotonique au service de l'identification de marqueurs pronostiques intracellulaires

J. Klossa, S. Daliphard, X. Troussard, P. Vielh, M. Manfait, J. Angulo, G. Flandrin, A. Beljebbar, V. Untereiner, **T. Happillon**, C. Gobinet, S. Velasco-Forero, S. Roux, V. Saada, R. Dagiral, V. Coomans, E. Froigneux, P. Rideau, A. Vievard.

IRBM **2011**, *32*, 72-75.

Présentation de la publication n°4

Contexte

Les approches diagnostiques et pronostiques de la LLC, principalement morphologiques et moléculaires, sont complexes. Afin de répondre à une problématique de pénurie en personnel, d'automatisation et de standardisation des analyses, le projet IHMO (Imagerie Hybride en Microscopie Oncologique) à été mis en place. Ce projet vise à développer une solution d'imagerie hybride combinant la microspectroscopie Raman et la microscopie multispectrale afin d'apporter une aide au diagnostic de la LLC en différenciant les états cancéreux et non cancéreux des cellules analysées, *i.e.* lymphocytes B tumoraux et sains respectivement.

Objectif

L'objectif de cette étude est dans un premier temps de décrire les solutions techniques retenues pour construire le microscope multimodale IHMO, issu du projet éponyme. Dans un second temps, le but est de montrer par une étude préliminaire que le module Raman est capable de réaliser un tri des éléments nucléés et d'établir un diagnostic de la LLC en s'appuyant sur des outils de traitement et de classification supervisée des données. Les résultats obtenus avec le module de microscopie multispectrale, non encore aboutis, n'ont pas été présentés dans ce papier.

Matériels et Méthodes

Les échantillons sanguins considérés dans cette étude se présentent sous forme de frottis sanguins non colorés et non fixés. La détection et la localisation des globules blancs sur le frottis sont réalisées de façon automatique par microscopie conventionnelle (ECLIPSE FN1 Nikon SA, Champigny Sur Marne, France) avec un objectif à sec 40x (NA 0,6, Nikon SA, Champigny Sur Marne, France). Les frottis de 24 patients atteints d'une LLC hypercytaire et 11 volontaires ont été considérés et 90 cellules en moyenne ont été analysées par microspectroscopie Raman pour chaque prélèvement. Les spectres sont acquis au centre du noyau des globules blancs précédemment repérés avec un temps d'accumulation de 2 x 10 secondes, *via* un objectif 100x (NA 0,9, Nikon SA, Champigny Sur Marne, France) et un module Raman (Horiba Jobin Yvon, Villeneuve d'Ascq, France). Ce module est composé

d'une source laser à 532 nm et d'un spectromètre associant un détecteur CCD de 1024 x 256 pixels (Andor Technology, South Windsor, CT, USA), refroidi à -70°C par effet Pelletier. Les spectres Raman sont ensuite prétraités avant d'être classifiés par l'algorithme Support Vector Machine. Après analyse Raman, les frottis sont colorés par MGG afin d'établir la catégorie cellulaire de chaque élément analysé (lymphocyte, monocyte ou polynucléaire) et ainsi obtenir une référence permettant de valider les résultats obtenus après classification.

Résultats

Les classifications de spectres Raman, menées dans le cadre d'une distinction du type cellulaire des leucocytes analysés, ont montré un excellent pouvoir discriminant en présentant des précisions de 99,6% pour les lymphocytes et de 98,8% pour les polynucléaires et monocytes. Dans un second temps, la discrimination entre lymphocytes pathologiques et sains, a elle aussi montré un réel potentiel de reconnaissance des spectres avec une sensibilité de 95% et une spécificité de 87,5%.

Conclusion

Le projet IHMO a permis de mettre au point une machine multimodale, combinant différentes technologies de pointe et permettant l'acquisition de données de différentes natures (dont la microspectroscopie Raman) sur des frottis sanguins non colorés et dont la préparation demeure simple, rapide et peu couteuse. Des outils de prétraitements et de classification des spectres ont également été développés et intégrés au module, permettant ainsi à l'utilisateur de mener une étude complète depuis l'acquisition jusqu'au résultat de prédiction de spectres qu'il enregistre.

Cette approche diagnostic du projet a été réalisé sur la LLC mais peut également être appliquée à différents domaines de cytohématologie, comme d'autres catégories de leucémies, l'étude des hématies, ou encore l'étude de parasites sanguins.

Provided for non-commercial research and education use.
Not for reproduction, distribution or commercial use.



This article appeared in a journal published by Elsevier. The attached copy is furnished to the author for internal non-commercial research and education use, including for instruction at the authors institution and sharing with colleagues.

Other uses, including reproduction and distribution, or selling or licensing copies, or posting to personal, institutional or third party websites are prohibited.

In most cases authors are permitted to post their version of the article (e.g. in Word or Tex form) to their personal website or institutional repository. Authors requiring further information regarding Elsevier's archiving and manuscript policies are encouraged to visit:

http://www.elsevier.com/copyright

Author's personal copy

Disponible en ligne sur ScienceDirect www.sciencedirect.com

Elsevier Masson France EM|consulte www.em-consulte.com

IRBM

IRBM 32 (2011) 72–75

Article original

La biophotonique au service de l'identification de marqueurs pronostiques intracellulaires

Using biophotonics techniques to retrieve prognostic intracellular signatures

J. Klossa [a,*], S. Daliphard [b], X. Troussard [c], P. Vielh [d], M. Manfait [e], J. Angulo [f], G. Flandrin [a], A. Beljebbar [e], V. Untereiner [e], T. Happillon [e], C. Gobinet [e], S. Velasco-Forero [f], S. Roux [a], V. Saada [d], R. Dagiral [g], V. Coomans [g], E. Froigneux [h], P. Rideau [i], A. Vievard [a]

[a] *TRIBVN, 39, rue Louveau, 92320 Châtillon, France*
[b] *CHU de Reims, 51100 Reims, France*
[c] *CHU de Caen, 14033 Caen, France*
[d] *IGR-Institut Gustave-Roussy, 94805 Villejuif, France*
[e] *MEDyC CNRS URCA n° 6237, 51096 Reims, France*
[f] *CMM-ARMINES, 77300 Fontainebleau, France*
[g] *Réactifs RAL, 33651 Martillac, France*
[h] *HORIBA Jobin-Yvon, 59650 Villeneuve D'ascq, France*
[i] *NIKON, 94500 Champigny-sur-Marne, France*

Résumé

Le projet vise à développer une solution d'imagerie hybride combinant en un même instrument microspectroscopie Raman et imagerie multispectrale pour le diagnostic et l'établissement des facteurs pronostiques d'une tumeur. Les lymphocytes de 24 patients présentant une leucémie lymphoïde chronique hyperleucocytaire et ceux de 11 individus sains ont été étudiés avec en moyenne 90 cellules par prélèvement. Le repérage s'effectue en microscopie optique et les noyaux sont analysés par microspectroscopie Raman à partir de frottis sanguins étalés sur lame de verre. Finalement, les cellules correspondantes sont imagées sous forme de piles multi-Z pour huit bandes spectrales réparties dans le visible. Une classification des données Raman par un algorithme de classification supervisé (séparateur à vastes marges) a permis d'identifier une signature moléculaire permettant de classer les lymphocytes et les autres éléments nucléés du sang avec une sensibilité de 99,6 % et une spécificité de 98,8 %. Cet algorithme a été utilisé pour mettre au point un modèle permettant de distinguer individus sains et leucémiques et caractérisé par une sensibilité de 95 % pour une spécificité de 87,5 % sur les spectres constituant la base de tests (1540 spectres issus de 11 patients atteints d'une leucémie lymphoïde chronique et 516 spectres issus de sept individus sains). Le projet IHMO a montré le potentiel de la spectroscopie Raman pour réaliser une classification cellulaire. Les descripteurs morphologiques extraits de l'imagerie produisent une deuxième classification qui reste à évaluer. La solution pourra être étendue en cytohématologie, mais nécessitera des développements complémentaires pour son application en anatomo-cytopathologie.
© 2011 Elsevier Masson SAS. Tous droits réservés.

Mots clés : Leucémie lymphoïde chronique (LLC) ; Plate-forme d'imagerie microscopique multimodale ; Microspectroscopie Raman ; Signature moléculaire ; Imagerie multispectrale

Abstract

IHMO project aims at developing a multimodal microscopy platform that includes in a single machine Raman micro spectroscopy and multispectral imaging used for tumor diagnosis and prognosis. Lymphocytes from 24 leukaemic patients suffering from hyper leukocytosis Chronic Lymphoid Leukemia and from 11 healthy individual have been studied, using around 90 cells per blood sample. Blood smears were prepared on microscopy slides; cells were localized by optical microscopy, and Raman microspectroscopy spectra were acquired on cell nuclei. Afterward

* Auteur correspondant.
Adresse e-mail : jklossa@tribvn.com (J. Klossa).

1959-0318/$ – see front matter © 2011 Elsevier Masson SAS. Tous droits réservés.
doi:10.1016/j.irbm.2011.01.039

Author's personal copy

multi-Z stacks images of each cell were acquired for eight bands distributed in the visible spectrum. Raman data's were classified using a Support Vector Machine algorithm that provided a molecular signature that allowed for distinguishing lymphocytes from other nucleated blood components with 99.6% sensibility and 98.8% specificity. Then the algorithm was used for developing a classification model splitting leukaemic and healthy smears; sensibility was 95% and specificity was 87.5% among the spectra used for evaluation: 1540 from 11 leukaemic patients and 516 from seven healthy individual. IHMO project has demonstrated the power of Raman microspectroscopy for cell classification. Morphological descriptors obtained from multi-Z and multispectral images provide another independent classification that still needs to be assessed. The microscopy platform can be used more generally in the field of cytohaematology, however application to cytological and histological pathology would need further developments.
© 2011 Elsevier Masson SAS. All rights reserved.

Keywords: Chronic Lymphoid Leukemia (CLL); Microscopy multimodal platform; Raman micro spectroscopy; Molecular signature; Multispectral imaging

1. Introduction

Les approches diagnostiques et pronostiques des cancers sont principalement morphologiques et moléculaires ; elles conduisent à l'orientation thérapeutique personnalisée des patients. Ces approches multidisciplinaires restent complexes [1,2] et gagneraient à être standardisées et simplifiées. Le projet vise à développer une solution d'imagerie hybride combinant en un même instrument microspectroscopie Raman (MSR) et microscopie multispectrale [3,4] dans le but de compléter, voire à long terme remplacer, l'analyse multidisciplinaire classique utilisée jusqu'à présent pour le diagnostic et l'établissement des facteurs pronostiques d'une tumeur sur un patient donné.La MSR confocale est une technique rapide, non invasive et non destructive d'analyse de cellules ou de tissus [5]. Très sensible aux changements moléculaires et associée à des méthodes statistiques multivariées, cette technique s'impose comme un outil de diagnostic puissant pour la caractérisation des pathologies tumorales, particulièrement à l'échelle cellulaire [6,7]. L'imagerie multispectrale (IMS) dans le visible vise à confirmer les classifications de données de MSR et à identifier des caractéristiques morphologiques fines associées [8].

2. Patients et méthode

Les lymphocytes préalablement repérés en microscopie optique (au 40×) ont été analysés par MSR à partir de frottis sanguins réalisés sur lame de verre. Les lymphocytes de 24 patients présentant une LLC hyperleucocytaire et ceux de 11 individus sains ont été étudiés avec en moyenne 90 cellules par prélèvement.

Le microscope IHMO (Fig. 1) est composé d'un châssis de microscopie Nikon FN1 équipé de deux objectifs NIKON à sec : 40× pour le repérage des cellules et 100× pour l'acquisition de données MSR et d'IMS. Le module Raman (Horiba, France) couplé au microscope utilise une source laser à 532 nm (puissance à l'échantillon 8–10 mW) et un spectromètre associant un réseau de dispersion 1200 traits par milimètre et un détecteur CCD de 1024 × 256 pixels (Technologies Andor, South Windsor, CT, États-Unis) refroidi par effet Pelletier. Chaque spectre Raman est enregistré avec un temps d'accumulation de 2 × 10 secondes, dans la gamme spectrale 700–3170 cm^{-1}, avec une résolution spectrale de 4 cm^{-1}. L'IMS est acquise avec une source à diode blanche Schott filtrée selon huit bandes d'environ 50 nm réparties entre 440 et 650 nm. Le Z est piloté avec un actuateur piézo dont la répétitivité est de 30 nm. La configuration est exploitée avec la plate-forme translationnelle TRIBVN ICS FW qui pilote l'ensemble des axes motorisés du système : X-Y-Z et Roue à filtres, ainsi que le module de MSR. La plate-forme IHMO (Fig. 2) permet la production de grands champs d'images avec une caméra, le repérage et l'annotation manuelle ou automatique dans l'espace XY de la région de la lame étudiée, l'acquisition, le traitement ainsi que la classification de données Raman de même que l'acquisition à différents Z de données d'IMS dans les huit bandes spectrales.

2.1. Acquisition et traitement des données de MSR

Les données Raman sont acquises au centre du noyau des cellules avec une résolution spatiale de l'ordre du micron en XY et de deux microns en Z, chaque spectre Raman enregistré est prétraité par soustraction du signal du verre, correction de la ligne de base, lissage de Savitzky-Golay et normalisation SNV (Standard Normal Variate).

2.2. Acquisition et traitement des données d'IMS

L'IMS est acquise sur les cellules repérées dans ICS FW. Pour chaque cellule, on acquiert une pile en Z de 20 images

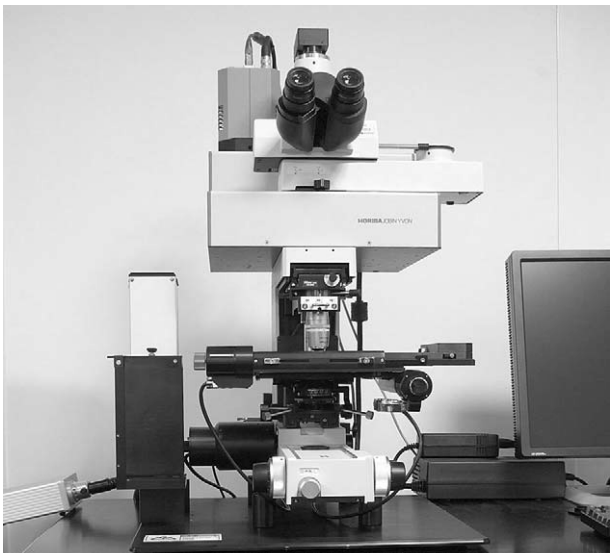


Fig. 1. Plate-forme d'imagerie multimodale IHMO.

Author's personal copy

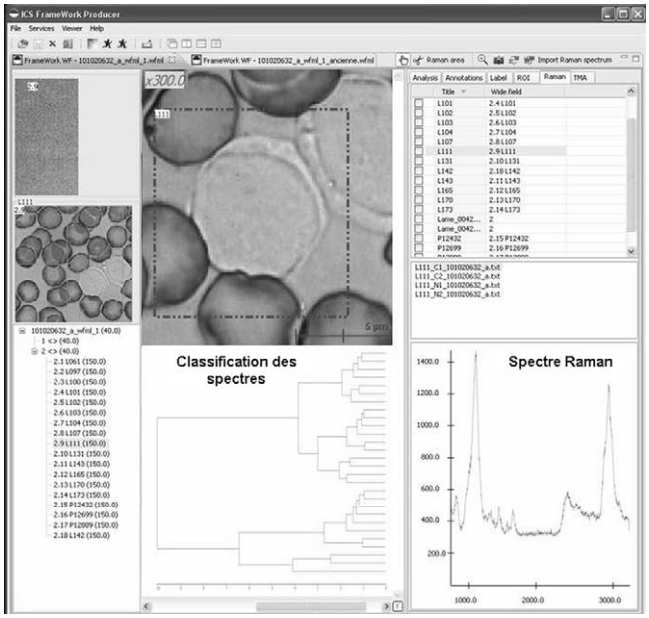


Fig. 2. Présentation à l'écran des données morphologique et spectroscopiques.

monochromes espacées de 100 nm, pour chacune des huit bandes spectrales.

Un algorithme combinant techniques de morphologie mathématique et de régression parcimonieuse a été développé pour produire une image couleur (RVB) avec une profondeur de champ étendue à partir des huit séries d'images multi-Z/multispectrale. Cette image couleur est exploitée pour confirmer visuellement la classe des cellules caractérisées par MSR et IMS [9].

L'ensemble des images scalaires multi-Z/multispectrale de chaque cellule est ensuite traité par des techniques de morphologie mathématique pour produire des descripteurs morphologiques des cellules utilisés pour produire une classification non supervisée [3,9].

3. Resultats et discussion

3.1. Classification des spectres des éléments nucléés

La comparaison entre les spectres issus de lymphocytes, de polynucléaires et de monocytes, montre des différences dans les intensités de certaines bandes spectrales. Une analyse statistique multivariée de ces spectres a permis de déterminer des marqueurs spectroscopiques spécifiques à ces 3 types cellulaires (lymphocytes, polynucléaires et monocytes) (Fig. 3). Une classification des données par un algorithme de classification supervisé (séparateur à vastes marges [SVM]), tout en considérant ces marqueurs, a permis de mettre au point un modèle caractérisé par une sensibilité de 99,6 % et une spécificité de 98,8 % dans la reconnaissance des lymphocytes. Cette étude a montré le potentiel de la spectroscopie Raman pour réaliser, sans coloration, une classification cellulaire. Pour les études ultérieures, cette classification permettra d'effectuer sur la lame un prétricellulaire afin de localiser les seuls lymphocytes et ainsi

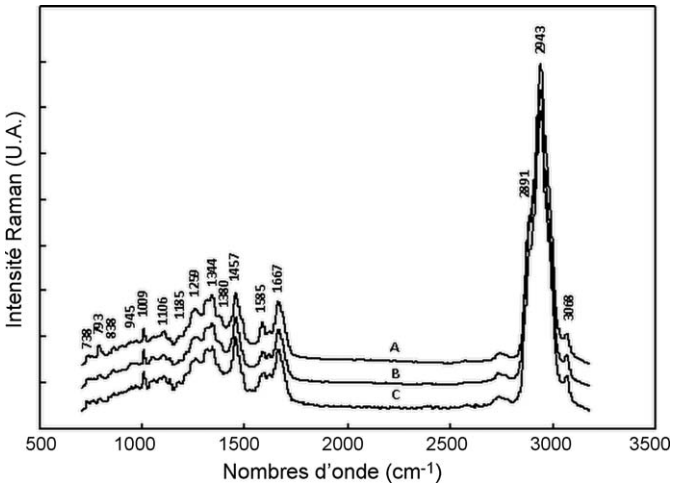


Fig. 3. Spectres de Microspectroscopie Raman d'éléments nucléés d'un frottis sanguin : spectres moyens enregistrés sur 100 lymphocytes (A), 100 polynucléaires (B) et dix monocytes (C). Les nombres d'onde des marqueurs spectroscopiques spécifiques à ces trois types cellulaires (lymphocytes, polynucléaires et monocytes) sont indiqués sur les courbes.

éviter la reconnaissance des cellules à posteriori sur l'échantillon coloré.

3.2. Distinction entre lymphocytes normaux et cellules lymphoïdes de leucémie lymphoïde chronique (LLC)

Dans notre étude, et à partir de cellules isolées d'individus, nous avons montré que les spectres Raman permettaient de distinguer des lymphocytes normaux et tumoraux (LLC) en déterminant des marqueurs intrinsèques Raman discriminant entre les états sain et pathologique à partir d'une méthode non supervisée. Ces résultats sont en accord avec des études antérieures sur des lignées cellulaires [10]. Les variations spectrales sont essentiellement associées aux acides nucléiques et aux protéines. L'algorithme SVM a donc été utilisé pour confirmer la pertinence de ces marqueurs en mettant au point un modèle permettant d'identifier des individus sains et leucémiques, avec une sensibilité de 95 % et une spécificité de 87,5 % sur les spectres constituant la base de tests (1540 spectres issus de 11 patients atteints d'une LLC et 516 spectres issus de sept individus sains).

Les descripteurs morphologiques extraits de l'imagerie produite sur les cellules ayant fourni les descripteurs Raman produisent une deuxième classification qui reste à évaluer.

4. Conclusion

Le projet a permis la constitution d'un instrument original d'acquisition de données d'IMS et de MSR, d'une technique de préparation et de repérage facilitant l'acquisition de données interprétables, d'un outillage spécifique de traitement de données de MSR et de traitement de données multispectrales multidimensionnelles. Enfin, la preuve de concept sous la forme d'une corrélation entre données morphologiques et spectroscopiques, et données cliniques relatives aux LLC est actuellement en cours.

Author's personal copy

L'approche IHMO pourra être appliquée en cytohématologie par exemple à d'autres leucémies, à l'étude des hématies et de leurs inclusions et parasites. Les applications en cyto-anatomopathologie nécessiteront des développements complémentaires.

Ce travail a bénéficié du soutien financier de l'Agence nationale de la recherche (ANR) dans le cadre de l'appel à projets TECSAN 2007.

Conflit d'intérêt

Aucun.

Références

[1] Cramer P, Hallek M. Prognostic factors in chronic lymphocytic leukemia-what do we need to know? Nat Rev Clin Oncol 2011;8(1):38–47.

[2] Zenz T, Mertens D, Küppers R, Döhner H, Stilgenbauer S. From pathogenesis to treatment of chronic lymphocytic leukaemia. Nat Rev Cancer 2010;10(1):37–50.

[3] Velasco-Forero S, Angulo J. Parameters selection of morphological scale-space decomposition for hyperspectral images using tensor modeling. In: Proc. of SPIE symposium on Defense, Security, and Sensing: Algorithms and Technologies for Multispectral, Hyperspectral, and Ultraspectral Imagery XVI, SPIE, Vol. 7695. 2010, 12 p.

[4] Malik Z, Rothmann C, Cycowitz T, Cycowitz ZJ, Cohen AM. Spectral morphometric characterization of B-CLL cells versus normal small lymphocytes. J Histochem Cytochem 1998;46(10):1113–8.

[5] Anthony T. Raman spectroscopy in biology. New York: John Wiley & Sons; 1982.

[6] Draux F, Jeannesson P, Beljebbar A, Tfayli A, Fourre N, Manfait M, Sulé-Suso J, Sockalingum GD. Raman spectral imaging of single living cancer cells: a preliminary study. Analyst 2009;134:542–8.

[7] Krishna CM, Sockalingum GD, Vadhiraja BM, Maheedhar K, Rao AC, Rao L, Venteo L, Pluot M, Fernandes DJ, Vidyasagar MS, Kartha VB, Manfait M. Vibrational spectroscopy studies of formalin-fixed cervix tissues. Biopolymers 2007;85:214–21.

[8] Angulo J, Klossa J, Flandrin G. Ontology-based lymphocyte population description using mathematical morphology on colour blood images. Cell Mol Biol 2006;52(6):2–15.

[9] Katzilakis N, Stiakaki E, Papadakis A, Dimitriou H, Stathopoulos E, Markaki EA, Balas C, Kalmanti M. Spectral characteristics of acute lymphoblastic leukemia in childhood. Leukemia Res 2004;28:1159–64.

[10] Chan JW, Taylor DS, Lane SM, Zwerdling T, Tuscano J, Huser T. Nondestructive Identification of Individual Leukemia Cells by Laser Trapping Raman Spectroscopy. Anal Chem 2008;80:2180–7.

Article #V et #VI

Publication n°5

Label free technologies: Raman micro-spectroscopy and multi-spectral imaging for lymphocyte classification

T. Happillon, V. Untereiner, A. Beljebbar, C. Gobinet, M. Manfait, S. Daliphard, P. Cornillet-Lefebvre, X. Troussard, J. Angulo, S. Velasco-Forero, V. Saada, G. Flandrin, J. Klossa.

Diagnostic Pathology 2013, 8 (Suppl 1) : S32 doi : 10.1186/1746-1596-8-S1-S32

Publication n° 6

Diagnosis approach of Chronic Lymphocytic Leukemia on unstained blood smears using Raman microspectroscopy and supervised classification

T. Happillon, V. Untereiner, A. Beljebbar, C. Gobinet, S. Daliphard, P. Cornillet-Lefebvre, A. Quinquenel, A. Delmer, X. Troussard, J. Klossa, M. Manfait.

Article à soumettre à Analyst, 2013

Présentation des publications n°5 et 6

Contexte

La Leucémie Lymphoïde Chronique, et plus précisément la LLC de type B, est une maladie cancéreuse du sang touchant les lymphocytes B. Elle est la plus répandue des leucémies en Occident et touche essentiellement les personnes âgées de plus de 50 ans. Ne présentant aucun symptôme, son diagnostic est très souvent réalisé de façon fortuite lors d'une simple analyse sanguine où un taux anormalement élevé de globules blancs laissera suspecter la maladie. Après différentes analyses morphologiques et moléculaires, le diagnostic peut être posé, le stade d'évolution de la maladie déterminé, et si besoin un traitement est administré. La microspectroscopie Raman quant à elle est un outil puissant, non destructif et non invasif, permettant d'enregistrer des informations moléculaires sur les échantillons analysés. Son utilisation dans une optique d'approche diagnostique de la LLC apparaît donc pertinente.

Objectif

L'objectif de cette étude est dans un premier temps de décrire précisément les procédures d'acquisition et de traitement numérique, des spectres acquis avec le module Raman du microscope multimodale IHMO.

Dans un second temps, l'objectif de ce travail est d'approfondir les résultats préliminaires obtenus dans le cadre du projet IHMO (publication n°4) et de confirmer le potentiel de la microspectroscopie Raman dans le diagnostic de la LLC sur une plus grande cohorte de patients.

Matériels et Méthodes

Les frottis sanguins considérés sont préparés dans les mêmes conditions que pour l'étude précédente (publication n°4) à la différence qu'ils proviennent d'un ensemble d'échantillons plus conséquent, à savoir 49 patient LLC et 27 volontaires sains. En termes de quantité de spectres, 1127 lymphocytes et 997 polynucléaires sont considérés pour la discrimination du type cellulaire, et 4257 lymphocytes tumoraux et 2628 lymphocytes sains sont utilisés pour la distinction Sain vs LLC. Ces spectres suivent les mêmes protocoles d'acquisition et de prétraitement que précédemment. Une sélection de variables assistée par classification

supervisée a été appliquée aux spectres afin d'estimer les régions spectrales d'intérêt (ROI pour Region Of Interest) la plus discriminante entre les individus sains et patients leucémiques, permettant ainsi d'améliorer les résultats de classification. Comparativement à la publication n°4, les classifications supervisées sont complétées par une étude basée sur la technique du 10-Fold Cross Validation, permettant de déterminer les capacités de généralisation en prédiction d'un algorithme de classification (ici SVM), lors de son application à un jeu de données particulier. Dans un deuxième temps, l'augmentation de la base de données pour chaque catégorie de spectres autorise une procédure de classification en trois temps, qui a été réalisée en prenant en compte des jeux d'entrainement, de validation et de prédiction.

Résultats

Les premières classifications, cherchant à distinguer les spectres issus de lymphocytes et de polynucléaires se sont révélées quasi parfaites, que ce soit avec la technique du 10-Fold Cross Validation, dont les résultats atteignent 98,9% et 96,9% de reconnaissance pour chaque catégorie cellulaire (lymphocytes et polynucléaires respectivement), ou pour la procédure de classification en trois temps, avec des résultats obtenus sur le jeu d'entrainement de 98,8% et 96,7%, sur le jeu de validation de 100% et 99,7%, et sur le jeu de prédiction de 98,5% et 99,1%. En ce qui concerne les classifications de spectres issus de lymphocytes sains et tumoraux, les valeurs de sensibilité et spécificité obtenues avec la technique du 10-Fold Cross Validation sont respectivement de 97,4% et 88,2%. Celles obtenues par la suite en couplant la procédure en trois temps avec la ROI estimée par sélection de variables sont de 83,3% et 100% sur le jeu d'entrainement, 90,9% et 84,6% sur le jeu de validation, et 80% et 100% sur le jeu de prédiction.

Conclusion

Lors de cette étude, les classifications réalisées dans le cadre d'une approche diagnostique de la LLC, ont montré le fort pouvoir discriminant de la microspectroscopie Raman dans la différenciation de spectres issus de cellules de différentes natures, mais également dans la distinction de spectres en fonction de l'état pathologique ou non des cellules dont ils sont issus.

DIAGNOSTIC PATHOLOGY

PROCEEDINGS Open Access

Label free technologies: Raman micro-spectroscopy and multi-spectral imaging for lymphocyte classification

Teddy Happillon[1†], Valérie Untereiner[1†], Abdelilah Beljebbar[1†], Cyril Gobinet[1†], Michel Manfait[1†], Sylvie Daliphard[2†], Pascale Cornillet-Lefebvre[2†], Xavier Troussard[3†], Jesus Angulo[4†], Santiago Velasco-Forero[4†], Véronique Saada[5†], Georges Flandrin[6†], Jacques Klossa[6*]

From 11th European Congress on Telepathology and 5th International Congress on Virtual Microscopy
Venice, Italy. 6-9 June 2012

Background

Current diagnostic and prognostic approaches in oncology use morphological and molecular techniques which lead to patient personalized therapies. However, they are still complex and hard to standardize. This is also true for Chronic Lymphocytic Leukemia (CLL) that has been chosen for the IHMO project [1,2]. Simplifying diagnostic processes and making easier the standardization would be highly suitable. In order to develop such a simpler automated method, the IHMO project, funded by the French National Research Agency, proposed to develop a multimodal microscopy scanning platform that includes in a single machine a Raman micro-spectrometer (RMS) combined with a multispectral imager (MSI) [3,4].

RMS is a quick non-invasive and non-destructive technique for tissues and cells analysis [5]. It is very sensible to molecular changes and it could be used as a powerful diagnostic and prognostic tool when used in association with multivariate statistical methods. It is particularly useful for characterizing pathological tumors especially at the cellular level [6,7]. Multispectral imaging in the visible spectrum could confirm RMS classifications and provide new morphometric findings [8].

CLL disease is characterized by the proliferation of lymphocytes (lymphocytosis). It is the most common leukemia, preferentially affecting people aged over 50 years old. It is incurable and in most cases shows no clinical signs. Thus, it is often discovered by chance during a blood test. If necessary, morphological and immunological studies are led by analyzing blood smears colored with May-Grünwald Giemsa, by making a complete blood count and by computing a Matutes score. Such studies are necessary because it is impossible to distinguish a healthy cell from a cancerous one only using a conventional microscope.

* Correspondence: jklossa@tribvn.com
† Contributed equally
[6]TRIBVN, 39, rue Louveau, 92320 Châtillon, France
Full list of author information is available at the end of the article

Material and methods

Blood smears were prepared on classical glass slides commonly used in laboratories for microscopy; cells were localized by optical microscopy. A new multimodal machine which has been developed combining i) a 10 bands multispectral imager using the full spectrum of transmitted visible light ii) a Raman micro-spectrometer, equipped with a 532nm diode laser excitation source delivering 13.5mW of power on the sample; iii) a microscope stand with 40x and 150x lenses suited with an xyz motorized stage for scanning the blood smear, and localizing x-y coordinates of a representative series (~100 for each patient) of lymphocyte cells before registering Raman spectra on their nuclei and their individual multi-spectral images. An additional piezo actuator allowed for precise z stack recording. Figure 1 shows each step from the screening of the smears to the final results.

Raman micro-spectroscopy

More precisely, 997 polymorphonuclears, 95 monocytes and 1127 lymphocytes from 12 different blood samples have been considered in the first part of this study, and a Raman spectrum has been acquired on each of them. In the second part, 4257 spectra have been registered on lymphocytes of 49 leukemic patients suffering from hyper leukocytosis Chronic Lymphocytic Leukemia, and 2596 spectra have been recorded on lymphocytes of 27 healthy

BioMed Central

© 2013 Happillon et al; licensee BioMed Central Ltd. This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/2.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

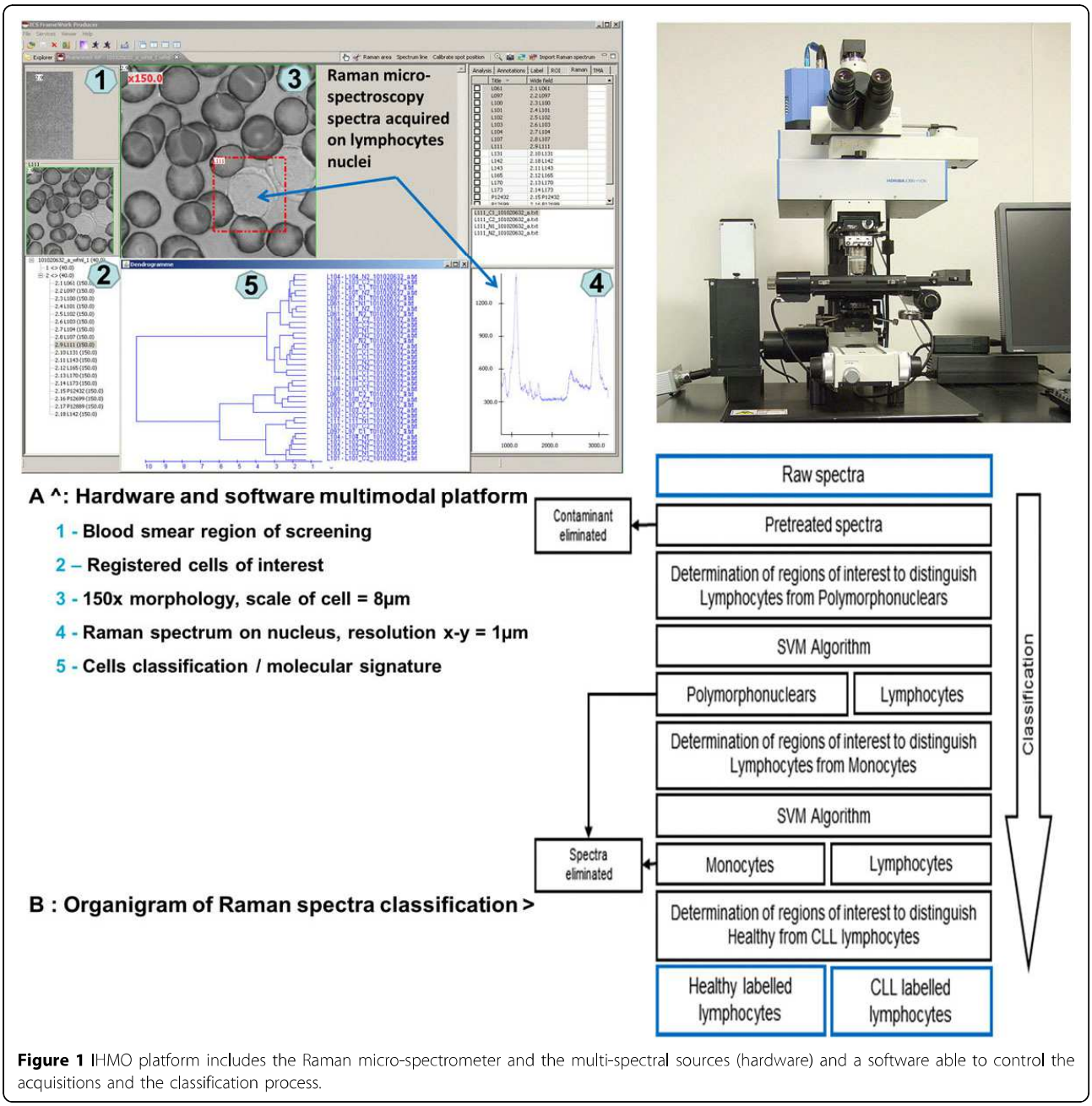


Figure 1 IHMO platform includes the Raman micro-spectrometer and the multi-spectral sources (hardware) and a software able to control the acquisitions and the classification process.

individuals. These spectra have been then studied, using around 90 cells per blood sample.

Raman data were first pretreated to erase contaminant information into the spectra and then were analyzed using a multivariate statistical method which put forward the relevant information needed to distinguish in a first time lymphocytes from polymorphonuclears. The spectra are then reduced to their relevant information and classified using a Support Vector Machine algorithm [9]. Then this algorithm was used to develop a classification model splitting leukemic and healthy lymphocytes; 3 sets of data were created, the first one, the training set, composed of the spectra from 6 leukemic patients (513 spectra) and 4 healthy individuals (315 spectra) which was used to create different prediction models, the second set, the validation set, composed of spectra from 33 leukemic patients (2820 spectra) and 13 healthy individuals (1106 spectra) was used to select the best model among all previously computed. Finally, the third set, the test set, composed of spectra from 20 different blood samples (2099 spectra) was used to validate the selected classification model in a blind way.

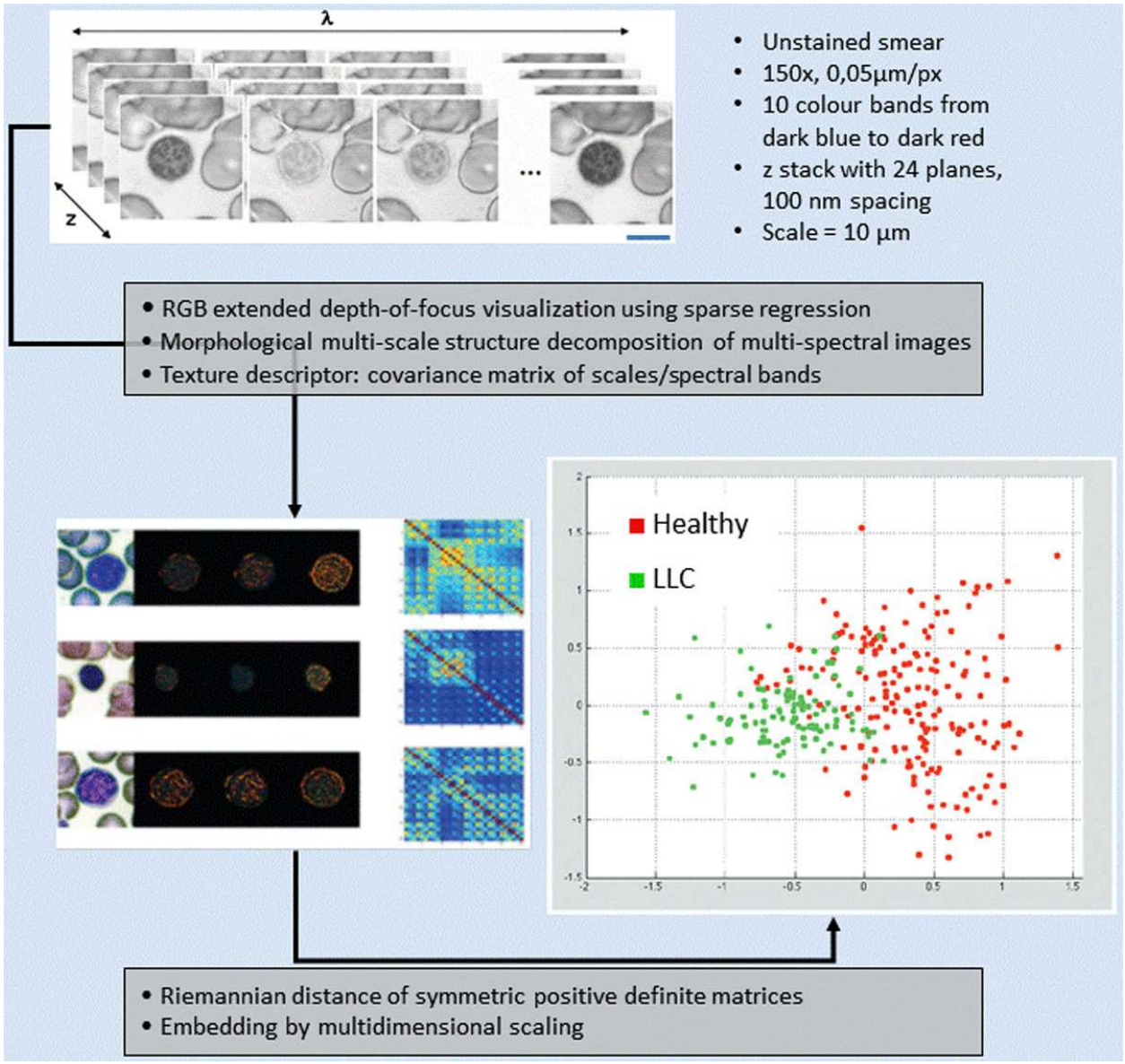


Figure 2 multi-spectral imaging feature extraction and classification method

Multi-spectral imaging

After acquisition of Raman spectra, the slides are stained using a RAL_DIAGNOSTICS™ standardized staining protocol. Then, for each target cell, a multi-z/multi-spectral image is acquired: a z-stack of 100 nm spaced of 24 monochrome images for each of the ten wavelength bands. An algorithm combining mathematical morphology techniques and sparse regression was developed to produce a color RGB image with an extended-depth-of-focus from the multi-z/multispectral image. This color image is used to visually confirm the classification of cells characterized by RMS. The set of scalar images multi-Z/multi-spectral of each cell is also processed by mathematical morphology techniques to produce morpho-spectral texture descriptors reducing the 240 images from each color band to a set of 4 representative images. Such data reduction process allowed representing each cell by a 40x40 square symmetrical correlation matrix. Then, using tools from information geometry and multivariate statistics, cells are embedded into a dimensionality reduced space used to produce an unsupervised classification into two classes: leukemia patients vs. healthy individuals.

Results and discussion

First stage molecular classification with Raman spectrometry aimed at identifying the nucleated components

from a blood smear, mainly polymorphonuclears and lymphocytes. Indeed such differentiation is straightforward with morphological methods like MSI even on unstained samples. Molecular classification gave a sensitivity of 99.3% and a specificity of 99.6%.

The second stage aimed at classifying spectra from leukemic and healthy lymphocytes and provided a sensitivity of 80% and a specificity of 100% on an extended set of blind samples.

It has to be mentioned that in a first step of this work, we tried to match the classification results to the percentages obtained with flow cytometry in term of T, B and NK lymphocytes of a blood sample. However, these results could not be compared with each other since in this study the information which is taken into account is obtained from the nucleus of each white blood cell. As a result, this is a molecular and nuclear information while, in classical cytology, this is the membrane of the cells which is considered.

Morpho-spectral texture classification from multispectral imaging in the visible is used to complement and consolidate the RMS data classifications and in on-going work will be used to identify specific morpho-spectral characteristics associated to leukemic cells (see figure 2).

Conclusion

IHMO project has demonstrated the power of Raman micro-spectroscopy coupled with supervised classification algorithm such as Support Vector Machines for cell classification and more precisely for the diagnosis of CLL. Morphological descriptors obtained from multi-Z and multispectral images provide another independent classification that still needs to be assessed.

The multimodal microscopy platform can be used more generally in the field of cytohematology, however application to cytological and histological pathology would need further developments and could take profit from new methods in data classification.

List of abbreviations

RMS: Raman Micro-Spectroscopy; MSI: Multi-Spectral Imaging; CLL: Chronic Lymphoid Leukemia; SVM: Support Vector Machines

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

• TH carried out the acquisitions, pretreatments, the highlighting of relevant information into the spectra, the classification of the data and drafted the extended abstract.
• AB drove the implementation of pretreatments adapted to spectral data and found a way to highlight the relevant information into the spectra.
• VU realized the acquisitions of the data and drafted the extended abstract.
• CG contributes to the choice of the classification algorithm among several studied (supervised or not), and participated in the choice of each possible settings of these algorithms to improve the obtained results.
• JA and SVF developed the multi-spectral classification method and produced the multi-spectral classification.
• XT, JK and MM both designed and managed the full project with the contribution of GF.
• XT, SD, VS and PCL selected the best CLL patients, provided the corresponding samples and patient's data, and realized the flow cytometry for each of them.

Acknowledgements

• The French National Research Agency (ANR) for funding
• Chih-Chung Chang and Chih-Jen Lin, LIBSVM: a library for support vector machines, 2001. Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm
• Clinical staff of the university hospital of Reims
• Rodolphe Dagiral and Valérie Coomans from Réactifs RAL for developing and providing standardizes MGG stains used for multi-spectral imaging classification
• Philippe Rideau, David Pointu and Claire Vernois from NIKON France for defining and installing appropriate microscopy hardware
• Emmanuel Froigneux from HORIBA JOBIN YVON for developing and providing the micro-spectroscopy Raman module

Authors' details

[1]MEDyC FRE/CNRS 3481, 51096 Reims, France. [2]CHU de Reims, 51100 Reims, France. [3]CHU de Caen, 14033 Caen, France. [4]CMM-ARMINES, 77300 Fontainebleau, France. [5]IGR-Institut Gustave-Roussy, 94805 Villejuif, France. [6]TRIBVN, 39, rue Louveau, 92320 Châtillon, France.

Published: 30 September 2013

References

1. Cramer P, Hallek M: **Prognostic factors in chronic lymphocytic leukemia what do we need to know?** *Nat Rev Clin Oncol* 2011, **8(1)**:38-47.
2. Zenz T, Mertens D, Küppers R, Döhner H, Stilgenbauer S: **From pathogenesis to treatment of chronic lymphocytic leukaemia.** *Nat Rev Cancer* 2010, **10(1)**:37-50.
3. Velasco-Forero S, Angulo J: **Parameters selection of morphological scale space decomposition for hyperspectral images using tensor modeling.** *Proc. of SPIE symposium on Defense, Security, and Sensing: Algorithms and Technologies for Multispectral, Hyperspectral, and Ultraspectral Imagery XVI, SPIE* 2010, **7695**:12.
4. Malik Z, Rothmann C, Cycowitz T, Cycowitz ZJ, Cohen AM: **Spectral morphometric characterization of B-CLL cells versus normal small lymphocytes.** *J Histochem Cytochem* 1998, **46(10)**:1113-8.
5. Anthony T: **Raman spectroscopy in biology.** NewYork: JohnWiley&Sons; 1982.
6. Draux F, Jeannesson P, Beljebbar A, Tfayli A, Fourre N, Manfait M, Sulé-Suso J, Sockalingum GD: **Raman spectral imaging of single living cancer cells: a preliminary study.** *Analyst* 2009, **134**:542-8.
7. Krishna CM, Sockalingum GD, Vadhiraja BM, Maheedhar K, Rao AC, Rao L, Venteo L, Pluot M, Fernandes DJ, Vidyasagar MS, Kartha VB, Manfait M: **Vibrational spectroscopy studies of formalin-fixed cervix tissues.** *Biopolymers* 2007, **85**:214-21.
8. Angulo J, Klossa J, Flandrin G: **Ontology-based lymphocyte population description using mathematical morphology on colour blood images.** *Cell Mol Biol* 2006, **52(6)**:2-15.
9. Cortes C, Vapnik V: **Support-Vector Networks.** *Machine Learning* 1995, **20**:273-297.

doi:10.1186/1746-1596-8-S1-S32
Cite this article as: Happillon *et al.*: **Label free technologies: Raman micro-spectroscopy and multi-spectral imaging for lymphocyte classification.** *Diagnostic Pathology* 2013 **8**(Suppl 1):S32.

Diagnosis approach of Chronic Lymphocytic Leukemia on unstained blood smears using Raman microspectroscopy and supervised classification

HAPPILLON Teddy[1], UNTEREINER Valérie[1], BELJEBBAR Abdelilah[1], GOBINET Cyril[1+], DALIPHARD Sylvie[2], CORNILLET-LEFEBVRE Pascale[2], QUINQUENEL Anne[3], DELMER Alain[3], TROUSSARD Xavier[4], KLOSSA Jacques[5], MANFAIT Michel[1].

[+]Corresponding author: cyril.gobinet@univ-reims.fr

[1]Biophotonique et Technologies pour la Santé, unité MEDyC CNRS FRE 3481, UFR de Pharmacie, SFR-CAP Santé, Université de Reims Champagne-Ardenne, 51 rue Cognacq-Jay, 51096, Reims Cedex, France.

[2]Laboratoire d'Hématologie Biologique, Hôpital Robert Debré, 51100 Reims, France.

[3]Service d'Hématologie Clinique, Hôpital Robert Debré, 51100 Reims, France.

[4]Laboratoire d'Hématologie Biologique, Centre Hospitalier Universitaire de Caen, 14033 Caen, France.

[5]TRIBVN, 39 rue Louveau, 92320 Châtillon, France.

Keywords: Raman microspectroscopy, support vector machines, chronic lymphocytic leukemia, human leukocytes, unstained blood smear.

Abstract:

We have investigated the potential of Raman microspectroscopy coupled with supervised classification algorithms to diagnose a blood lymphoproliferative disease, namely Chronic Lymphocytic Leukemia (CLL). To correlate Raman classifications with cell morphology, samples were stained after Raman measurements. For this study, standard glass slides were more convenient, and the 532nm laser excitation was a good compromise to reduce glass interferences and to improve signal/noise ratio with no sample damages. Preliminary classifications allowed the discrimination of data measured on the two main leucocyte subpopulations (lymphocytes and polymorphonuclears). These classifications aim to retain only lymphocyte cells for further investigations, *i.e.* normal vs CLL cells. Prediction models were built using Support Vector Machine algorithm (with or without 10-Fold Cross-Validation procedure) on a dataset composed of 1127 lymphocytes and 997 polymorphonuclears spectra. Their sensitivity and specificity were calculated over 98.5%.

The mean difference spectrum between these two classes shows higher nucleic acid content in lymphocyte compared to polymorphonuclears. These differences could be correlated to their nuclear-to-cytoplasm ratio. In the second part of this study, we have developed powerful classification models based on 6885 lymphocytes spectra acquired from 49 CLL patient and 27 volunteer samples. The mean sensitivity and specificity reach 88% and 93% respectively. The main molecular differences between healthy and CLL cells were associated to DNA and proteins changes. These spectroscopic markers could lead in the future to develop a helpful medical tool for CLL diagnosis.

Introduction:

B-cell Chronic Lymphocytic Leukemia (B-CLL) (Zenz, Mertens et al. 2009, Campo, Swerdlow et al. 2011) is a blood neoplastic disease usually diagnosed in people over 50 years old. It is the most common leukemia among adults in Western countries often detected at an early stage in patients without symptoms. B-CLL is defined by the accumulation of monoclonal mature B-cell composed in its classic form of monomorphic small round lymphocytes with clumped chromatin and scanty cytoplasm, expressing a distinct immunophenotype. Previous studies reported that genetic alterations contribute to tumorigenesis, clinical progression, and refractoriness of CLL to treatment, which could help to understand the origin of CLL.

A CLL is suspected when an isolated peripheral lymphocytosis persists more than three months. CLL diagnosis is based on laboratory tests such as complete blood cell count, including differential count, and lymphocytes immunophenotyping. Differential count is made by automatic blood cell counters but lymphocytes morphology of blood smears, stained with May-Grünwald Giemsa (MGG), is validated by a cytologist. Because of the morphological similarity between healthy and neoplastic lymphocytes, flow cytometry immunophenotyping is used to quantify the different lymphocytes subsets, and compute a Matutes score (Matutes, Owusu-Ankomah et al. 1994) to establish the CLL diagnosis.

The Rai (Rai, Sawitsky et al. 1975) and Binet (Binet, Auquier et al. 1981) classifications are used to define disease stage. The evaluation of other prognosis factors, such as ZAP-70 gene expression and cell doubling time, contribute to predict patient survival and suitable treatment. Almost all CLL treated patients show clinical ameliorations but to date, no cure is available. New approach such as Raman microspectroscopy would be useful tool, in order to provide rapid and complementary information to diagnose CLL disease at molecular level.

Raman microspectroscopy (Tu 1982) is a process based on an inelastic light scattering, called Raman diffusion. When the laser beam interacts with the sample, a small proportion of incident photons (around 1 for 108) coming from the excitation source are scattered with a corresponding change in frequency. The difference between the incident and the scattered frequencies corresponds to the vibrational modes of molecules participating in the interaction. Raman spectra represent fingerprints of the biomolecular composition of a sample. Raman technology is non destructive, with a high spatial resolution (around 1 µm),

and probes intrinsic molecular and biochemical information. This technology is rapid, requires no staining or any marking of the samples. Raman microspectroscopy combined with multivariate statistical analysis is thus of valuable interest for biomedical applications (Ellis and Goodacre 2006) to highlight the physiopathological modifications in cancer cells and tissues (Diem, Romeo et al. 2004), such as skin (Sigurdsson, Philipsen et al. 2004), lung (Huang, McWilliams et al. 2003), cervix (Vargis, Brown et al. 2012) and colon (Beljebbar, Bouché et al. 2009).

The aim of our present study is to diagnose CLL directly from human blood smears, obtained in the same conditions as all cytological preparations without MGG staining, using Raman microspectroscopy. CLL being a lymphoproliferative disease, the first step of this study is focused on the distinction between the two main leukocyte categories, i.e. lymphocytes and polymorphonuclears. The second part investigates the molecular changes between CLL and healthy lymphocytes. Supervised classification procedures are developed in order to construct predictive models able to distinguish between different cell categories (lymphocytes vs polymorphonuclears and CLL vs healthy).

Material and methods:

Samples:

Blood samples from 27 healthy volunteers and 49 untreated B-CLL patients were collected with their consent by the Centre Hospitalier Universitaire of Reims Champagne-Ardenne in France. The untreated leukemia patients have been selected according to different criteria: stage A in the Binet classification, a Matutes score over 3 and a CD19 percentage above 80% (i.e. at least 80% of the lymphocytes are neoplastic B-lymphocyte cells). Smears were manually prepared by spreading a blood drop on conventional glass slides, without prior chemical treatments.

Raman data acquisition:

To conduct this study, a multimodal machine has been developed by the Tribvn compagny (Châtillon, France) which combined a Raman spectrometer (Horiba Jobin Yvon, Villeneuve d'Ascq, France) and a conventional microscope (ECLIPSE FN1 Nikon SA, Champigny Sur Marne, France). This microscope was equipped with two dry objectives (Nikon SA,

Champigny Sur Marne, France), 40x and 100x (NA 0.6 and 0.9 respectively), and a motorized stage (Ludl Electronic Products, Ltd).

The Raman spectrometer is equipped with a 532 nm laser excitation source (Toptica Photonics, Munich, Germany) delivering 22 mW of output power and 13.5 mW of power on the sample. With confocal hole size about 100 µm, the spatial resolution is 2 µm. The spectral resolution is about 4 cm-1 by using a 1200 lines/mm dispersive grating. The Raman signal is recorded on a multichanel CCD detector (Andor Technology, South Windsor, CT, USA) with 1024x256 pixels, cooled at -70°C by Peltier effect.

Approximately a hundred white blood cells were localized on the smear by using a 40x microscope objective. Raman measurements were performed on each cell nucleus by using a 100x objective with 2 accumulations of 10 seconds, in the 702-3187 cm-1 spectral range. After Raman measurements, samples were stained with MGG procedure, cytologists classified each cell as lymphocyte, monocyte or polymorphonuclears and the corresponding spectrum was labeled accordingly.

Preprocessing:

In order to keep only relevant cell information for further classifications, it is essential to correct Raman spectra from glass background. The contribution of glass signal is estimated by ordinary least squares and subtracted from each spectrum. Spectra are then smoothed with a Savitzky-Golay filter (Savitzky and Golay 1964) using a window length of 9 points and a polynomial function of order 1.

Three quality tests are realized on these spectra. The first one computes a signal to noise ratio (S/N), by dividing the 2800-3150 cm-1 integrated signal band with the 1800-2200 cm-1 integrated noise band. This quality test retains only the spectra whose S/N ratios are over three times the standard deviation. When white and red blood cells are geographically close or superimposed, the measured spectrum is a linear combination between leucocyte and hemoglobin signals. This contamination level was evaluated with a second quality test by computing the integrated band ratio 1656-1720 cm-1 / 1540-1656 cm-1, representing leucocyte to hemoglobin fraction (S/H). This quality test retains only spectra whose S/H ratios are over 1. This threshold was determined by simulation tests, using different

combinations between pure lymphocytes and hemoglobin spectra. The third quality test allows detecting and removing spectra containing spikes by analyzing mean-centered data.

Due to the intrinsic fluorescence emission of the biological sample, the spectra were baseline corrected using a polynomial function of order 2, and then, normalized using the Standard Normal Variate method (Barnes, Dhanoa et al. 1989).

Statistical analysis:

After these preprocessing steps, two classification algorithms were constructed to distinguish between different cell categories, polymorphonuclears vs lymphocytes (P\L classification) and CLL vs healthy (C\H classification). The first classification is applied on the P\L dataset composed of 997 polymorphonuclear and 1127 lymphocyte spectra. The second classification uses 4257 CLL and 2596 healthy spectra (C\H dataset, see table I). Each dataset is divided into training, validation and prediction sets.

Subsets	Categories	Patients/Volunteers	Spectra
Training set	**CLL**	6	513
	Healthy	4	347
Validation set	**CLL**	33	2820
	Healthy	13	1106
Prediction set	**CLL**	10	924
	Healthy	10	1175

Table I: Composition of CH dataset.

Regions Of Interest (ROI):

In order to improve data classification sensitivity, specificity and to reduce computational time, the most discriminant wavenumbers (that compose ROI) between spectra from C\H dataset were considered using the MATLAB randfeatures function (Liu and Motoda 1998).

Supervised classification:

Support Vector Machine (SVM) (Cortes and Vapnik 1995, Vapnik 1999) combined to Raman spectroscopy has already been used as a powerful supervised classification algorithm in medical diagnosis (Widjaja, Zheng et al. 2008, Sattlecker, Bessant et al. 2010, Yan, Li et al.

2011). The aim of this method is to discriminate classes by finding the optimal hyperplane which allows maximizing the margin between them. In this study, an improvement of SVM algorithm, ν-SVM (Chang and Lin 2011), was considered and the Gaussian kernel radial basis function was chosen as ν-SVM kernel (Rüping 2001). 720 combinations of (σ, ν) parameters were tested (σ between 10-4 and 100 and ν from 0 to 1) to optimize classification result. For both datasets, the efficiency of each estimated model was evaluated by its sensitivity (% of correctly classified lymphocyte cells for P\L dataset or % of correctly classified CLL patients for C\H dataset) and specificity (% of correctly classified polymorphonuclear cells for P\L dataset or % of correctly classified healthy volunteers for C\H dataset).

Preliminary classifications, combining a 10-Fold Cross-Validation procedure with the ν-SVM algorithm (FCV-SVM), were achieved on training and validation sets of P\L dataset (respectively C\H dataset). Then, robust training-validation-prediction ν-SVM classifications were led on the full P\L dataset (respectively full C\H dataset). For each (σ, ν) couple, a prediction model was computed with the training set and validated on the validation set. Among the 720 generated models, the best one was chosen according to the best validation results, and was blindly tested on the prediction set.

Results:

In healthy volunteers, polymorphonuclears represent 50-70 % of white blood cells, lymphocytes 20-40 % and finally monocytes 3-8 %. A first step of this work was to discriminate between Raman spectra associated to the two main cell categories (lymphocytes vs polymorphonuclears) using multivariate statistical analysis.

Classification of Lymphocytes vs polymorphonuclears

First, FCV-SVM classification was achieved on training and validation parts of P\L dataset on the whole spectral region 703-3187 cm^{-1}. The result of this classification shows sensitivity and specificity of 99% and 98% respectively. Second, the training-validation-prediction procedure, applied on the full P\L dataset gave similar result (see table II).

	Lymphocytes (%)	Spectra	Polymorphonuclears (%)	Spectra
N-Fold-Cross Validation	98.9	723/731	96.9	635/655
Training set	98.8	327/331	96.7	266/275
Validation set	100	400/400	99.7	379/380
Test set	98.5	390/396	99.1	339/342

Table II: Results obtained for PL classifications.

To identify the differences between these two classes in term of molecular composition, mean Raman spectra recorded on 1127 lymphocytes and 997 polymorphonuclears were compared (Figure 1a and 1b). These spectra displayed Raman bands associated to proteins, lipids, and nucleic acids as the main constituents of cellular components. The region 700-800 cm^{-1} provides information on nucleic acids. The band at 1010 cm^{-1} is associated to breathing mode of phenylalanine of the proteins and protein bands at 1665 cm^{-1} and 1258 cm^{-1} are assigned to the amide I and amide III respectively. The band at 1342 cm^{-1} is associated to CH_3/CH_2 *twisting* vibration of proteins and nucleic acids. The band at 1455 cm^{-1} is attributed to CH_2 deformation mode arising from both proteins and/or lipids, and 1103 cm^{-1} band is associated to PO_2 symmetric *stretching* vibration of DNA. In the high wavenumber region, the Raman bands at 2886 cm^{-1}, 2941 cm^{-1} and 3066 cm^{-1} correspond to *stretching* of CH_2 and CH_3 groups of lipids and proteins respectively, and also CH of aromatic amino acids. To identify the molecular variations between these two classes, mean difference spectrum was calculated to highlight their main molecular changes (Figure 1c). This spectrum displays positive bands related to the vibrations of nucleic acid and backbone of DNA, such as 791 cm^{-1}, 1103 cm^{-1}, 1378 cm^{-1} and 1583 cm^{-1} bands. Indeed, nucleic acid content was higher in lymphocyte compared to polymorphonuclears which could be correlated to lymphocyte nuclear-to-cytoplasm ratio.

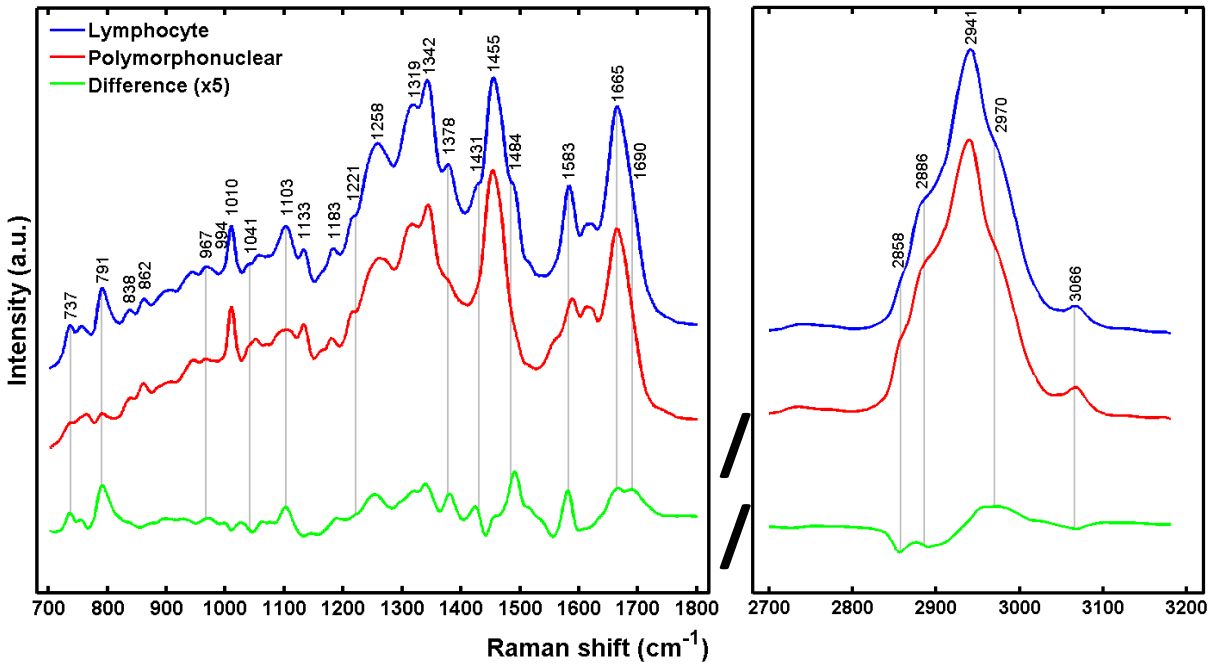


Figure 1: Mean spectra of lymphocyte and polymorphonuclear cells. Difference spectrum between them and annotated main spectral differences (gray lines).

CLL being a lymphoproliferative disease, the main leukocyte cells are lymphocytes with a small proportion of polymorphonuclears. The P\L classification model was then applied on the C\H dataset in order to check its discriminating power on CLL spectra. The classification result confirms that this model is able to distinguish between lymphocytes and polymorphonuclears as well as healthy or neoplastic samples. In this study we investigated B-CLL which affects only lymphocytes cells. The aim of this P\L classification was to identify and exclude all spectra associated to polymorphonuclears for further discrimination between healthy and neoplastic cells.

Classification of Healthy vs CLL lymphocytes

In the second part of this study, we have investigated the potential of Raman spectroscopy combined to SVM algorithm to discriminate between healthy and CLL lymphocyte cells recorded from 27 volunteers and 49 CLL patients.

FCV-SVM procedure was applied on the training and validation parts of C\H dataset on the full spectral range 703-3187cm^{-1}. Our classification result provides a sensitivity (related to CLL patients) of 97.4%, and a specificity (associated to healthy volunteers) of 88.2%. The second test, using the training-validation-prediction procedure on the full C\H dataset, was performed to create and validate a more robust SVM model than the FCV-SVM one. As shown in table IIIa, the sensitivity and the specificity reach respectively 80% and 50%. By using the whole spectral range, the discriminant power of the model is less efficient than the FCV-SVM one. In order to improve C\H data classification, the last procedure was repeated on spectra reduced to ROI. Our best model allowed a significant increase in sensitivity and specificity on training, validation and prediction sets except for the training set sensitivity (see Table III-B). This loss in training set sensitivity, between the full spectra and ROI, could be due to an overfitting of the full spectra training model.

		Sensitivity (%)	Patients	Specificity (%)	Volunteers
A					
Full spectra	N-Fold-Cross Validation	97,4	38/39	88.2	15/17
	Training set	100	6/6	75	3/4
	Validation set	81.8	27/33	61.5	8/13
	Test set	80	8/10	50	5/10
B					
ROI	Training set	83.3	5/6	100	4/4
	Validation set	90.9	30/33	84.6	11/13
	Test set	80	8/10	100	10/10

Table III: Results obtained for CH classifications by considering the full spectral range (A) and spectra reduced to ROI (B).

To correlate the C\H classification results with the molecular changes between normal and neoplastic cells, the difference spectrum between the mean Raman spectra measured on

2596 healthy lymphocytes and 4257 leukemic cells was performed (Figure 2). This spectrum shows positive peaks at 796 cm^{-1}, 1106 cm^{-1}, 1380 cm^{-1}, 1498 cm^{-1} and 1585 cm^{-1} indicating that the overall DNA concentration is higher in neoplastic B-lymphocyte cells compared to healthy ones. The presence of some positive bands at 1456 cm^{-1} and 1666 cm^{-1} due to vibrations of proteins shows an increase in protein content in the CLL cells. Concerning the phenylalanine band, a shift can be observed on the CLL mean spectra. In high wavenumber region, the difference spectrum shows a negative band at 2922 cm^{-1}, attributed to CH stretch of both lipids and proteins, indicates a little decrease in the difference spectrum. A positive peak at 2954 cm^{-1} indicates a higher stretching vibration of CH_3 into the mean CLL spectrum than in the healthy one which confirms the results obtained in the fingerprint region. Beyond 3000 cm^{-1}, Raman spectral band are attributed to CH stretching such for 3062 cm^{-1} and 3079 cm^{-1} bands which both present little variations between CLL and healthy mean spectra.

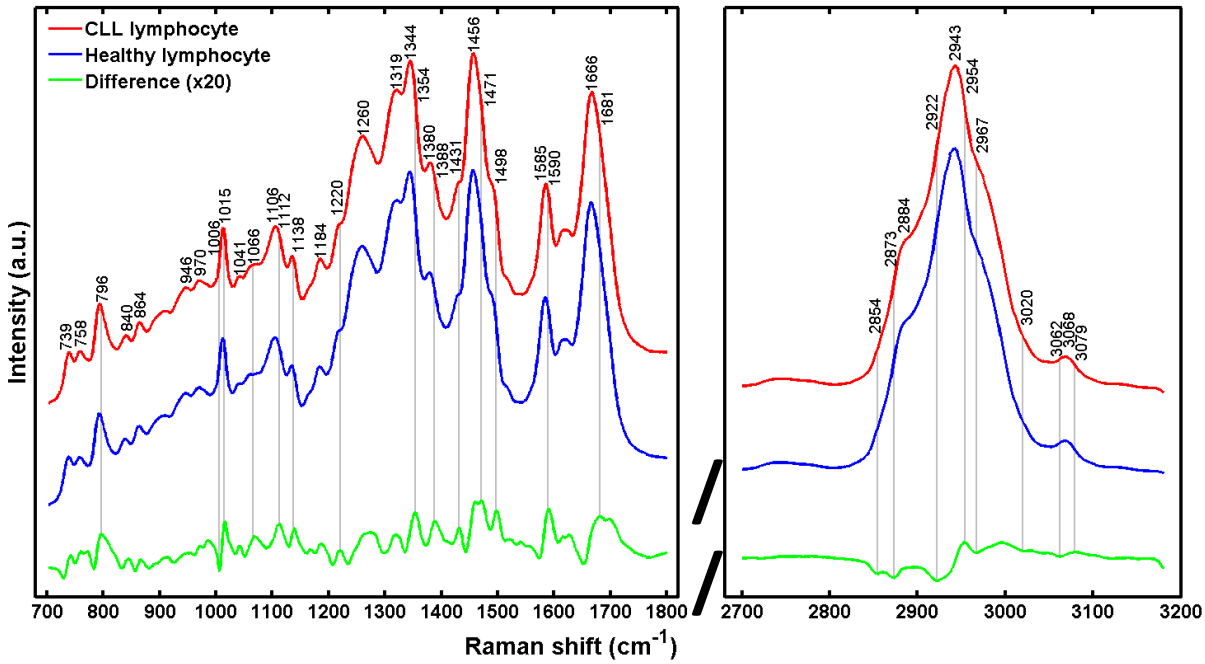


Figure 2: Mean spectra of CLL and healthy lymphocyte cells. Difference spectrum between them and annotated main spectral differences (gray lines).

These intrinsic Raman markers associated with vibration modes of DNA and proteins could be used as spectroscopic markers to discriminate between healthy and pathological states.

In the figure 3, ROIs are plotted in order to assign all discriminant wavenumbers to molecular characteristics. Most of ROIs are correlated to changes in nucleic acids and proteins content between healthy and CLL lymphocytes. This result corroborates with that observed using the difference spectrum (Figure 3c). Indeed, 831 cm^{-1} and 1464 cm^{-1} spectral bands are associated to DNA and RNA content. Other bands such as 1429-1468 cm^{-1}, 2874 cm^{-1}, and 2835 cm^{-1} can be correlated to proteins and/or lipids.

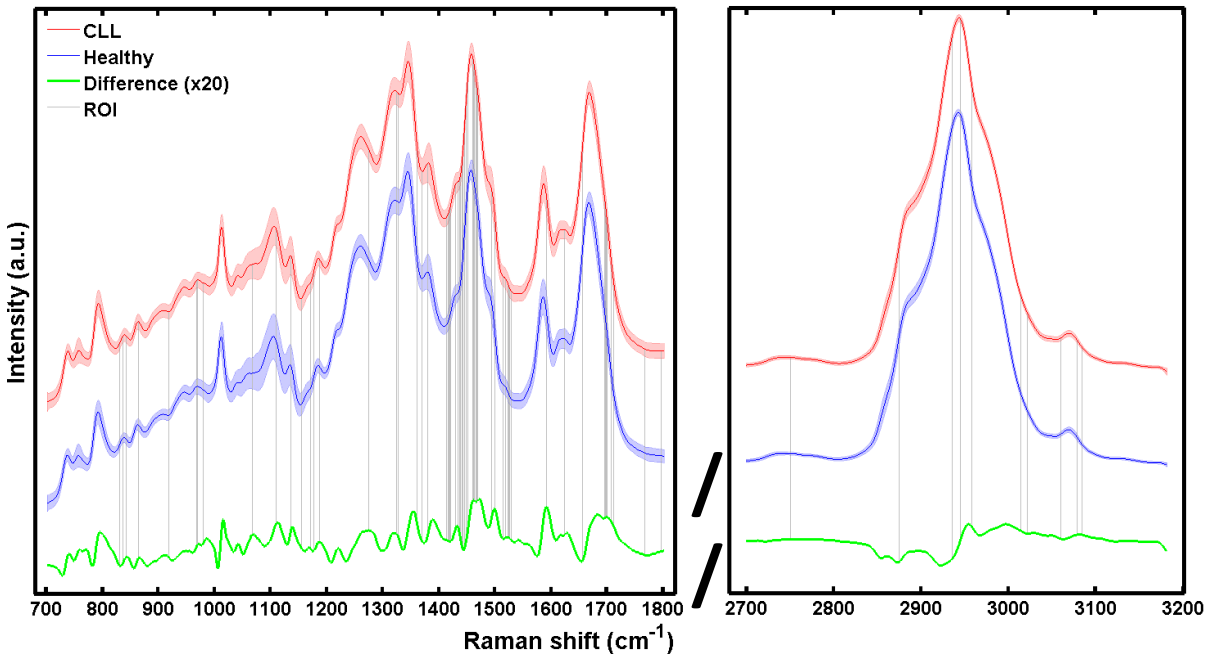


Figure 3: Mean spectrum and standard deviation for both CLL (red) and healthy (blue) lymphocytes (4257 and 2596 spectra respectively). Difference between the two mean spectra (green). Region Of Interest, composed of 74 wavenumbers, retained during the classification in three steps on CH dataset.

Discussion

This study demonstrated the potential of confocal Raman microspectroscopy as a clinical tool to discriminate the two main cell categories, lymphocyte and polymorphonuclear on healthy and CLL unstained blood smears, and to classify lymphocytes as normal or pathological based on intrinsic spectroscopic fingerprints. Our results showed that the main differences between these two main cell categories are the more prominent vibrational bands attributed to nucleic acids in the lymphocytes spectra than in the polymorphonuclears ones. These observations corroborate with previous studies (Sijtsema, Otto et al. 1998, van Manen, Kraan et al. 2005). Such spectral differences can be explained by the very large

nucleus to cell ratio of lymphocytes, as observed under histochemical staining (Ochiai and Eguchi 1987). Therefore, the nucleus occupying 80% of the cell likely affects the compactness of the chromatin and thus the concentration of DNA in the probe volume (Ramoji, Neugebauer et al. 2012).

We have also demonstrated that the main distinguishing spectral features between normal and CLL cells are due to different DNA and protein contents. These molecular changes could be used as spectroscopic markers to discriminate between healthy and pathological states. Chan et al. reported that Raman microspectroscopy is able to discriminate healthy cells of volunteers from transformed T Jurkat and B Raji culture cell lines or human healthy lymphocyte cells from CLL cells isolated by Ficoll procedure, based on their DNA, RNA and protein content (Chan, Taylor et al. 2006, Chan, Taylor et al. 2008, Poplineau, Trussardi-Régnier et al. 2011). Previous works have reported that the main differences between healthy and CLL cells from volunteers and patients are associated to a reduction in lipid content, an increase in the DNA to proteins and DNA to RNA ratios using Fourier transform infrared microspectroscopy (FTIR) (Benedetti, Papineschi et al. 1983, Benedetti, Palatresi et al. 1985, Schultz, Liu et al. 1996, Benedetti, Bramanti et al. 1997). Additionally, Schultz et al. demonstrated that the DNA content could be used as spectroscopic markers to provide a useful diagnostic tool for staging (disease progression) (Schultz, Liu et al. 1997). These studies had measured cell or pellet using FTIR spectroscopy whereas Raman method probes subcellular compartment with high spatial resolution (1µm) with a minimal preparation compared to FTIR.

Over the past several years, most of Raman studies at cellular and tissular levels have used 785nm as excitation wavelength to minimize sample fluorescence and prevent sample damage. The common supports adapted for this wavelength excitation are BaF2, CaF2, and quartz because of their low interference background with the sample. In this study, standard glass slide was required since it is used in clinical routine for sample preparation and MGG staining. The laser excitation wavelength at 785nm was not appropriate in our study because glass slides provide a very strong background signal which completely circumvent Raman signal originating from cells. A 532nm laser excitation represents a good compromise to improve signal/noise ratio, due to better scattering and detector efficiency, and to reduce glass slide interferences. With 13mW of laser power on the sample and 10 seconds

integration time, no damages (photobleaching) were observed on cells after Raman measurements. This observation was confirmed by cytological analysis where morphology of cells was preserved. Puppels et al. showed that integrity of living biological samples, especially cell viability, depends on the laser light wavelength and power on the sample (Puppels, Olminkhof et al. 1991). They also reported that the use of 632.8 or 660nm wavelengths excitation on human lymphocytes strongly diminished the spectral degradation, with cells surviving 5 minutes irradiations at 20mW of laser power (Puppels, Olminkhof et al. 1991). All Raman spectra were recorded on leucocyte nucleus (lymphocyte and polymorphonuclear). We have attempted to measure Raman spectra on cytoplasm to verify if they could bring complementary information improving the two above classifications. The results do not provide any useful or additional information because the Raman measured area is higher than cytoplasm in most of lymphocytes. Thus, Raman spectra recorded in this compartment are a linear combination between glass, membrane, cytoplasm and nucleus signals.

SVM combined to features selection was used to successfully diagnose CLL disease with high sensitivity and specificity. Feature selection methods (univariate variable ranking and multivariate variable subset selection) are commonly used to retain only efficient wavenumbers in order to reduce high dimensionality of Raman data and improve classification results (Song, Molckovsky et al. 2005, Teh, Zheng et al. 2009, Bergholt, Zheng et al. 2011). In our study, the specificity of validation and prediction sets were increased by +23% and +50% respectively by using a multivariate method instead of univariate counterpart (data not shown). Classification algorithms such as SVM, Artificial Neural Network (ANN) or Decision Trees are powerful and nonlinear supervised methods in medical diagnosis (Yegnanarayana 2004) (Buntine 2013) (Ellis and Goodacre 2006). In our study, the SVM algorithm was successfully applied on P\L and C\H datasets since its performances are equivalent to both ANN and Decision Trees with a short processing time (Byvatov, Fechner et al. 2003, Caruana and Niculescu-Mizil 2006, Trevisan, Angelov et al. 2012).

In the literature, some studies have investigated Raman and infrared vibrational spectroscopies as efficient tools able to discriminate either cell lines or human cells of CLL patients from healthy volunteers ones (Schultz, Liu et al. 1996, Jackson, Sowa et al. 1997, Liu, Schultz et al. 1997, Schultz, Liu et al. 1997, Chan, Taylor et al. 2006, Chan, Taylor et al. 2008).

In this aim, they considered classification algorithms, supervised or not, like the Linear Discriminant Analysis algorithm (LDA), coupled or not with a Principal Component Analysis (PCA), or the Hierarchical Clustering Algorithm (HCA). In a recent work, Trevisan et al (Trevisan, Angelov et al. 2012) reported that these algorithms belong to the most efficient statistical methods employed in a histopathology diagnosis goal. However, some weaknesses are also pointed out. Indeed, HCA is known to be time-consuming especially when the number of data increases, as well as LDA algorithm is not usually complex enough to correctly discriminate data with thin variations. Other efficient supervised classification algorithms were cited as Artificial Neural Network (ANN), Partial Least Square Discriminant Algorithm (PLS-DA), K-Nearest Neighbors (KNN) or Support Vector Machines. This latest appeared the most popular in bioinformatics because of its ability to offer different ways to find pertinent classification solutions, i.e. classifications are based either on linear or nonlinear kernel functions.

Conclusion

This study report the use of Raman spectroscopy combined to supervised statistical analysis to differentiate first the two main leukocyte subpopulations and to discriminate 29 human healthy volunteers from 49 CLL patients with a mean sensitivity and specificity of 88% and 93% respectively. These classifications are based on specific spectroscopic fingerprints obtained under standard experimental clinical conditions without external staining. We have highlighted that the main intrinsic molecular differences between normal and CLL cells are related to changes in nucleic acids and protein contents. These spectroscopic fingerprints were used to build robust predictive models that could potentially aid, in the future, as an automatic, rapid and objective diagnosis tool for CLL disease in clinical routine. This work represents a first step before the application of Raman spectroscopy to evaluate its potential for CLL prognosis by correlating spectroscopic data with all clinical analysis applied in routine to predict the disease progression and to determine the most efficient patient treatment.

References

Barnes, R., et al. (1989). "Standard normal variate transformation and de-trending of near-infrared diffuse reflectance spectra." Applied spectroscopy 43(5): 772-777.

Beljebbar, A., et al. (2009). "Identification of Raman spectroscopic markers for the characterization of normal and adenocarcinomatous colonic tissues." Critical reviews in oncology/hematology 72(3): 255-264.

Benedetti, E., et al. (1997). "Determination of the relative amount of nucleic acids and proteins in leukemic and normal lymphocytes by means of Fourier transform infrared microspectroscopy." Applied spectroscopy 51(6): 792-797.

Benedetti, E., et al. (1985). "New possibilities of research in chronic lymphatic leukemia by means of Fourier transform-infrared spectroscopy-II." Leukemia research 9(8): 1001-1008.

Benedetti, E., et al. (1983). "Analytical infrared spectral differences between human normal and leukaemic cells (CLL)—I." Leukemia research 8(3): 483-489.

Bergholt, M. S., et al. (2011). "In vivo diagnosis of gastric cancer using Raman endoscopy and ant colony optimization techniques." International journal of cancer 128(11): 2673-2680.

Binet, J., et al. (1981). "A new prognostic classification of chronic lymphocytic leukemia derived from a multivariate survival analysis." Cancer 48(1): 198-206.

Buntine, W. L. (2013). "Decision tree induction systems: a Bayesian analysis." arXiv preprint arXiv:1304.2732.

Byvatov, E., et al. (2003). "Comparison of support vector machine and artificial neural network systems for drug/nondrug classification." Journal of Chemical Information and Computer Sciences 43(6): 1882-1889.

Campo, E., et al. (2011). "The 2008 WHO classification of lymphoid neoplasms and beyond: evolving concepts and practical applications." Blood 117(19): 5019-5032.

Caruana, R. and A. Niculescu-Mizil (2006). An empirical comparison of supervised learning algorithms. Proceedings of the 23rd international conference on Machine learning, ACM.

Chan, J. W., et al. (2008). "Non-destructive Identification of Individual Leukemia Cells by Optical Trapping Raman Spectroscopy." Analytical Chemistry, vol. 80, no. 6, February 9, 2008, pp. 2180-2187 80(UCRL-JRNL-229152).

Chan, J. W., et al. (2006). "Micro-Raman spectroscopy detects individual neoplastic and normal hematopoietic cells." Biophysical journal 90(2): 648-656.

Chang, C.-C. and C.-J. Lin (2011). "LIBSVM: a library for support vector machines." ACM Transactions on Intelligent Systems and Technology (TIST) 2(3): 27.

Cortes, C. and V. Vapnik (1995). "Support-vector networks." Machine learning 20(3): 273-297.

Diem, M., et al. (2004). "A decade of vibrational micro-spectroscopy of human cells and tissue (1994–2004)." Analyst 129(10): 880-885.

Ellis, D. I. and R. Goodacre (2006). "Metabolic fingerprinting in disease diagnosis: biomedical applications of infrared and Raman spectroscopy." Analyst 131(8): 875-885.

Huang, Z., et al. (2003). "Near-infrared Raman spectroscopy for optical diagnosis of lung cancer." International journal of cancer 107(6): 1047-1052.

Jackson, M., et al. (1997). "Infrared spectroscopy: a new frontier in medicine." Biophysical chemistry 68(1): 109-125.

Kohavi, R. (1995). "A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection." Proceedings of the 14th international joint conference on Artificial intelligence 2: 1137-1143.

Liu, H. and H. Motoda (1998). Feature selection for knowledge discovery and data mining, Springer.

Liu, K.-Z., et al. (1997). "Comparison of infrared spectra of CLL cells with their< i> ex vivo</i> sensitivity (MTT assay) to chlorambucil and cladribine." Leukemia research 21(11): 1125-1133.

Matutes, E., et al. (1994). "The immunological profile of B-cell disorders and proposal of a scoring system for the diagnosis of CLL." Leukemia: official journal of the Leukemia Society of America, Leukemia Research Fund, UK 8(10): 1640.

Molckovsky, A., et al. (2003). "Diagnostic potential of near-infrared Raman spectroscopy in the colon: differentiating adenomatous from hyperplastic polyps." Gastrointestinal endoscopy 57(3): 396-402.

Ochiai, F. and M. Eguchi (1987). "Morphometrical evaluation of acute leukemic cells by electron microscopy. Discrepancy between morphological characteristics in FAB classification and electron microscopic morphometry." Virchows Archiv. B, Cell pathology including molecular pathology 52(5): 403.

Poplineau, M., et al. (2011). "Raman microspectroscopy detects epigenetic modifications in living Jurkat leukemic cells." Epigenomics 3(6): 785-794.

Puppels, G., et al. (1991). "Laser irradiation and Raman spectroscopy of single living cells and chromosomes: sample degradation occurs with 514.5 nm but not with 660 nm laser light." Experimental cell research 195(2): 361-367.

Rai, K. R., et al. (1975). "Clinical staging of chronic lymphocytic leukemia." Blood 46(2): 219-234.

Ramoji, A., et al. (2012). "Toward a Spectroscopic Hemogram: Raman Spectroscopic Differentiation of the Two Most Abundant Leukocytes from Peripheral Blood." Analytical chemistry 84(12): 5335-5342.

Rüping, S. (2001). SVM kernels for time series analysis. Dortmund, Citeseer.

Sattlecker, M., et al. (2010). "Investigation of support vector machines and Raman spectroscopy for lymph node diagnostics." Analyst 135(5): 895-901.

Savitzky, A. and M. J. Golay (1964). "Smoothing and differentiation of data by simplified least squares procedures." Analytical chemistry 36(8): 1627-1639.

Schultz, C. P., et al. (1996). "Study of chronic lymphocytic leukemia cells by FT-IR spectroscopy and cluster analysis." Leukemia research 20(8): 649-655.

Schultz, C. P., et al. (1997). "Prognosis of chronic lymphocytic leukemia from infrared spectra of lymphocytes." Journal of molecular structure 408: 253-256.

Sigurdsson, S., et al. (2004). "Detection of skin cancer by classification of Raman spectra." Biomedical Engineering, IEEE Transactions on 51(10): 1784-1793.

Sijtsema, N., et al. (1998). "Resonance Raman Microspectroscopy of Myeloperoxidase and Cytochrome< i> b</i>< sub> 558</sub> in Human Neutrophilic Granulocytes." Biophysical journal 74(6): 3250-3255.

Singh, S., et al. (2013). "In vivo Raman spectroscopy of oral buccal mucosa: a study on malignancy associated changes (MAC)/cancer field effects (CFE)." Analyst.

Song, L.-M. W. K., et al. (2005). Diagnostic potential of Raman spectroscopy in Barrett's esophagus. Proceeding of SPIE.

Teh, S. K., et al. (2009). "Near-infrared Raman spectroscopy for gastric precancer diagnosis." Journal of Raman Spectroscopy 40(8): 908-914.

Trevisan, J., et al. (2012). "Extracting biological information with computational analysis of Fourier-transform infrared (FTIR) biospectroscopy datasets: current practices to future perspectives." Analyst 137(14): 3202-3215.

Tu, A. T. (1982). Raman spectroscopy in biology: principles and applications, Wiley New York.

Van Manen, H.-J., et al. (2005). "Single-cell Raman and fluorescence microscopy reveal the association of lipid bodies with phagosomes in leukocytes." Proceedings of the National Academy of Sciences of the United States of America 102(29): 10159-10164.

Vapnik, V. (1999). The nature of statistical learning theory, springer.

Vargis, E., et al. (2012). "Detecting Biochemical Changes in the Rodent Cervix During Pregnancy Using Raman Spectroscopy." Annals of biomedical engineering: 1-11.

Widjaja, E., et al. (2008). "Classification of colonic tissues using near-infrared Raman spectroscopy and support vector machines." International journal of oncology 32(3): 653-662.

Yan, B., et al. (2011). "Discrimination of parotid neoplasms from the normal parotid gland by use of Raman spectroscopy and support vector machine." Oral oncology 47(5): 430-435.

Yegnanarayana, B. (2004). Artificial neural networks, PHI Learning Pvt. Ltd.

Zenz, T., et al. (2009). "From pathogenesis to treatment of chronic lymphocytic leukaemia." Nature Reviews Cancer 10(1): 37-50.

Article #VII

Publication n°7

Quantitative phase imaging and Raman micro-spectroscopy applied to Malaria

J. Klossa, B. Wattelier, **T. Happillon**, D. Toubas, L. de Laulanie, V. Untereiner, P. Bon, M. Manfait.

Diagnostic Pathology, 2013 ;8 (Suppl 1) : S42 doi:10.1186/1746-1596-8-S1-S42

Présentation de la publication n°7

Contexte

La Malaria (ou Paludisme selon la région du globe) est une maladie qui peut être mortelle chez l'homme. Elle est due à un parasite, le *Plasmodium*, transmis par les moustiques qui en sont porteurs. Il existe 4 types de *Plasmodium* (*falciparum, vivax, malariae* et *ovale*) parmi lesquels *falciparum* est à la fois le plus répandu et le plus mortel. Lors d'une infection, le parasite se multiplie dans le foie, puis infecte les hématies du porteur. Il s'écoule généralement entre 10 et 15 jours avant que les premiers symptômes n'apparaissent, notamment sous forme de fièvre, maux de tête et vomissements. Si elle n'est pas traitée suffisamment tôt, l'infection des globules rouges par les parasites entraine des troubles de la circulation sanguine menant inévitablement au décès. Afin d'établir le diagnostic et ainsi statuer sur le traitement approprié, trois informations sont nécessaires, à savoir l'espèce, le taux de globules rouges infectés (parasitémie) et le stade de développement du parasite. Dans ce but, une goutte épaisse et un frottis sanguin colorés au MGG sont observés au microscope par un spécialiste, nécessitant à la fois du temps et des ressources humaines disponibles.

Objectif

L'objectif de cette étude préliminaire est donc de montrer qu'il est possible de détecter et dénombrer les parasites présents dans les hématies infectées. Ceci en les repérant par Imagerie de Phase Quantitative (IPQ) d'une part, et en enregistrant sur chacun d'eux un spectre Raman d'autre part. La microspectroscopie Raman est ici considérée comme un outil de validation des repérages préalablement réalisés. En effet, représentant de véritables signatures moléculaires, les spectres Raman peuvent contenir un signal propre à un biocristal formé par le parasite, l'hémozoïne, et ainsi confirmer sa présence dans les cellules repérées.

Matériels et Méthodes

L'échantillon analysé se présente sous la forme d'un frottis sanguin non coloré et non fixé, issu d'un patient infecté par *Plasmodium falciparum* (parasitémie de 3,5%). 10 zones d'intérêt sur le frottis sont sélectionnées en champ large (Objectif 40x, ON=0,6 et Zoom=1,

Nikon SA, Champigny Sur Marne, France) puis analysées par IPQ (SID4bio, Phasics, Palaiseau). Pour chacune, une image de phase de 300x400 pixels est obtenue avec une résolution de 0,74µm/pixel, autorisant le repérage spatial des cellules infectées sur les zones d'intérêt. Le module Raman utilisé dans cette étude est similaire à celui des études menées dans le cadre du diagnostic de la LLC présenté précédemment (publications 4, 5 et 6). Les spectres Raman sont enregistrés avec un temps d'accumulation de 10 secondes, sur 34 parasites repérés et 35 hématies non infectées. La distorsion de leur ligne de base est corrigée, puis ces spectres sont normalisés, et un signal Raman représentatif de l'hémoglobine pure leur est soustrait. Les spectres corrigés sont ensuite réduits aux gammes spectrales propres au signal de l'hémozoïne (*i.e.* 1380-1385, 1594-1597 et 1644-1646 cm^{-1}), et classifiés par Hierarchical Clustering Analysis (HCA).

Résultats

Dans cette étude, 640 globules rouges en moyenne ont été repérés par champ observé en IPQ (avec un écart type de 23). 4,1% des hématies ont été détectées en moyenne comme infectées (avec un écart type de 1,1%), ce qui corrobore le pourcentage de parasitémie de la lame étudiée (3,5%). La classification par HCA des spectres Raman prétraités a montré une distinction parfaite avec une sensibilité et une spécificité de 100% pour les deux populations spectrales représentées (*i.e.* hématies saines et parasites).

Conclusion

Les résultats obtenus lors de cette étude préliminaire ont montré le fort potentiel des deux technologies employées (*i.e.* l'IPQ et la micro-spectroscopie Raman) lorsqu'elles sont couplées dans le but de détecter, puis confirmer, la présence du parasite *Plasmodium* dans les hématies d'un patient. Les conditions de préparation des échantillons (simples et similaires à celles réalisées en routine), ainsi que la rapidité d'analyse par ces technologies montrent le fort potentiel du module IHMO pour être employé comme outil diagnostique de la Malaria, en cliniques et laboratoires. Ces résultats prometteurs nécessitent cependant des études plus approfondies, prenant en compte une quantité d'échantillons beaucoup plus conséquente et des méthodes chimiométriques plus fines, afin de pouvoir non seulement détecter la présence de *Plasmodium*, mais également son espèce et son stade de développement.

Klossa et al. Diagnostic Pathology 2013, 8(Suppl 1):S42
http://www.diagnosticpathology.org/content/8/S1/S42

DIAGNOSTIC PATHOLOGY

PROCEEDINGS Open Access

Quantitative phase imaging and Raman micro-spectroscopy applied to Malaria

Jacques Klossa[1*†], Benoit Wattelier[2†], Teddy Happillon[3†], Dominique Toubas[3,4†], Lucie de Laulanie[2†], Valérie Untereiner[3†], Pierre Bon[2†], Michel Manfait[1†]

From 11th European Congress on Telepathology and 5th International Congress on Virtual Microscopy
Venice, Italy. 6-9 June 2012

Background

Malaria is due to parasitism of red blood cells (RBC) by protozoan parasites of the genus Plasmodium. Three main parameters have to be determined for patient treatment: parasite species, the rate of infected blood cells (parasitemia), and development stage. Even if a series of laboratory techniques are available, a suited treatment needs microscopy skills [1]. Microscopic observation needs a specialist and is time consuming (e. g. observation of hundreds fields of view at 100x immersion objective) and automating 100x slide scanning of white light imaging of thin film stained blood smears is not straightforward.

Seeking for an easy to automatize alternative, we thought that combining two microscopy techniques: quantitative phase imaging for quick detection and Raman micro-spectroscopy for molecular characterization could appear to be an efficient multipurpose solution.

Raman micro-spectroscopy is a good candidate to identify molecular species in microscopic. Laser light is focused in a tiny volume. Due to molecular vibrations, part of the light is non-linearly scattered to longer (Stokes) or shorter (Anti-Stokes) wavelength. The wavelength shift is directly linked to the vibration energy. The scattered light spectrum contains lines typical of the molecular binding in the focused volume. However imaging is very time consuming as each pixel needs to be acquired individually (between 1 and 10s per pixel).

Quantitative phase imaging is an imaging technique that measures the optical path difference of light travelling through different part of a semi-transparent medium. If a biological tissue has local different index of refraction, we obtain an image that reflects these index changes. This kind of maps usually has a contrast more than two orders of magnitude higher than bright field microscopic images. In addition, they give information about the relative index change in the tissue.

Combining a morphological technique suited for rare events detection with a molecular technique suited for local molecular signature acquisition could provide an easy way to address automation of parasite parameters acquisition.

In the following sections, we present our first level attempt for proof of concept.

Material and methods

Sample preparation and hardware

We used for our experiments the solution developed during the IHMO program (see "Label free technologies: Raman micro-spectroscopy and multi-spectral imaging for Lymphocyte classification" in the current publication). It uses the Calopix TRIBVN^{TM} software platform integrated to the hospital workflow and off the shelves components NIKON FN1^{TM} microscope fitted with a Raman micro-spectroscopy (RMS) HORIBA module whose excitation source is a 532nm diode laser. In addition a commercial quadriwave lateral shearing interferometer (QWLSI Phasics SID4Bio^{TM}) is plugged on one of the microscope exit port; it measures the quantitative phase shift in the visible spectral band. This is an easy-to-integrate and compact solution (dimension of a simple camera) and gives a 300x400px^2 (px=pixel) phase and intensity image with a lateral pitch of p=29.6 μm in the image plane, which corresponds to 0.74μm/px for 40x magnification and 0.197μm/px for 150x magnification.

Blood smears are prepared on a standard glass slide with no slide cover and without any previous staining to

* Correspondence: jklossa@tribvn.com
† Contributed equally
[1]TRIBVN, 39, rue Louveau, 92320 Châtillon, France
Full list of author information is available at the end of the article

BioMed Central
© 2013 Klossa et al; licensee BioMed Central Ltd. This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/2.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

avoid contamination of the recorded Raman spectra with fluorescence contribution. Spectra are acquired on infected RBC and finally, slides are stained with May-Grünwald Giemsa for morphological validation.

Quantitative phase imaging

The presence of the parasite inside the RBC implies a synthesis of hemozoin (complex molecular structure) from hemoglobin. These two molecules present a different refractive index; therefore the parasite should appear with a different quantitative phase inside the RBC. We have checked this assumption at high magnification (150x) and we found out that the parasite induces lower phase values (figure 1). Afterward, we studied large fields of view at lower magnification: 40x proved to be efficient: see a typical field of view figure 2. In order to automatically detect the infected cells, we first applied a high pass filter. This removes the slowly varying cell shape while keeping the plasmodium image. Then low phase values due to the plasmodium were detected by three-level segmentation. One level corresponds to the extracellular medium; the second level corresponds to the RBC cytoplasm containing hemoglobin and the third one to the hemozoin. The second level segmentation leads to the number of the RBC contained in the field. The third level segmentation gives the number of plasmodia and thus of infested RBCs. The ratio between these two numbers is the parasitemia. We have studied 10 fields of a single blood smear containing falciparum parasites.

Raman micro-spectroscopy

Raman spectra are recorded on parasite from infected RBC localized thank to QPI at 40x. Raw spectra need first a pre-treatment phase to make them eligible in the classification process. Indeed, the presence of hemoglobin into the cells, and more precisely the fact that hemoglobin contains cyclic structures, implies distortions of the baseline of each spectrum. This distortion effect is called a fluorescence background. To correct the spectra, a function based on a polynomial estimation and correction of the baseline is used for each spectrum independently of each other. Then, the normalization function, standard normal variate, is used on each spectrum. This function eliminates the variation of the absolute values into the spectra, and makes them comparable avoiding scale differences. Then, a first classification could be realized. For such purpose, we used the classical Hierarchical Clustering Analysis (HCA). HCA is based on the Euclidean distance between each observation (each spectrum) and from these distances a hierarchical binary tree is created, putting forward the different groups of spectra. Representative peaks of the hemozoin biocrystal, previously identified in literature, are taken into account to classify the spectra, by reducing them to the corresponding region of interest. In a second classification, and to improve the first results, a representative spectrum of pure hemoglobin, which has been obtained by averaging RBC spectra from other samples, has then been estimated and subtracted from each recorded spectrum, thanks to a mean squared based function.

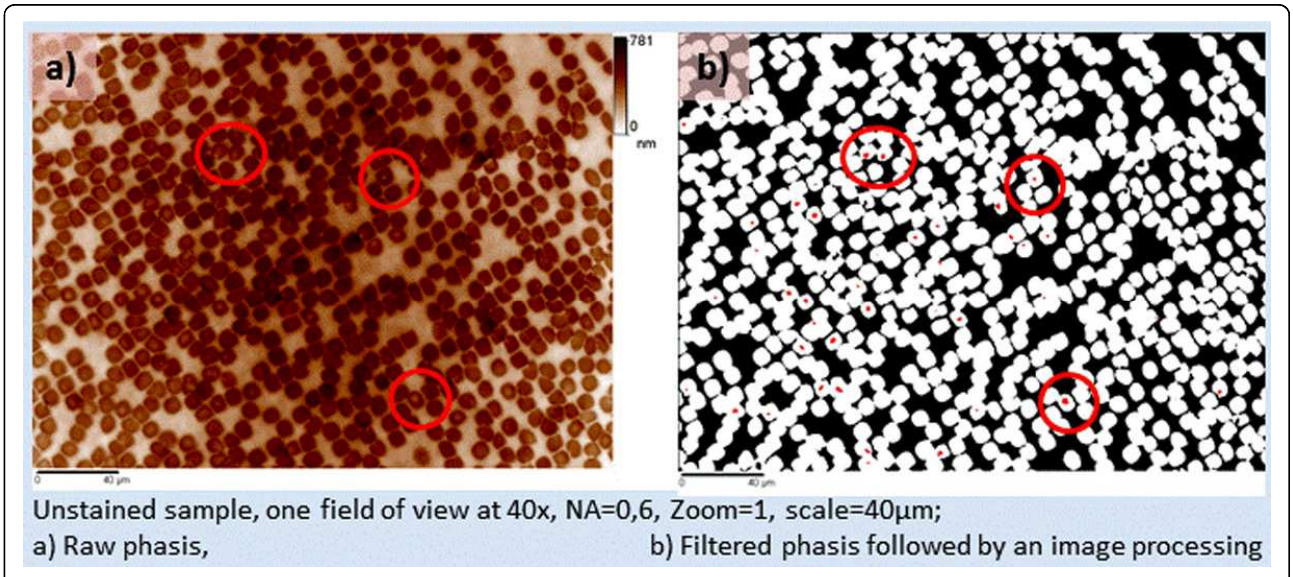


Figure 1 parasite characterization: a) microscope image, b)quantitative phase image, c) RMS spectra comparison between infected RBC and healthy RBC after hemoglobin mean spectrum subtraction

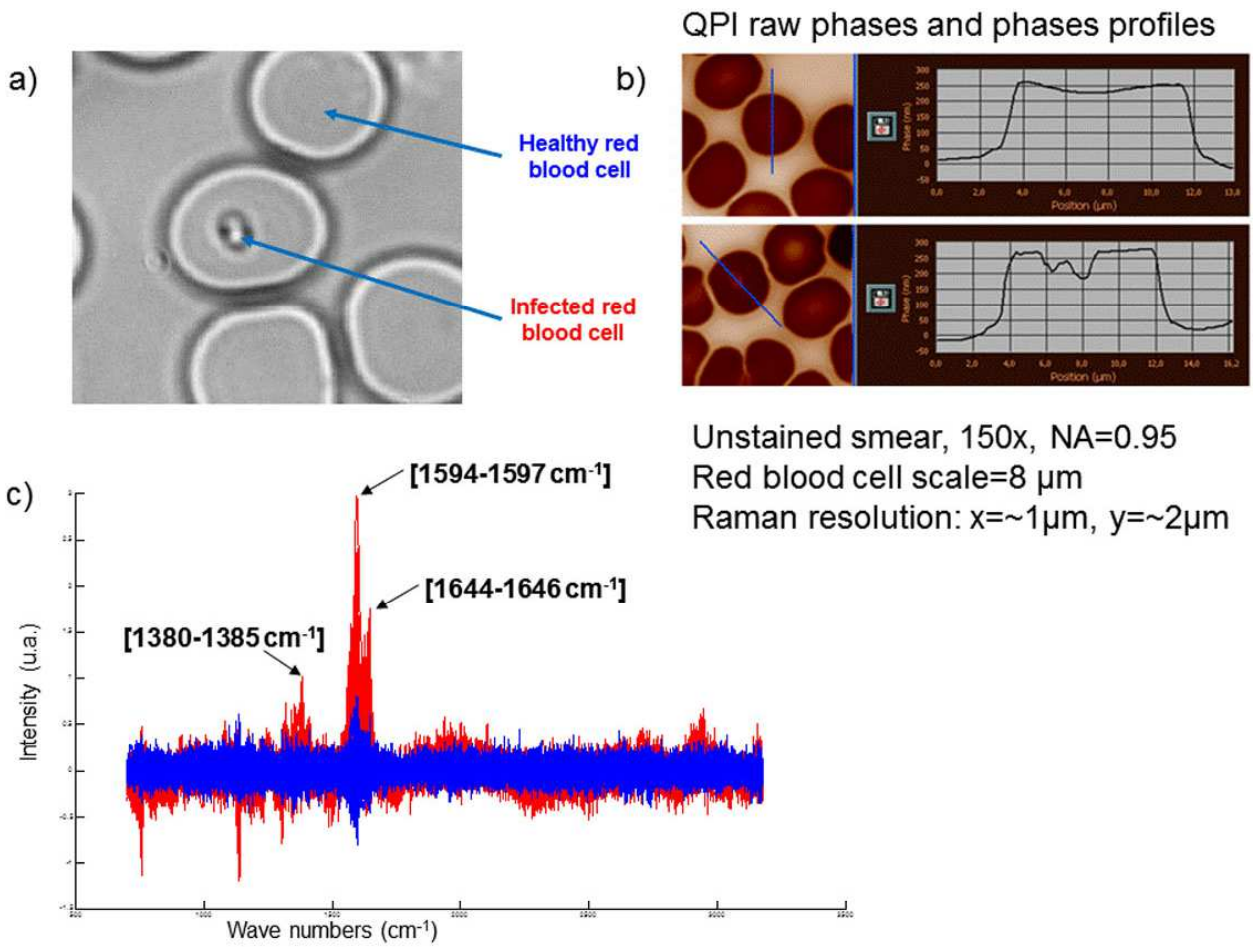


Figure 2 blood smear exploration and parasite detection with quantitative phase imaging

Results and discussion

Quantitative phase imaging results

QPI analytical process detected an average of 640 blood cells per field at 40x (standard deviation of 23) and an average of 4.1% of infected cells (standard deviation is 1.1%) for sample labeled as a 3.5% parasitemia.

Raman micro-spectroscopy results

The first classification realized considered the pre-treated spectra reduced to the representative peaks of the hemozoin biocrystal (i.e. 1380-1385, 1594-1597 and 1644-1646 cm-1). After a HCA classification the full 35 spectra obtained from sane red blood cells were classified in the first cluster of the HCA dendrogram, but only 20 spectra on the 34 registered on parasites were present in the second cluster, which gave a sensitivity of 100% and a specificity of 58.8%.

To improve these first results, a second classification was realized, considering every spectrum without the contribution of the hemoglobin signal as mentioned previously. Then, the full 35 spectra from the parasites where present in the first class of the dendrogram and the full 34 spectra from sane red blood cells were classified into the second cluster, giving a sensitivity and a specificity of 100%.

Discussion

This study proved the ability of QPI to detect RBC and infected RBC at low magnification without any previous staining. Each detected infected RBC can then be easily confirmed with RMS spectra acquisition. Making such global process easy to standardize and easy to automate.

The proposed concept is very efficient on this practical use case thanks to QPI efficiency that allows quick retrieving of a representative cell population without any previous staining. However during further studies, detection potential of poorly differentiated parasite should be compared to 100x classical immersion microscopy on stained sample.

RMS highlights thin molecular information and has been 100% sensible and specific in the current study. However, the presence of hemozoin into the parasite

could be variable in quantity or even nonexistent which in further studies could lead to spectra misclassifications. Finally, it will also be necessary to differentiate species: this should be assessed on following studies while using different high resolution (150x) techniques on infected cells: i) RMS spectra, ii) optical volume measure with QPI and iii) multi-spectral classification (see IHMO project).

Following further investigations, combination of quick detection with Quantitative Phase Imaging and selective molecular imaging with Raman Micro-Spectroscopy has a good potential to automate parasite detection and counting. Such toolkit could be applied to any kind of parasite, bacteria or bacteria inside vesicles and it also could be extended to other cytology use cases and likely to the field of histopathology.

List of abbreviations
HCA: ; QPI: Quantitative phase imaging; Px: pixel; QPI: Quantitative phase imaging; QWLSI: ; RBC: Red blood cell; RMS: Raman micro-spectroscopy

Competing interests
The authors declare that they have no competing interests.

Authors' contributions
• TH carried out the acquisitions, pretreatments, the highlighting of relevant information into the spectra, the classification of the data and drafted the extended abstract.
• VU realized the acquisitions of the Raman spectra and drafted the extended abstract.
• PB, LdL and BW developed the multi-spectral classification method and produced the multi-spectral classification.
• JK, BW and MM both designed and managed the study.
• DT selected the infected patients, provided the corresponding samples and patient's data, and did microscopy assessments.

Authors' details
[1]TRIBVN, 39, rue Louveau, 92320 Châtillon, France. [2]Phasics, Campus de l'Ecole Polytechnique, 91128 Palaiseau, France. [3]MEDyC FRE/CNRS 3481, 51096 Reims, France. [4]CHU de Reims, Laboratoire de parasitologie-mycologie, 51100 Reims, France.

Published: 30 September 2013

Reference
1. Noppadon Tangpukdee, Chatnapa Duangdee, Polrat Wilairatana, Srivicha Krudsood: **"Diagnosis: A Brief Review".** *Korean J Parasitol* 2009, **47(2)**:93-102.

doi:10.1186/1746-1596-8-S1-S42
Cite this article as: Klossa *et al.*: Quantitative phase imaging and Raman micro-spectroscopy applied to Malaria. *Diagnostic Pathology* 2013 **8**(Suppl 1):S42.

Submit your next manuscript to BioMed Central and take full advantage of:

- Convenient online submission
- Thorough peer review
- No space constraints or color figure charges
- Immediate publication on acceptance
- Inclusion in PubMed, CAS, Scopus and Google Scholar
- Research which is freely available for redistribution

Submit your manuscript at
www.biomedcentral.com/submit

BioMed Central

Chapitre 4 : Conclusions et perspectives

Au cours de ces travaux de thèse, des outils d'aide au diagnostic de cancers cutanés et de la leucémie lymphoïde chronique ont été développés. Les conclusions et perspectives relatives à ces deux axes de recherche sont présentées dans ce chapitre.

I) Aide au diagnostic de cancers cutanés par spectroscopie infrarouge

La spectroscopie infrarouge à transformée de Fourier couplée à l'algorithme de classification Fuzzy C-Means a montré de fortes capacités comme outil potentiel d'aide au diagnostic de cancers cutanés. Dans cette première étude, les améliorations apportées à l'algorithme RBA ont tout d'abord permis de diviser d'un facteur 8 le temps de calcul nécessaire à son exécution, puis, d'améliorer les résultats obtenus après classification d'images spectrales infrarouge de cancers cutanés, et enfin, d'étendre son domaine d'application à des données de nature différente. Sa performance a également été comparée à celle d'outils ayant le même objectif (*i.e.* trouver une classification FCM optimale pour chaque jeu de données), à savoir les indices de validité, les règles en *m* et les méthodes *K*-*m*. Les résultats obtenus ont montré que l'algorithme URBA obtenait des résultats de qualité au moins comparable sinon supérieure à ceux de ces outils.

Parmi les perspectives de ce travail, on peut imaginer améliorer d'avantage l'algorithme URBA en y intégrant une méthode permettant de palier le choix aléatoire des centroïdes à l'initialisation de l'algorithme FCM, ce qui permettrait d'apporter une robustesse supplémentaire à l'algorithme. Différentes stratégies sont envisageables et ont été publiées dans la littérature. Tout d'abord, on pourra envisager d'adapter des méthodes développées initialement pour les classifications supervisées en s'appuyant uniquement sur les données. Parmi ces méthodes, la méthode de Kennard-Stone choisit comme centroïdes initiaux des objets uniformément repartis dans leur espace de représentation (170), alors que la méthode D-optimale détermine les centroïdes initiaux comme étant les objets induisant le plus grand volume dans leur espace (170). Des méthodes spécifiquement développées dans le cadre de classification non-supervisée, et reposant principalement sur l'exploitation de la densité des données (171, 172), pourront également être utilisées. Enfin des algorithmes spécifiquement développés pour surmonter le problème lié à l'initialisation aléatoire des FCM pourront être testés, comme par exemple, une méthode hybride optimisant la matrice d'appartenance des FCM par essaims particulaires (173).

Il serait également intéressant dans le cas de classification d'images, de prendre en compte le voisinage géographique de chaque donnée, afin de minimiser les effets du bruit de mesure sur les partitions estimées et ainsi limiter le phénomène de pixellisation observé sur

les images reconstruites après classification. Différentes méthodes ont été développées dans la littérature. Certaines d'entre elles travaillent sur les résultats du FCM, c'est-à-dire en post-traitement. Ces méthodes utilisent des filtres 2D (filtres moyen, gaussien, …) pour lisser les images ou une régularisation par champs de Markov. Cependant, ces méthodes ne prennent pas en compte la structure des données. C'est pourquoi, nous envisageons d'utiliser des méthodes qui incorporent des contraintes spatiales au sein même des itérations de l'algorithme FCM, telles des méthodes modifiant la mesure de distance entre les objets et les centroïdes afin d'y incorporer des termes dépendants des objets voisins (174, 175), ou encore des techniques modifiant la fonction objectif à optimiser pour y incorporer ces contraintes spatiales (176).

En ce qui concerne l'algorithme URBA, de par sa structure cet algorithme impose un nombre de classes de départ maximum (fixé dans cette étude à la racine carrée du nombre d'objets +1). Or, cette initialisation peut mener à un nombre de classes de départ très grand si le nombre d'objets à classer est important. De plus, l'algorithme est plus coûteux en temps de calcul lors des premières itérations où ce nombre de classes est important. Une étude complémentaire permettant de trouver une alternative à cette estimation en déterminant un nombre de classes de départ plus pertinent serait donc une perspective intéressante à envisager.

D'autre part, dans ces travaux de thèse l'algorithme URBA est basé sur l'utilisation de l'algorithme Fuzzy C-Means. Cependant, sa structure est très facilement adaptable à des algorithmes dérivés des FCM qui sont proposés en nombre dans la littérature, comme par exemple les kernel FCM, qu'il serait intéressant de tester au sein de l'algorithme URBA.

II) Aide au diagnostic de la LLC par spectroscopie Raman

Dans cette seconde étude dont le but était d'apporter une aide au diagnostic de la leucémie lymphoïde chronique, des méthodes chimiométriques de traitement de données et l'algorithme de classification supervisée Support Vector Machine ont montré un fort potentiel discriminant lors de leur application à des données Raman acquises sur des cellules sanguines saines et tumorales. Les résultats de prédiction, obtenus après classification des spectres Raman acquis sur des cellules de différentes catégories de globules blancs (lymphocytes et polynucléaires), ont montré qu'une excellente prédiction (99% en moyenne)

pouvait être faite entre ces catégories cellulaires. D'autre part, les valeurs de sensibilité et de spécificité obtenues lors de classifications cherchant à distinguer l'état pathologique ou non de chaque échantillon atteignent 80% et 100% respectivement, confirmant le pouvoir discriminant de ces outils chimiométriques lorsqu'ils sont appliqués à des données acquises par microspectroscopie Raman. Ainsi, le module IHMO développé dans le cadre de ce projet, apparaît comme un puissant outil pouvant apporter une aide conséquente dans les laboratoires et en clinique.

En perspective de ce travail, et dans le but d'améliorer ces premiers résultats, il serait intéressant de tester d'autres méthodes proposées dans la littérature comme par exemple les « system ensembles » (110, 177) basés sur la fusion de résultats obtenus par différentes méthodes de classification.

Une autre approche de fort intérêt serait la fusion de données hétérogènes qui, dans le cas de la leucémie lymphoïde chronique, permettrait d'intégrer au processus de classification les données patients, *i.e.* les informations du dossier médical, ou des données acquises par d'autres modalités (imagerie multispectrale, MALDI-TOF, etc.), ce qui apporterait un ensemble de variables supplémentaires pouvant améliorer les prédictions. Dans la littérature, différentes méthodes sont proposées comme la fusion de résultats SVM de différents types, « early », « intermediate » ou « late » (178), la somme pondérée de plusieurs noyaux SVM (179), ou encore des méthodes multi-blocs ou multi-tableaux (180).

Des études récentes (181, 182) ont montré le fort potentiel discriminatoire de fonctions noyaux SVM spécifiquement développées pour la spectroscopie. Il serait donc intéressant de tester l'apport de ces noyaux pour la classification de spectres Raman acquis sur différents états physiopathologiques de cellules/tissus.

Au cours de ce travail, une étape de sélection de nombres d'onde a été réalisée dans le but d'améliorer les résultats de classification entre lymphocytes sains et leucémiques. Une interprétation biologique de ces nombres d'ondes discriminants (on parle alors d'attribution et d'interprétation de bandes) serait une perspective intéressante dès lors que des mécanismes biologiques liés au développement de la maladie pourraient être mis en lumière ou mieux compris. En effet, une telle analyse pourrait conforter la pertinence de certains facteurs pronostiques encore à l'étude et en mettre de nouveaux en exergue.

De manière générale, ces travaux de thèse ont permis de mettre en avant le potentiel de la microspectroscopie vibrationnelle comme outil d'aide au diagnostic de différents types de cancers, lorsqu'elle est couplée à des analyses numériques multivariées, que ce soit à l'échelle tissulaire ou cellulaire. Des bases de données conséquentes, en comparaison du nombre de patients/échantillons considérés dans la littérature, ont pu être enregistrées et ont permis de mettre en place des outils efficaces et pertinents, de par leur simplicité d'utilisation et leur rapidité, pour réaliser des discriminations sur les échantillons en se basant sur les signatures moléculaires propres à ces derniers.

Dans la continuité de ces travaux, il serait intéressant de s'impliquer dans une approche multimodale et multi-échelle (cellulaire/tissulaire) afin de pouvoir étudier un comportement cellulaire dans son environnement tissulaire, comme par exemple la réponse de la cellule tumorale à un médicament sachant que cette réponse est modulée par son environnement.

Cependant, d'autres pathologies, autres que cancéreuses, pourraient également être étudiées *via* les technologies employées dans ces travaux de thèse, et ceci qu'elles soient de type cellulaires ou infectieuses, comme ont pu le prouver les résultats obtenus lors de l'étude prospective cherchant à diagnostiquer la malaria.

La mise à disposition d'informations biologiques nouvelles vis à vis de celles obtenues par les techniques employées au quotidien dans le domaine médical, autorise donc à penser que de nouveaux outils très performants d'aide au diagnostic de maladies pourraient, à terme, considérablement aider les spécialistes à se prononcer sur un diagnostic (voire pronostic) patient. De plus, la précision de ces matériels, leur rendement et leur faible coût d'investissement en comparaison des méthodes communément employées font que la spectroscopie vibrationnelle sera certainement une technologie implantée et utilisée au quotidien en laboratoire dans un futur proche.

Bibliographie

1. Gagnon V. Étude des interactions entre les nerfs sensoriels et les follicules pileux dans un modèle in vitro de peau reconstruite par génie tissulaire. 2005.

2. Ly E. Caractérisations des lésions cutanées par microspectroscopies optiques vibrationnelles: applications ex vivo et in vivo 2009.

3. Tfayli A. Caractérisation structurale et moléculaire de la peau par microspectroscopies optiques vibrationnelles. Applications au diagnostic précoce des tumeurs cutanées et à l'étude de la diffusion de principes actifs à visée thérapeutique 2007.

4. Wysocki AB. Skin anatomy, physiology, and pathophysiology. The Nursing Clinics of North America. 1999;34(4):777.

5. Basset N, Renaud-Vilmer C. Carcinomes épidermoïdes. Manifestations dermatologiques des maladies du système hématopoïétique et oncologie dermatologique: Springer; 2008. p. 137-46.

6. Lourari S, Paul C, Meyer N. Traitements non chirurgicaux des carcinomes cutanés et de leurs précurseurs. La Presse Médicale. 2011;40(7):690-6.

7. Mélissopoulos A, Levacher C, Robert L. La peau: structure et physiologie: Tec & Doc Lavoisier Paris; 1998.

8. HAS. Haute Autorité de Santé. Available from: http://www.has-sante.fr/.

9. MASS. Ministère des affaires sociales et de la santé 2012. Available from: http://www.sante.gouv.fr/mieux-prevenir-et-detecter-tot-les-cancers-de-la-peau.html.

10. SFD. Société Française de Dermatologie France. Available from: http://www.sfdermato.org/.

11. Martin L, Bonerandi J-J, editors. Carcinome épidermoïde cutané (carcinome spinocellulaire) recommandations de pratique clinique pour la prise en charge diagnostique et thérapeutique. Annales de Dermatologie et de Venereologie; 2009: Elsevier Masson.

12. SFD. Carcinome basocellulaire. Recommandations pour la pratique clinique (Anaes 2004). 2004.

13. HAS. Tumeur maligne, affection maligne du tissu lymphatique ou hématopoïétique. Mélanome Cutané. Guide - affection de longue durée. 2012.

14. Saiag P, Grob J, Grosshans E. Tumeurs cutanées épithéliales et mélaniques. Ann Dermatol Venereol. 2002;129:2S143-2S8.

15. HAS. Stratégie de diagnostic précoce du mélanome. Recommandation en santé publique (Rapport d'évaluation). 2006.

16. HAS. La prise en charge du mélanome cutané. Guide - Affectation de longue durée. 2010.

17. Lavie A, Desouches C, Casanova D, Bardot J, Grob J-J, Legré R, et al., editors. Mise au point sur la prise en charge chirurgicale du mélanome malin cutané. Revue de la littérature. Annales de chirurgie plastique esthetique; 2007: Elsevier.

18. Staub G, Revol M, May P, Bayol J-C, Verola O, Servant J-M, editors. Marges d'exérèse chirurgicale et taux de récidive des carcinomes cutanés: Étude prospective de 844 cas. Annales de chirurgie plastique esthetique; 2008: Elsevier.

19. SOR. Standards, Options et Recommandations. Mélanome. France2005. Available from: www.fnclcc.fr.

20. HAS. Conditions de réalisation des actes d'exérèse de lésion superficielle de la peau. 2007.

21. Rallan D, Harland C. Skin imaging: is it clinically useful? Clinical and experimental dermatology. 2004;29(5):453-9.

22. Mogensen M, Jemec GB. Diagnosis of nonmelanoma skin cancer/keratinocyte carcinoma: a review of diagnostic accuracy of nonmelanoma skin cancer diagnostic tests and technologies. Dermatologic Surgery. 2007;33(10):1158-74.

23. Ghassemi P, Lemaillet P, Germer TA, Shupp JW, Venna SS, Boisvert ME, et al. Out-of-plane Stokes imaging polarimeter for early skin cancer diagnosis. Journal of biomedical optics. 2012;17(7):0760141-9.

24. HAS. Dermatoscopie [dermoscopie] pour surveillance cutanée. 2007.

25. Janda M, Loescher LJ, Soyer HP. Enhanced Skin Self-examination: A Novel Approach to Skin Cancer Monitoring and Follow-up. Jama Dermatology. 2013;149(2):231-6.

26. Gufler H, Franke FE, Rau WS. High-resolution MRI of basal cell carcinomas of the face using a microscopy coil. American Journal of Roentgenology. 2007;188(5):W480-W4.

27. Alex A, Považay B, Hofer B, Popov S, Glittenberg C, Binder S, et al. Multispectral in vivo three-dimensional optical coherence tomography of human skin. Journal of biomedical optics. 2010;15(2):026025--15.

28. Levy J-L. Evaluations et mesures en dermatologie: nouvelles approches. Dermatologie. 2011;289(13):759-63.

29. Pavlova P, Borisova E, Avramov L, Petkova E, Troyanova P. Investigation of Relations Between Skin Cancer Lesions 'Images and Their Reflectance and Fluorescent Spectra. 2011.

30. Gerger A, Hofmann-Wellenhof R, Samonigg H, Smolle J. In vivo confocal laser scanning microscopy in the diagnosis of melanocytic skin tumours. British Journal of Dermatology. 2009;160(3):475-81.

31. Lassau N, Chami L, Peronneau P. Imagerie des mélanomes: performances de l'échographie sans et avec injection de produit de contraste pour le diagnostic et l'évaluation des traitements. Bulletin du cancer. 2007;94(1):93-8.

32. Machet L, Ossant F, Bleuzen A, Grégoire J-M, Machet M-C, Vaillant L. L'échographie cutanée haute résolution: utilité pour le diagnostic, le traitement et la surveillance des maladies dermatologiques. Journal de Radiologie. 2006;87(12):1946-61.

33. Lee K-S, Zhao H, Ibrahim SF, Meemon N, Khoudeir L, Rolland JP. Three-dimensional imaging of normal skin and nonmelanoma skin cancer with cellular resolution using Gabor domain optical coherence microscopy. Journal of biomedical optics. 2012;17(12):126006-.

34. Hofmann-Wellenhof R, Pellacani G, Malvehy J, Soyer HP. Reflectance confocal microscopy for skin diseases: Springer; 2012.

35. Joseph CS, Yaroslavsky AN, Al-Arashi M, Goyette TM, Dickinson JC, Gatesman AJ, et al., editors. Terahertz spectroscopy of intrinsic biomarkers for non-melanoma skin cancer. SPIE OPTO: Integrated Optoelectronic Devices; 2009: International Society for Optics and Photonics.

36. Wilmink GJ, Ibey BL, Tongue T, Schulkin B, Peralta X, Rivest BD, et al., editors. Measurement of the optical properties of skin using terahertz time-domain spectroscopic techniques. BiOS; 2010: International Society for Optics and Photonics.

37. Perry SW, Burke RM, Brown EB. Two-photon and second harmonic microscopy in clinical and translational cancer research. Annals of biomedical engineering. 2012;40(2):277-91.

38. Mehta AB, Hoffbrand AV. Hématologie: De Boeck Supérieur; 2003.

39. Encyclopédie. Larousse médical. Available from: http://www.larousse.fr/encyclopedie/divers/sang/89933.

40. Jentsch-Ullrich K, Koenigsmann M, Mohren M, Franke A. Lymphocyte subsets' reference ranges in an age-and gender-balanced population of 100 healthy adults--a monocentric German study. Clinical immunology (Orlando, Fla). 2005;116(2):192.

41. EFS. Établissement Français du Sang. Available from: www.dondusang.net.

42. ARC. Fondation ARC pour la recherche sur le cancer. 2012.

43. InVS. Institut de Veille Sanitaire 2011. Available from: www.invs.sante.fr.

44. HAS. Tumeur maligne, affection maligne du tissu lymphatique ou hématopoïétique. Leucémie lymphoïde chronique. Guide - Affection de longue durée. 2011.

45. HAS. La prise en charge de la leucémie lymphoïde chronique. Guide patient - Affection de longue durée. 2011.

46. Troussard X. Diagnostic, pronostic et traitement chez les patients avec une leucémie lymphoïde chronique. Immuno-analyse & Biologie Spécialisée. 2007;22(5):313-8.

47. Troussard X, Malet M, Cornet E. Phase précoce de la leucémie lymphoïde chronique (LLC) et concept de lymphocytose B monoclonale (MBL). Revue Francophone des Laboratoires. 2011;2011(433):47-53.

48. Aurrant T, Callet-Bauchu E, Cymbalista F, Delmer A, Dreyfus B, Nguyen Khac F, et al. Recommandations 2012 de la SFH pour le diagnostic, le traitement et le suivi de la leucémie lymphoïde chronique. Hématologie. 2013;19:4-9.

49. Cymbalista F. LLC: physiopathologie, diagnostic et approche thérapeutique. Revue Francophone des Laboratoires. 2013;2013(452):61-71.

50. Matutes E, Owusu-Ankomah K, Morilla R, Garcia MJ, Houlihan A, Que T, et al. The immunological profile of B-cell disorders and proposal of a scoring system for the diagnosis

of CLL. Leukemia: official journal of the Leukemia Society of America, Leukemia Research Fund, UK. 1994;8(10):1640.

51. Fissore-Magdelein C, Riss I, Dahmani B, Malval R, Taillan B. Une hyperlymphocytose chez le sujet âgé n'est pas toujours une leucémie lymphoïde chronique (LLC). Revue Francophone des Laboratoires. 2009;2009(410):71-2.

52. SFH. Société Française d'Hématologie - Recommandations patients. Available from: sfh.hematologie.net.

53. Rai KR, Sawitsky A, Cronkite EP, Chanana AD, Levy RN, Pasternack B. Clinical staging of chronic lymphocytic leukemia. Blood. 1975;46(2):219-34.

54. Binet J, Auquier A, Dighiero G, Chastang C, Piguet H, Goasguen J, et al. A new prognostic classification of chronic lymphocytic leukemia derived from a multivariate survival analysis. Cancer. 1981;48(1):198-206.

55. Leblanc R-M. Quelle place pour les marqueurs pronostiques de la leucémie lymphoïde chronique en 2010? Option/Bio. 2010;21(432):16-7.

56. Troussard X. Facteurs pronostiques utiles dans l'évaluation personnalisée d'un patient avec une leucémie lymphoïde chronique (LLC). Immuno-analyse & Biologie Spécialisée. 2009;24(1):1-8.

57. Eclache V, Baran-Marszak F. Intérêt de la cytogénétique et de la biologie moléculaire pour le diagnostic des syndromes lymphoprolifératifs B. Revue Francophone des Laboratoires. 2013;2013(452):49-59.

58. Laffitte A, Ghiringhelli C-B, Roger-Schmeltz J, Rivière E, Hennequin N, Lippa N, et al. Leucémie lymphoïde chronique, alemtuzumab et risque d'anémie hémolytique. La Revue de médecine interne. 2008;29:S409-S10.

59. Dalibart M, Servant S. Spectroscopie dans l'infrarouge, techniques de l'ingénieur. Citée aux pages: 77, 78 et. 2000;80.

60. Barbillat J, Bougeard D, Buntinx G, Delhaye M, Dhamelincourt P, Fillaux F. Spectrométrie raman. Techniques de l'ingénieur Analyse et caractérisation. 1999;9(P2865):1-31.

61. Kendall C, Isabelle M, Bazant-Hegemark F, Hutchings J, Orr L, Babrah J, et al. Vibrational spectroscopy: a clinical tool for cancer diagnostics. Analyst. 2009;134(6):1029-45.

62. Diem M, Romeo M, Boydston-White S, Miljković M, Matthäus C. A decade of vibrational micro-spectroscopy of human cells and tissue (1994–2004). Analyst. 2004;129(10):880-5.

63. Lui H, Zhao J, McLean D, Zeng H. Real-time Raman spectroscopy for in vivo skin cancer diagnosis. Cancer research. 2012;72(10):2491-500.

64. Seddon AB. Mid-infrared (IR) A hot topic: The potential for using mid-IR light for non-invasive early detection of skin cancer in vivo. physica status solidi (b). 2013:1020-7.

65. Silveira FL, Pacheco MTT, Bodanese B, Zângaro RA, Silveira L, editors. Discriminating model for skin cancer diagnosis in vivo through Raman spectroscopy. SPIE BiOS; 2013: International Society for Optics and Photonics.

66. Mohs AM, Mancini MC, Singhal S, Provenzale JM, Leyland-Jones B, Wang MD, et al. Hand-held spectroscopic device for in vivo and intraoperative tumor detection: contrast enhancement, detection sensitivity, and tissue penetration. Analytical chemistry. 2010;82(21):9058-65.

67. Schultz CP, Liu K-z, Johnston JB, Mantsch HH. Study of chronic lymphocytic leukemia cells by FT-IR spectroscopy and cluster analysis. Leukemia research. 1996;20(8):649-55.

68. Nguyen CT, Nguyen JT, Rutledge S, Zhang J, Wang C, Walker GC. Detection of chronic lymphocytic leukemia cell surface markers using surface enhanced Raman scattering gold nanoparticles. Cancer letters. 2010;292(1):91-7.

69. Chan JW, Taylor DS, Lane SM, Zwerdling T, Tuscano J, Huser T. Non-destructive Identification of Individual Leukemia Cells by Optical Trapping Raman Spectroscopy. Analytical Chemistry, vol 80, no 6, February 9, 2008, pp 2180-2187. 2008;80(UCRL-JRNL-229152).

70. Chan JW, Taylor DS, Zwerdling T, Lane SM, Ihara K, Huser T. Micro-Raman spectroscopy detects individual neoplastic and normal hematopoietic cells. Biophysical journal. 2006;90(2):648-56.

71. Benedetti E, Bramanti E, Papineschi F, Rossi I, Benedetti E. Determination of the relative amount of nucleic acids and proteins in leukemic and normal lymphocytes by means of Fourier transform infrared microspectroscopy. Applied spectroscopy. 1997;51(6):792-7.

72. Lasch P. Spectral pre-processing for biomedical vibrational spectroscopy and microspectroscopic imaging. Chemometrics and Intelligent Laboratory Systems. 2012;117:100-14.

73. Heraud P, Wood BR, Beardall J, McNaughton D. Effects of pre-processing of Raman spectra on in vivo classification of nutrient status of microalgal cells. Journal of chemometrics. 2006;20(5):193-7.

74. James TM, Schlösser M, Lewis RJ, Fischer S, Bornschein B, Telle HH. Automated Quantitative Spectroscopic Analysis Combining Background Subtraction, Cosmic Ray Removal, and Peak Fitting. Applied spectroscopy. 2013;67(8):949-59.

75. Wold S, Esbensen K, Geladi P. Principal component analysis. Chemometrics and Intelligent Laboratory Systems. 1987;2(1):37-52.

76. Martens H, Nielsen JP, Engelsen SB. Light scattering and light absorbance separated by extended multiplicative signal correction. Application to near-infrared transmission analysis of powder mixtures. Analytical chemistry. 2003;75(3):394-404.

77. Cappel UB, Bell IM, Pickard LK. Removing cosmic ray features from Raman map data by a refined nearest neighbor comparison method as a precursor for chemometric analysis. Applied spectroscopy. 2010;64(2):195-200.

78. Li S, Dai L. An improved algorithm to remove cosmic spikes in Raman spectra for online monitoring. Applied spectroscopy. 2011;65(11):1300-6.

79. Schulze HG, Turner RF. A Fast, Automated, Polynomial-Based Cosmic Ray Spike-Removal Method for the High-Throughput Processing of Raman Spectra. Applied spectroscopy. 2013;67(4):457-62.

80. Savitzky A, Golay MJ. Smoothing and differentiation of data by simplified least squares procedures. Analytical chemistry. 1964;36(8):1627-39.

81. Cao A, Pandya AK, Serhatkulu GK, Weber RE, Dai H, Thakur JS, et al. A robust method for automated background subtraction of tissue fluorescence. Journal of Raman Spectroscopy. 2007;38(9):1199-205.

82. Lieber CA, Mahadevan-Jansen A. Automated method for subtraction of fluorescence from biological Raman spectra. Applied spectroscopy. 2003;57(11):1363-7.

83. Mazet V, Carteret C, Brie D, Idier J, Humbert B. Background removal from spectra by designing and minimising a non-quadratic cost function. Chemometrics and Intelligent Laboratory Systems. 2005;76(2):121-33.

84. Zhao J, Lui H, McLean DI, Zeng H. Automated autofluorescence background subtraction algorithm for biomedical Raman spectroscopy. Applied spectroscopy. 2007;61(11):1225-32.

85. Hu Y, Jiang T, Shen A, Li W, Wang X, Hu J. A background elimination method based on wavelet transform for Raman spectra. Chemometrics and Intelligent Laboratory Systems. 2007;85(1):94-101.

86. Zhang ZM, Chen S, Liang YZ, Liu ZX, Zhang QM, Ding LX, et al. An intelligent background-correction algorithm for highly fluorescent samples in Raman spectroscopy. Journal of Raman Spectroscopy. 2010;41(6):659-69.

87. Liland KH, Almøy T, Mevik B-H. Optimal choice of baseline correction for multivariate calibration of spectra. Applied spectroscopy. 2010;64(9):1007-16.

88. Baek S-J, Park A, Kim J, Shen A, Hu J. A simple background elimination method for Raman spectra. Chemometrics and Intelligent Laboratory Systems. 2009;98(1):24-30.

89. Zhang Z-M, Chen S, Liang Y-Z. Baseline correction using adaptive iteratively reweighted penalized least squares. Analyst. 2010;135(5):1138-46.

90. Barnes R, Dhanoa M, Lister SJ. Standard normal variate transformation and de-trending of near-infrared diffuse reflectance spectra. Applied spectroscopy. 1989;43(5):772-7.

91. Heinz DC. Fully constrained least squares linear spectral mixture analysis method for material quantification in hyperspectral imagery. Geoscience and Remote Sensing, IEEE Transactions on. 2001;39(3):529-45.

92. Beier BD, Berger AJ. Method for automated background subtraction from Raman spectra containing known contaminants. Analyst. 2009;134(6):1198-202.

93. Tfayli A, Gobinet C, Vrabie V, Huez R, Manfait M, Piot O. Digital dewaxing of Raman signals: discrimination between nevi and melanoma spectra obtained from paraffin-embedded skin biopsies. Applied spectroscopy. 2009;63(5):564-70.

94. Geladi P, MacDougall D, Martens H. Linearization and scatter-correction for near-infrared reflectance spectra of meat. Applied spectroscopy. 1985;39(3):491-500.

95. Kohler A, Böcker U, Warringer J, Blomberg A, Omholt S, Stark E, et al. Reducing inter-replicate variation in Fourier transform infrared spectroscopy by extended multiplicative signal correction. Applied spectroscopy. 2009;63(3):296-305.

96. Afseth NK, Kohler A. Extended multiplicative signal correction in vibrational spectroscopy, a tutorial. Chemometrics and Intelligent Laboratory Systems. 2012;117:92-9.

97. Nallala J, Piot O, Diebold MD, Gobinet C, Bouché O, Manfait M, et al. Infrared imaging as a cancer diagnostic tool: Introducing a new concept of spectral barcodes for identifying molecular changes in colon tumors. Cytometry Part A. 2013;83(3):294-300.

98. Jachtenberg JW, Schut TB, French P, Kros M, Lamfers M, Leenstra S. Raman spectroscopy can discriminate distinct glioma subtypes as defined by RNA expression profiling. Journal of Raman Spectroscopy. 2013.

99. Ly E, Piot O, Durlach A, Bernard P, Manfait M. Differential diagnosis of cutaneous carcinomas by infrared spectral micro-imaging combined with pattern recognition. Analyst. 2009;134(6):1208-14.

100. Travo A, Desplat V, Barron E, Poychicot-Coustau E, Guillon J, Déléris G, et al. Basis of a FTIR spectroscopy methodology for automated evaluation of Akt kinase inhibitor on leukemic cell lines used as model. Analytical and bioanalytical chemistry. 2012;404(6-7):1733-43.

101. Saeys Y, Inza I, Larrañaga P. A review of feature selection techniques in bioinformatics. Bioinformatics. 2007;23(19):2507-17.

102. Xiaobo Z, Jiewen Z, Povey MJ, Holmes M, Hanpin M. Variables selection methods in near-infrared spectroscopy. Analytica chimica acta. 2010;667(1):14-32.

103. Guyon I, Elisseeff A. An introduction to variable and feature selection. The Journal of Machine Learning Research. 2003;3:1157-82.

104. Nallala J, Gobinet C, Diebold M-D, Untereiner V, Bouché O, Manfait M, et al. Infrared spectral imaging as a novel approach for histopathological recognition in colon cancer diagnosis. Journal of biomedical optics. 2012;17(11):116013-.

105. Schmitt J, Beekes M, Brauer A, Udelhoven T, Lasch P, Naumann D. Identification of scrapie infection from blood serum by Fourier transform infrared spectroscopy. Analytical chemistry. 2002;74(15):3865-8.

106. Zhang X, Xu Y, Zhang Y, Wang L, Hou C, Zhou X, et al. Intraoperative detection of thyroid carcinoma by fourier transform infrared spectrometry. Journal of Surgical Research. 2011;171(2):650-6.

107. Lasch P, Schmitt J, Beekes M, Udelhoven T, Eiden M, Fabian H, et al. Antemortem identification of bovine spongiform encephalopathy from serum using infrared spectroscopy. Analytical chemistry. 2003;75(23):6673-8.

108. Duraipandian S, Zheng W, Ng J, Low JJ, Ilancheran A, Huang Z. In vivo diagnosis of cervical precancer using Raman spectroscopy and genetic algorithm techniques. Analyst. 2011;136(20):4328-36.

109. Song L-MWK, Molckovsky A, Wang KK, Burgart LJ, Dolenko B, Somorjai RL, et al., editors. Diagnostic potential of Raman spectroscopy in Barrett's esophagus. Proceeding of SPIE; 2005.

110. Trevisan J, Angelov PP, Carmichael PL, Scott AD, Martin FL. Extracting biological information with computational analysis of Fourier-transform infrared (FTIR) biospectroscopy datasets: current practices to future perspectives. Analyst. 2012;137(14):3202-15.

111. Duby C, Robin S. Analyse en composantes principales. Institut National Agronomique, Paris-Grignon. 2006;80.

112. Hyvärinen A, Karhunen J, Oja E. What is Independent Component Analysis? Independent Component Analysis. 2001:145-64.

113. Lee DD, Seung HS. Learning the parts of objects by non-negative matrix factorization. Nature. 1999;401(6755):788-91.

114. Mallat SG. A theory for multiresolution signal decomposition: the wavelet representation. Pattern Analysis and Machine Intelligence, IEEE Transactions on. 1989;11(7):674-93.

115. Fisher R. Linear discriminant analysis. Ann Eugenics. 1936;7:179-88.

116. Kruger U, Xie L. Partial Least Squares. Statistical Monitoring of Complex Multivariate Processes: With Applications in Industrial Process Control. 2012:375-409.

117. Talukder A, Casasent D. General methodology for simultaneous representation and discrimination of multiple object classes. Optical Engineering. 1998;37(3):904-13.

118. Ostrovsky E, Zelig U, Guskova I, Ariad S, Mordechai S, Nisky I, et al. Detection of Cancer Using Advanced Computerized Analysis of Infrared Spectra of Peripheral Blood. IEEE Transactions on Biomedical Engineering. 2013;60 (2):343-53.

119. Molckovsky A, Song L-MWK, Shim MG, Marcon NE, Wilson BC. Diagnostic potential of near-infrared Raman spectroscopy in the colon: differentiating adenomatous from hyperplastic polyps. Gastrointestinal endoscopy. 2003;57(3):396-402.

120. Argov S, Ramesh J, Salman A, Sinelnikov I, Goldstein J, Guterman H, et al. Diagnostic potential of Fourier-transform infrared microspectroscopy and advanced computational methods in colon cancer patients. Journal of biomedical optics. 2002;7(2):248-54.

121. Majumder SK, Gebhart S, Johnson MD, Thompson R, Lin W-C, Mahadevan-Jansen A. A probability-based spectroscopic diagnostic algorithm for simultaneous discrimination of brain tumor and tumor margins from normal brain tissue. Applied spectroscopy. 2007;61(5):548-57.

122. Vargis E, Kanter EM, Majumder SK, Keller MD, Beaven RB, Rao GG, et al. Effect of normal variations on disease classification of Raman spectra from cervical tissue. Analyst. 2011;136(14):2981-7.

123. Ma Z, Tavares JMR, Jorge RN, Mascarenhas T. A review of algorithms for medical image segmentation and their applications to the female pelvic cavity. Computer Methods in Biomechanics and Biomedical Engineering. 2010;13(2):235-46.

124. Ward Jr JH. Hierarchical grouping to optimize an objective function. Journal of the American statistical association. 1963;58(301):236-44.

125. Roux M. Algorithmes de classification. 1985.

126. MacQueen J, editor Some methods for classification and analysis of multivariate observations. Proceedings of the fifth Berkeley symposium on mathematical statistics and probability; 1967: California, USA.

127. Bezdek JC. Pattern recognition with fuzzy objective function algorithms: Kluwer Academic Publishers; 1981.

128. Tosi G, Conti C, Giorgini E, Ferraris P, Garavaglia MG, Sabbatini S, et al. FTIR microspectroscopy of melanocytic skin lesions: a preliminary study. Analyst. 2010;135(12):3213-9.

129. Hughes C, Iqbal-Wahid J, Brown M, Shanks JH, Eustace A, Denley H, et al. FTIR microspectroscopy of selected rare diverse sub-variants of carcinoma of the urinary bladder. Journal of biophotonics. 2013;6(1):73-87.

130. Kong K, Rowlands CJ, Varma S, Perkins W, Leach IH, Koloydenko AA, et al. Diagnosis of tumors during tissue-conserving surgery with integrated autofluorescence and Raman scattering microscopy. Proceedings of the National Academy of Sciences. 2013;110(38):15189-94.

131. Wang XY, Garibaldi J, Ozen T, editors. Application of the fuzzy C-means clustering method on the analysis of non pre-processed FTIR data for cancer diagnosis. Internat Conf on Australian and New Zealand Intelligent Information Systems (ANZIIS); 2003.

132. Krafft C, Diderhoshan MA, Recknagel P, Miljkovic M, Bauer M, Popp J. Crisp and soft multivariate methods visualize individual cell nuclei in Raman images of liver tissue sections. Vibrational Spectroscopy. 2011;55(1):90-100.

133. Mallot HA. Artificial Neural Networks. Computational Neuroscience: Springer; 2013. p. 83-112.

134. Cover T, Hart P. Nearest neighbor pattern classification. Information Theory, IEEE Transactions on. 1967;13(1):21-7.

135. Vapnik V. The nature of statistical learning theory: springer; 1999.

136. Cortes C, Vapnik V. Support-vector networks. Machine learning. 1995;20(3):273-97.

137. Gniadecka M, Philipsen PA, Sigurdsson S, Wessel S, Nielsen OF, Christensen DH, et al. Melanoma diagnosis by Raman spectroscopy and neural networks: structure alterations in proteins and lipids in intact cancer tissue. Journal of Investigative Dermatology. 2004;122(2):443-9.

138. Kim SB, Temiyasathit C, Bensalah K, Tuncel A, Cadeddu J, Kabbani W, et al. An effective classification procedure for diagnosis of prostate cancer in near infrared spectra. Expert Systems with Applications. 2010;37(5):3863-9.

139. Lin K, Cheng DLP, Huang Z. Optical diagnosis of laryngeal cancer using high wavenumber Raman spectroscopy. Biosensors and Bioelectronics. 2012;35(1):213-7.

140. Li Q, Wang W, Ling X, Wu JG. Detection of Gastric Cancer with Fourier Transform Infrared Spectroscopy and Support Vector Machine Classification. BioMed research international. 2013;2013.

141. Bergner N, Bocklitz T, Romeike BF, Reichart R, Kalff R, Krafft C, et al. Identification of primary tumors of brain metastases by Raman imaging and support vector machines. Chemometrics and Intelligent Laboratory Systems. 2012;117:224-32.

142. Nachabé R, Evers DJ, Hendriks BH, Lucassen GW, van der Voort M, Rutgers EJ, et al. Diagnosis of breast cancer using diffuse optical spectroscopy from 500 to 1600 nm: comparison of classification methods. Journal of biomedical optics. 2011;16(8):087010--12.

143. Ly E, Cardot-Leccia N, Ortonne JP, Benchetrit M, Michiels JF, Manfait M, et al. Histopathological characterization of primary cutaneous melanoma using infrared microimaging: a proof-of-concept study. British Journal of Dermatology. 2010;162(6):1316-23.

144. Sebiskveradze D, Gobinet C, Vrabie V, Jeannesson P, Piot O, Manfait M. Highlighting peritumoral areas in human skin cancer biopsies by infrared micro-spectroscopy. Diagnostic Pathology. 2013;8(Suppl 1):S33.

145. Sebiskveradze D, Vrabie V, Gobinet C, Durlach A, Bernard P, Ly E, et al. Automation of an algorithm based on fuzzy clustering for analyzing tumoral heterogeneity in human skin carcinoma tissue sections. Laboratory Investigation. 2011;91(5):799-811.

146. Kohler A, Kirschner C, Oust A, Martens H. Extended multiplicative signal correction as a tool for separation and characterization of physical and chemical information in Fourier transform infrared microscopy images of cryo-sections of beef loin. Applied spectroscopy. 2005;59(6):707-16.

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

148. Wolthuis R, Travo A, Nicolet Cl, Neuville As, Gaub M-P, Guenot D, et al. IR spectral imaging for histopathological characterization of xenografted human colon carcinomas. Analytical chemistry. 2008;80(22):8461-9.

149. Chen W, Giger ML, Bick U. A fuzzy c-means (FCM)-based approach for computerized segmentation of breast lesions in dynamic contrast-enhanced MR images. Academic radiology. 2006;13(1):63.

150. Marec A, Thomas J-H, El Guerjouma R. Damage characterization of polymer-based composite materials: Multivariable analysis and wavelet transform for clustering acoustic emission data. Mechanical Systems and Signal Processing. 2008;22(6):1441-64.

151. Parthasarathy R, Thiagarajan G, Yao X, Wang Y-P, Spencer P, Wang Y. Application of multivariate spectral analyses in micro-Raman imaging to unveil structural/chemical features of the adhesive/dentin interface. Journal of biomedical optics. 2008;13(1):014020--9.

152. Wang X-Y, Garibaldi JM, Bird B, George MW. A novel fuzzy clustering algorithm for the analysis of axillary lymph node tissue sections. Applied Intelligence. 2007;27(3):237-48.

153. Tan KS, Lim WH, Isa NAM. Novel initialization scheme for Fuzzy C-Means algorithm on color image segmentation. Applied Soft Computing. 2013;13(4):1832-52.

154. Benaichouche A, Oulhadj H, Siarry P. Improved spatial fuzzy c-means clustering for image segmentation using PSO initialization, Mahalanobis distance and post-segmentation correction. Digital Signal Processing. 2013.

155. Chaira T. A novel intuitionistic fuzzy C means clustering algorithm and its application to medical images. Applied Soft Computing. 2011;11(2):1711-7.

156. Genitha CH, Vani K, editors. Classification of satellite images using new Fuzzy cluster centroid for unsupervised classification algorithm. Information & Communication Technologies (ICT), 2013 IEEE Conference on; 2013: IEEE.

157. Lasch P, Haensch W, Naumann D, Diem M. Imaging of colorectal adenocarcinoma using FT-IR microspectroscopy and cluster analysis. Biochimica et Biophysica Acta (BBA)-Molecular Basis of Disease. 2004;1688(2):176-86.

158. Wang W, Zhang Y. On fuzzy cluster validity indices. Fuzzy sets and systems. 2007;158(19):2095-117.

159. Dembele D, Kastner P. Fuzzy C-means method for clustering microarray data. Bioinformatics. 2003;19(8):973-80.

160. Choe H, Jordan JB, editors. On the optimal choice of parameters in a fuzzy c-means algorithm. Fuzzy Systems, 1992, IEEE International Conference on; 1992: IEEE.

161. Okeke F, Karnieli A. Linear mixture model approach for selecting fuzzy exponent value in fuzzy< i> c</i>-means algorithm. Ecological Informatics. 2006;1(1):117-24.

162. Schwämmle V, Jensen ON. A simple and fast method to determine the parameters for fuzzy c–means cluster analysis. Bioinformatics. 2010;26(22):2841-8.

163. Futschik ME, Carlisle B. Noise-robust soft clustering of gene expression time-course data. Journal of bioinformatics and computational biology. 2005;3(04):965-88.

164. Burges CJ. A tutorial on support vector machines for pattern recognition. Data mining and knowledge discovery. 1998;2(2):121-67.

165. Joachims T. Text categorization with support vector machines: Learning with many relevant features: Springer; 1998.

166. Sattlecker M, Bessant C, Smith J, Stone N. Investigation of support vector machines and Raman spectroscopy for lymph node diagnostics. Analyst. 2010;135(5):895-901.

167. Widjaja E, Zheng W, Huang Z. Classification of colonic tissues using near-infrared Raman spectroscopy and support vector machines. International journal of oncology. 2008;32(3):653-62.

168. Yan B, Li Y, Yang G, Wen Z-n, Li M-l, Li L-j. Discrimination of parotid neoplasms from the normal parotid gland by use of Raman spectroscopy and support vector machine. Oral oncology. 2011;47(5):430-5.

169. Cover TM. Geometrical and statistical properties of systems of linear inequalities with applications in pattern recognition. Electronic Computers, IEEE Transactions on. 1965 (3):326-34.

170. Wu W, Walczak B, Massart D, Heuerding S, Erni F, Last I, et al. Artificial neural networks in classification of NIR spectral data: design of the training set. Chemometrics and Intelligent Laboratory Systems. 1996;33(1):35-46.

171. Redmond SJ, Heneghan C. A method for initialising the K-means clustering algorithm using kd-trees. Pattern Recognition Letters. 2007;28(8):965-73.

172. Al-Daoud MdB, Roberts SA. New methods for the initialisation of clusters. Pattern Recognition Letters. 1996;17(5):451-5.

173. Izakian H, Abraham A. Fuzzy C-means and fuzzy swarm for fuzzy clustering problem. Expert Systems with Applications. 2011;38(3):1835-8.

174. Liew A, Leung S, Lau W. Fuzzy image clustering incorporating spatial continuity. IEE Proceedings-Vision, Image and Signal Processing. 2000;147(2):185-92.

175. Wang J, Kong J, Lu Y, Qi M, Zhang B. A modified FCM algorithm for MRI brain image segmentation using both local and non-local spatial constraints. Computerized Medical Imaging and Graphics. 2008;32(8):685-98.

176. Hou Z, Qian W, Huang S, Hu Q, Nowinski WL. Regularized fuzzy c-means method for brain tissue clustering. Pattern Recognition Letters. 2007;28(13):1788-94.

177. Gajjar K, Trevisan J, Owens G, Keating PJ, Wood NJ, Stringfellow HF, et al. Fourier-transform infrared spectroscopy coupled with a classification machine for the analysis of blood plasma or serum: a novel diagnostic approach for ovarian cancer. Analyst. 2013.

178. Pavlidis P, Weston J, Cai J, Grundy WN, editors. Gene functional classification from heterogeneous data. Proceedings of the fifth annual international conference on Computational biology; 2001: ACM.

179. Lanckriet GR, Deng M, Cristianini N, Jordan MI, Noble WS, editors. Kernel-based data fusion and its application to protein function prediction in yeast. Pacific symposium on biocomputing; 2004.

180. Hassani S, Martens H, Qannari EM, Hanafi M, Borge GI, Kohler A. Analysis of-omics data: Graphical interpretation-and validation tools in multi-block methods. Chemometrics and Intelligent Laboratory Systems. 2010;104(1):140-53.

181. Kyriakides A, Kastanos E, Hadjigeorgiou K, Pitris C. Classification of Raman spectra using the correlation kernel. Journal of Raman Spectroscopy. 2011;42(5):904-9.

182. Jiang H, Ching W-K. Correlation Kernels for support vector machines classification with applications in cancer data. Computational and mathematical methods in medicine. 2012;2012.

Publications et communications

Publications scientifiques

- **Label free technologies: Raman micro-spectroscopy and multi-spectral imaging for lymphocyte classification.**
  Happillon T, Untereiner V, Beljebbar A, Gobinet C, Manfait M, Daliphard S, Cornillet-Lefebvre P, Troussard X, Angulo J, Velasco-Forero S, Saada V, Flandrin G, Klossa J.
  Diagnostic Pathology. 2013;8:S32

- **Quantitative phase imaging and raman micro-spectroscopy applied to malaria.**
  Klossa J, Wattelier B, **Happillon T**, Toubas D, de Laulanie L, Untereiner V, Bon P, Manfait M.
  Diagnostic Pathology. 2013;8:S42

- **Raman comparison of skin dermis of different ages: Focus on spectral markers of collagen hydration.**
  Nguyen TT, **Happillon T**, Feru J, Brassart-Passco S, Angiboust J-F, Manfait M, Piot O.
  Journal of Raman Spectroscopy. 2013;44:1230-1237

- **Probing non-enzymatic glycation of type I collagen: A novel approach using Raman and infrared biophotonic methods.**
  Guilbert M, Said G, **Happillon T**, Untereiner V, Garnotel R, Jeannesson P, Sockalingum GD.
  Biochimica Biophysica Acta. 2013;6:3525-3531

- **FCM parameter estimation methods: Application to infrared spectral histology of human skin cancers.**
  Happillon T, Sebiskveradze D, Vrabie V, Piot O, Jeannesson P, Manfait M, Gobinet C.
  Proceedings of the 20th European Signal Processing Conference (EUSIPCO), 2012 2012:1831-1835

- **Raman microspectroscopy detects epigenetic modifications in living jurkat leukemic cells.**
  Poplineau M, Trussardi-Régnier A, **Happillon T**, Dufer J, Manfait M, Bernard P, Piot O, Antonicelli F.
  Epigenomics. 2011;3:785-794

- **La biophotonique au service de l'identification de marqueurs pronostiques intracellulaires.**
  Klossa J, Daliphard S, Troussard X, Vielh P, Manfait M, Angulo J, Flandrin G, Beljebbar A, Untereiner V, **Happillon T**, Gobinet C, Velasco-Forero S, Roux S, Saada V, Dagiral R, Coomans V, Froigneux E, Rideau P, Vievard A.
  IRBM. 2011;32:72-75

Publications en cours de finalisation

- **A new method for the simultaneous estimation of the number of clusters and the fuzziness parameter in FCM,**
  Happillon T, Sebiskveradze D, Vrabie V, Manfait M, Gobinet C.
  Article à soumettre à Pattern Recognition

- **Review of validity indices and m-rules determining optimal Fuzzy C-Means parameters, applied on FTIR spectral images of skin cancer**
  Happillon T, Sebiskveradze D, Vrabie V, Manfait M, Gobinet C.
  Article en cours de rédaction

- **Diagnosis approach of Chronic Lymphocytic Leukemia on unstained blood smears using Raman microspectroscopy and supervised classification**
  Happillon T, Untereiner V, Beljebbar A, Gobinet C, Daliphard S, Cornillet-Lefebvre P, Quinquenel A, Delmer A, Troussard X, Klossa J, Manfait M.
  Article à soumettre à Analyst

Communications orales

- **Label free technologies: Raman micro-spectroscopy and multi-spectral imaging for lymphocyte classification.**
  Happillon T, Untereiner V, Beljebbar A, Gobinet C, Manfait M, Daliphard S, Cornillet-Lefebvre P, Troussard X, Angulo J, Velasco-Forero S, Saada V, Flandrin G, Klossa J.
  European congress on Telepathology, Venise, Italia, June 6-9, 2012.

- **Diagnostic de la Leucémie Lymphoïde Chronique sur des frottis sanguins par microspectroscopie Raman et classification supervisée**
  Happillon T, Untereiner V, Gobinet C, Beljebbar A, Daliphard S, Cornillet-Lefebvre P, Klossa J, Manfait M.
  OptDiag, Paris, France, du 9 au 11 mai 2012.

- **Diagnostic de la Leucémie Lymphoïde Chronique sur frottis sanguins non colorés par microspectroscopie Raman et classification supervisée**
  Happillon T, Untereiner V, Gobinet C, Beljebbar A, Daliphard S, Cornillet-Lefebvre P, Klossa J, Manfait M.
  Cancéropôle du Grand-Est, Strasbourg, France, du 2 au 3 novembre 2011.

- **Diagnostic de la Leucémie Lymphoïde Chronique sur frottis sanguins par microspectroscopie Raman et classification supervisée**
  Happillon T, Untereiner V, Gobinet C, Beljebbar A, Daliphard S, Cornillet-Lefebvre P, Klossa J, Manfait M.
  Journée des jeunes chercheurs, Reims, France, le 20 mai 2011.

- **Comparaison de méthodes d'automatisation du paramétrage de l'algorithme Fuzzy C-Means appliqué à l'imagerie spectrale infrarouge de tumeurs cutanées**
  Happillon T, Sebiskveradze D, Gobinet C, Manfait M, Jeannesson J, Vrabie V.
  Chimiométrie, Paris, France, du 31 novembre au 1er décembre 2009.

Communications par voie d'affiche

- **Diagnosis of chronic lymphoid leukemia on unstained blood smears using Raman microspectroscopy and supervised classification**
  Happillon T, Untereiner V, Gobinet C, Beljebbar A, Daliphard S, Cornillet-Lefebvre P, Klossa J, Manfait M.
  SPEC, Chiang Mai, Thailand, November 11-16, 2012.

- **FCM parameter estimation methods : Application to infrared spectral histology of human skin cancers**
  Happillon T, Sebiskveradze D, Vrabie V, Piot O, Jeannesson P, Manfait M, Gobinet C.
  Cancéropôle Grand-Est, Strasbourg, France, du 13 au 14 novembre 2012

- **FCM parameter estimation methods : Application to infrared spectral histology of human skin cancers**
  Happillon T, Sebiskveradze D, Vrabie V, Piot O, Jeannesson P, Manfait M, Gobinet C.
  European Signal Processing Conference, Bucarest, Roumania, August 27-31, 2012

- **Quantitative phase imaging and Raman micro-spectroscopy applied to Malaria**
  J. Klossa, B. Wattelier, **T. Happillon**, D. Toubas, L. de Laulanie, V. Untereiner, P. Bon, M. Manfait.
  European congress on Telepathology, Venise, Italia, June 6-9, 2012.

- **Diagnosis of chronic lymphoid leukemia on unstained blood smears using Raman microspectroscopy and supervised classification**
  Happillon T, Untereiner V, Gobinet C, Beljebbar A, Daliphard S, Cornillet-Lefebvre P, Klossa J, Manfait M.
  American Association for Cancer Research (AACR), Chicago, United-States, from March 31 to April 4, 2012.

- **Automatic clustering of infrared tissues images using an improved Fuzzy C-Means algorithm**
  T. Happillon, D. Sebiskveradze, J.K. Nallala, G.D. Sockalingum, V. Vrabie, D. Farley, C. Gobinet, M. Manfait
  Centre de Recherche Public (CRP), Reims, France, le 28 novembre 2011.

- **Automatic clustering of infrared tissues images using an improved Fuzzy C-Means algorithm**
  T. Happillon, D. Sebiskveradze, J.K. Nallala, G.D. Sockalingum, V. Vrabie, D. Farley, C. Gobinet, M. Manfait
  European Conference on the Spectroscopy of Biological Molecules (ECSBM), Coïmbra, Portugal, from August 29 to September 3, 2011.

Annexes

Dans cette partie figurent trois publications issues de collaborations scientifiques réalisées en dehors des thématiques de cette thèse.

Préambule

Dans le cadre de trois collaborations réalisées au sein du laboratoire, des outils de traitement, de sélection de variables et de classification de données présentés dans ces travaux de thèse ont été employés lors d'études portant sur trois thèmes :

1. La détection par microspectroscopie Raman de modifications épigénétiques au sein de lignées cellulaires leucémiques de type Jurkat.
2. La mise en évidence de marqueurs spectroscopiques Raman permettant une distinction entre des échantillons de peau de différents âges.
3. La mise en évidence par spectroscopies infrarouge et Raman de modifications au sein de collagène de type 1 glyqué *in vitro.*

Pour ces trois études, un support technique a été apporté concernant les différents prétraitements et classifications à appliquer aux données.

Une sélection de variables a également été réalisée par la méthode randfeatures dans les études 1 et 2.



Micro-Raman spectroscopy detects epigenetic modifications in living Jurkat leukemic cells

Aims: Classical biochemical and molecular methods for discerning cells with epigenetic modifications are often biologically perturbing or even destructive. We wondered whether the noninvasive laser tweezer Raman spectroscopy technique allowed the discrimination of single living human cells undergoing epigenetic modifications. **Materials & Methods:** Human Jurkat leukemic cells were treated with inhibitors of histone deacetylases (trichostatin A and MS-275). Epigenetic changes were monitored through histone electrophoresis, nuclear image cytometry and laser tweezer Raman spectroscopy. **Results:** Treatment of Jurkat cells with histone deacetylase inhibitors increased histone acetylation and induced chromatin organization changes. Characteristic vibrations, issued from laser tweezer Raman spectroscopy analyses, mostly assigned to DNA and proteins allowed discerning histone deacetylase inhibitor-treated cells from control with high confidence. Statistical processing of laser tweezer Raman spectroscopy data led to the definition of specific biomolecular fingerprints of each cell group. **Conclusion:** This original study shows that laser tweezer Raman spectroscopy is a label-free rapid tool to identify living cells that underwent epigenetic changes.

KEYWORDS: acetylation epigenetics histone MS-275 Raman microspectroscopy trichostatin A

Mathilde Poplineau[1‡], Aurélie Trussardi-Régnier*[1‡], Teddy Happillon[1], Jean Dufer[1], Michel Manfait[1], Philippe Bernard[2], Olivier Piot[1] & Frank Antonicelli[2]

[1]Unité MEDyC, UMR URCA CNRS 6237, IFR53, Faculté de Pharmacie, Université de Reims, 1 avenue du Maréchal Juin, 51096 Reims, France
[2]Université de Reims Champagne-Ardenne, Laboratoire de Dermatologie, UMR URCA CNRS 6237, IFR53, Faculté de médecine, Reims, France
**Author for correspondence:*
Tel.: +33 032 691 8016
Fax : +33 032 691 3550
aurelie.trussardi@univ-reims.fr
‡These authors contributed equally

Most of current biochemical and molecular techniques used to detect, identify and quantify epigenetic modifications are destructive. In order to avoid these shortfalls, new technologies based on light scattering methods are used to measure cell alterations, which are associated with changes in cell physical properties. Image cytometry study of pathological specimens or cell lines, through nuclear texture analysis, revealed that epigenetic mechanisms play a major role in determining chromatin patterns [1,2]. Therefore, this type of methodology was applied to evaluate the effects of epigenetic events on the chromatin higher order spatial organization. Several studies revealed variations in chromatin condensation and modifications in the spatial arrangement of chromatin packing states, either in response to histone acetylation [1–3] and/or DNA methylation [4,5] in various cell lines. However, information arising from such methodologies are only related to gross genomic changes observed on 2D preparations of fixed cells, and do not take into account the finely tuned and highly dynamic molecular modifications induced by these epigenetic modulations in cell suspension. Alternative methodologies such as confocal imaging of fluorescent probes enable more integrated approaches to study the nuclear organization [6]. However, this less invasive method is only applicable to characterize fixed cells, and often requires confirmatory biological evaluation, which is subject to the aforementioned shortfalls. Therefore, there is still a need for new techniques that overcome such limitations to provide an easily automated and accurate method for the detection of epigenetic alterations in both research and clinical domains. Thus, research development has now focused on new noninvasive biophysical approaches able to detect these molecular changes on living cells. Amongst these new technologies, Raman microspectroscopy could offer a promising way to analyze the chromatin structure [7].

Raman microspectroscopy is an optical laser-based technique which relies on the inelastic interaction between light (a monochromatic source) and matter. In complex samples such as cells or tissues, this tool provides information about the vibrational energies of the biomolecular constituents of the sample. Raman microspectroscopy does not require any sample labeling and appears to be noninvasive and nondestructive when suitable laser wavelengths and intensities are used. Recently, this method has been used in combination with microscopy and optical 'traps' or 'tweezers' to study isolated living cells [8,9]. This approach, known as laser tweezers Raman spectroscopy (LTRS), permits probing of a single cell over an extended period of time by 'trapping' the cell in suspension, provided

Author

future medicine part of fsg

that the cell is nonadherent and the trapping forces can overcome the size and weight of this cell. Previous studies have shown the potential of LTRS to sort living microorganisms [10], to monitor protein expression in living bacteria [11], to detect hyperosmotic stress in optically trapped yeast cells [12] and also to study dynamic cellular processes such as cell bacterial lysis [13]. The originality and efficiency of this physical noninvasive technique prompted us to investigate its potential to evaluate biological changes induced by epigenetic modulators in living cells.

Aberrant transcriptional gene regulations, related to genetic and epigenetic alterations, are observed during tumorigenesis and cancer progression. Amongst histone post-translational modifications associated with epigenetic regulations, the balance of histone acetylation/deacetylation has received a critical focus for transcriptional regulation since it controls chromatin accessibility [14–16]. Histone deacetylases (HDACs) play a key role in the regulation of gene expression by catalyzing the removal of acetyl groups on the N-terminal tail of histones, triggering chromatin condensation and promoting gene silencing [16–18]. Since aberrant epigenetic changes are nowadays a hallmark of cancer, drugs targeting these mechanisms are promising tools for pharmacological gene regulation. HDAC inhibitors (HDIs) can induce cell cycle arrest, promote differentiation and stimulate tumor cell death. In leukemia models, HDIs used either alone or in combination with other drugs, have demonstrated their ability to modulate gene expression, and for instance, it was shown that the HDI MS-275 could promote differentiation and apoptosis of human leukemic cells [19].

Thus, the purpose of our study was to evaluate the potential of LTRS to detect cellular changes induced by HDI treatment. To our knowledge, no study using Raman microspectroscopy has so far been performed to evaluate the effects of epigenetic drugs in individual living leukemic cells. In this study, we demonstrated that LTRS can be used to monitor epigenetic changes induced by the HDIs MS-275 and trichostatin A (TSA) in living unfixed Jurkat T-lymphocyte-derived cells. Statistical analysis of spectra revealed changes in the spectral fingerprint of HDI-treated cells that displayed distinct clusters between HDI-treated cells and untreated cells.

Materials & methods

Cell culture & treatments

Human Jurkat cell line (American Type Culture Collection, TIB-152) was obtained by Weiss *et al.* [20] from leukemic T cells. Cells were grown at 37°C in Roswell Park Memorial Institute medium supplemented with 10% fetal calf serum, 100 U/ml penicillin, and 100 µg/ml streptomycin in 95% humidified air, 5% CO_2 atmosphere. Cells were treated or not for 24 h with the following HDIs: 165 nM (50 ng/ml) of TSA (Sigma Aldrich, France) or 26.6 µM (10 µg/ml) of MS-275 (Alexis, San Diego, CA, USA). These concentrations were defined according to their Jurkat cell toxicity, and with respect to their known EC_{50} ranges [21].

Histone extraction & acid-urea polyacrylamide gel electrophoresis

Details of the procedure are previously described in El-Khoury *et al.* [1]. Briefly, 1.10^7 cells were kept in 500 µl of ice cold phosphate buffered saline. 2 ml of cytoplasmic lysis buffer (320 mM saccharose, 5 mM $MgCl_2$, 10 mM Tris–HCL pH 7.5, 1% Triton™ X100, 10 mM butyric acid, 1 mM benzamidine, 1 mM phenylmethylsulfonylfluoride) were added and incubated for 10 min on ice. After centrifugation, the nuclear pellet was washed with cytoplasmic lysis buffer and resuspended in 500 µl of nuclear lysis buffer (10 mM Tris–HCL pH7.4, 4 mM $MgCl_2$, 10 mM butyric acid, 0.1 mM phenylmethylsulfonylfluoride and 0.1 mM benzamidine). Histones were acid-extracted for 2 h with 500 µl of ice cold 0.8 M HCl. Acid-insoluble material was removed by centrifugation at 12,000 g for 10 min at 4°C. Supernatants were harvested, precipitated with eight volumes of acetone, kept at -20°C overnight, and histones were collected by centrifugation at 2000 g for 10 min at 4°C. Pellets were air-dried, suspended in distilled water and protein content evaluated using the Quant-it Assay Kit (Invitrogen, France).

The level of acetylation was analyzed by polyacrylamide gel electrophoresis (PAGE) using acid-urea-Triton X100 (AUT). The migration system consisted of a separating gel containing 13.1% acrylamide, 0.09% bisacrylamide, 4.5% glacial acetic acid, 3.2% Triton X100 and 5.2 M urea. Histones (10 µg) were separated for 16 h at 4°C at 150 V in a running buffer (17.4 M acetic acid). Histones were revealed using a staining solution (50% methanol, 10% acetic acid, 40% distilled water and 0.1% Coomassie blue R250) and a destaining solution (50% methanol, 10% acetic acid and 40% distilled water).

Conventional image cytometry

Cells were smeared by cytocentrifugation, air-dried, fixed in alcohol-formalin and stained by

Author Proof

the Feulgen method. Image cytometry was performed with an image analysis system (Samba Technologies, Meylan, France) as previously reported [1]. In brief, at least 300 interphase cell nuclei were analyzed on each microscopic slide. Eleven parameters were computed from each nuclear image: one geometric feature (nuclear area), one densitometric parameter (the DNA integrated optical density [DNA-IOD]) and nine textural parameters. Geometric and textural features estimated by image cytometry are closely related to the location of a given cell in the cell cycle. G0/G1 cells were therefore isolated on the basis of their DNA-IOD value as previously reported and analyzed for morphometry and chromatin texture.

Nine textural parameters were calculated on the nuclear image. Four features were calculated on the gray-levels co-occurrence matrix: local mean, energy, entropy and inertia. Five parameters were calculated on the run-length matrix: short run-length emphasis, long run-length emphasis, gray levels distribution, run-length distribution and run-length percentage.

The distribution, mean and standard deviation of the nuclear parameters were calculated for each cell population. Significance of the differences between parameter values were estimated by student T-test after Bonferroni correction for multiple variables.

To compare multiple cell groups, a deviation index was computed and an average nuclear abnormality was then calculated for each cell line as the arithmetic mean of the absolute values of the deviation indexes (z-scores) observed over all included features for all nuclei in a data set as previously reported [22–24].

Raman microspectroscopy

In order to avoid spectral interference with the cell culture medium, 1.10^7 cells were washed twice with phosphate buffered saline and resuspended in physiological serum. More than 20 living cells were randomly analyzed for each sample and the acquisition procedure was repeated three times. Two spectra per cell were collected consecutively, thus leading to approximately 120 spectra per sample. These spectra were obtained using a Labram microspectrometer (Horiba Scientific, Villeneuve d'Ascq, France) coupled to a charge-coupled device camera (Andor, Belfast, Northern Ireland) and a right microscope. The excitation light was provided by a titanium-sapphire laser tuned at 785 nm. A 100× water immersion objective (numerical aperture = 1) was used to focus the laser and to trap the probed cell. The laser beam waist was evaluated to 4 µm, and the laser power at the objective output was adjusted to 35 mW, resulting in an irradiance of approximately $2.8.10^8$ mW/cm^2 [25]. Spectra were collected with an integration time of 30 s and the acquisition was performed in a temperature-controlled room. Observations of the sample before and just after the acquisition allowed for the checking that only one cell was trapped at once and that the cell stayed trapped all along the measurement. Rayleigh elastic scattering and reflected laser were rejected by placing an edge filter before the spectroscopic part of the instrument. The Raman Stokes scattering was spectrally analyzed in the 590–1750 cm^{-1} spectral region with a 950 grooves/mm holographic grating, assuring a spectral resolution of 4 cm^{-1}.

Data processing & statistical analysis

A Raman spectrum is composed of 1024 points along the wavenumber axis. A value of the Raman scattering intensity is measured to each wavenumber. Raman spectra were first corrected for the apparatus response by dividing the raw spectra by a National Institute of Standard spectrum. The National Institute of Standard was used to mimic a black body in order to control and correct the dependency of the instrument efficiency with the light frequency. After this first step, the spectra were slightly smoothed and baseline corrected using toolboxes of Labspec software (Horiba Scientific, Villeneuve d'Ascq, France). Parameters used to treat spectra were: three points and average mode for filtering and line segment baseline using manually selected wavenumbers (17 minima points were used). Spectra were afterwards vector normalized on the whole 590–1750 cm^{-1} spectral range.

After this correction step, spectra were processed by a random features method for selection of discriminant wavenumbers, and by hierarchical cluster analysis (HCA) for spectra classification. Random features algorithm permitted to identify in a nonsupervised way the most relevant wavenumbers in order to discriminate two groups of spectra. For this processing, the MatLab randfeatures function (MathWorks, Inc., Natick, MA, USA) was used with standard parameters. A score plot allowed visualizing, in a decreasing order, the discriminant weight of the wavenumbers.

HCA of spectra was then performed on the wavenumbers selected by the random features algorithm, using a customized-software developed in MatLab. Grouping of spectra was carried out using the Euclidean distance and the Ward's

Author Proof

method for linkage. This method proved to be efficient to reveal distinctive classes of Raman spectra, with relevance to biological properties of samples [26].

Results

Biomolecular effects of TSA & MS-275 on histone post-translational modifications

The first step to analyze the ability of LTRS to detect cellular changes induced by HDIs has been to validate a cellular model displaying different levels of histone acetylation. For that purpose, Jurkat cells were treated or not with nontoxic doses of either, TSA or MS-275 (data not shown). Separation of histones H1, H2B, H3 and H4 was performed using AUT-PAGE (Figure 1). In these stimulatory conditions, treatment of Jurkat cells with MS-275 markedly increased acetylation of histones H4 and H3, whereas the effects of TSA was mainly observed on histone H3, and weekly on histone H4.

Effects of TSA & MS-275 on nuclear morphometry & texture

The effects of these two HDIs on the nuclear organization of fixed Jurkat cells were analyzed using conventional image cytometry. Values of the analyzed features were estimated in G0/G1 cells and are listed in Table 1. Control and HDI-treated cell populations exhibited the same nuclear areas. Values of the textural features showed significant variations in MS-275-treated cells only. These changes corresponded to a slight, but significant, chromatin global decondensation (decrease in local mean by 7% compared with control). Besides, the main chromatin alteration concerned spatial arrangements with a less compact (increase in energy, whilst decreases in entropy and inertia by 73, 20 and 23% compared with control, respectively) and more homogeneously and finely distributed chromatin (decrease in short run-length emphasis [4%] and run-length percentage [10%], whilst increases in long run-length emphasis and gray level distribution by 35 and 38% with respect to controls). Meanwhile, TSA did not induce any significant modification of this chromatin supraorganization compared with untreated cells. Lack of effects of TSA on the structural organization relatively to MS-275 treatments was visualized in the integrative average nuclear abnormality representation (Figure 2).

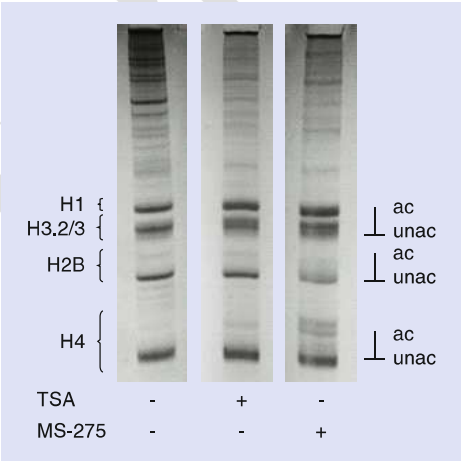


Figure 1. Histone deacetylase inhibitor treatment modulates histone acetylation in Jurkat cancer cells. Jurkat cancer cells were treated either with TSA 165 nM (50 ng/ml) or MS-275 26.6 µM (10 µg/ml) for 24 h. Total histones were acid extracted and analyzed by acid-urea-TritonX100 polyacrylamide gel electrophoresis.
ac: Acetyled; MS-275: ; TSA: Trichostatin A; unac: Unacetyled.

HDI treatment induces Raman spectral differences in Jurkat cells

Effects of HDI treatment on living Jurkat cells in suspension was then investigated using the LTRS technique. First, aberrant spectra were eliminated. Then, spectra issued from MS-275- (n = 107) and TSA- (n = 119) treated cells were merged to form one group (HDI-treated cells). These 226 Raman spectra from HDI-treated cells and 107 Raman spectra from untreated cells were processed by random features selection of discriminant wavenumbers and HCA. This processing resulted in the generation of a score plot (Supplementary Figure 1; www.futuremedicine.com/doi/suppl/10.2217/epi.11.102) and in the identification of 16 discriminant wavenumbers (Table 2) leading to a classification in two separate clusters (Figure 3). 77.5% of spectra from control cells were localized in cluster 1, and 76.1% of spectra from HDI-treated cells appeared in cluster 2 (Table 3). Adding more wavenumbers in the discriminant analysis did not result in a better classification of HDI-treated and control cells (data not shown). The most discriminant wavenumbers belonged mainly to one spectral range, located between 978 and 989 cm^{-1}, as shown in Figure 4 & Table 2. Other wavenumbers (i.e., 839, 846, 848, 1355, 1358 or 1413) have also been identified by the selection algorithm, but with lower discriminant scores (Table 2). Besides these wavenumbers defined by random features selection with a very high discriminant power ($p < 1.10^{-8}$), we also observed a significant decrease in the intensity

Author Proof

Table 1. Nuclear geometric and textural parameter values in G0/G1 Jurkat cells.

Parameter	Treatment		
	Control	*TSA*	*MS-275*
A	404.6 ± 82.6	405.1 ± 83.9	404.7 ± 70.0
LM	199.1 ± 47.6	192.3 ± 45.3	185.4 ± 39.5†
E	8.6 ± 2.6	8.0 ± 2.4	14.9 ± 4.3†
ENT	279.2 ± 31.2	285.8 ± 30.5	224.2 ± 30.2†
I	302.0 ± 52.4	310.9 ± 52.9	233.6 ± 20.9†
SRE	61.6 ± 2.8	62.2 ± 2.9	59.2 ± 1.1†
LRE	5.6 ± 0.9	5.4 ± 0.9	7.6 ± 0.8†
GLD	20.3 ± 4.2	19.3 ± 3.9	28.1 ± 5.3†
RLD	36.4 ± 3.1	37.2 ± 3.1	33.5 ± 1.1†
RPC	104.8 ± 6.9	106.7 ± 6.9	94.4 ± 3.6†

Values are mean ± standard error of mean, and data are expressed in arbitrary units.
†Significantly different from control cells (T-test after Bonferroni correction for multiple variables).
A: Nuclear area; E: Energy; ENT: Entropy; GLD: Graylevel distribution; I: Inertia; LM: Local mean of gray levels; LRE: Long run-length emphasis; RLD: Run-length distribution; RPC: Run-length percentage; SRE: Short run-length emphasis; TSA: Trichostatin A.

of several peaks (720–725; 785; 1093–1094; 1315–1320; and 1575 cm^{-1}) ($p < 0.001$; Kruskal Wallis test).

The potential of LTRS to discriminate TSA-treated cells from control cells and MS-275-treated cells from control cells was then assessed using the 16 wavenumbers previously retained by random features and shown in the Table 2. This analysis showed that 73.1 and 80.4% of spectra from TSA- and MS-275-treated cells, respectively, were localized in cluster 2 (Table 3).

This type of analysis was applied in order to discriminate TSA- from MS-275-treated cells. Random features selection followed by HCA was unable to efficiently separate these two cell groups (data not shown).

Discussion

In this study, we focused our investigation on the histone acetylation/deacetylation balance, which is kept under control in the cell by the expression of the histone acetylase and HDAC enzymes. Accordingly, we found that the biological consequence of HDAC inhibition by TSA and MS-275 was an accumulation of acetylated histones in the cell nucleus. Histone H1 remained unaffected by both treatments. Accordingly, this histone form is not acetylated through post-translational processes *in vivo* [27]. Nevertheless, modification of histone acetylation upon MS-275 treatment appeared more pronounced compared with that induced by TSA. Several matters could be implicated in these differences. MS-275 was used at higher concentration than TSA. However, both HDI concentrations were adjusted at sublethal levels and within their EC_{50} ranges [21]. Therefore, differences observed could be mainly due to their substrate specificity. In that line, use of other HDIs such as suberoylanilide hydroxamic acid and sodium butyrate also displayed specific patterns of histone acetylation (data not shown). Concomitantly to these histone acetylation changes, we showed that these HDI treatments were associated with modifications of the nuclear morphometry and texture, and with Raman spectral differences.

The introduction of epigenetic strategies in the therapeutic arsenal against cancer led to the

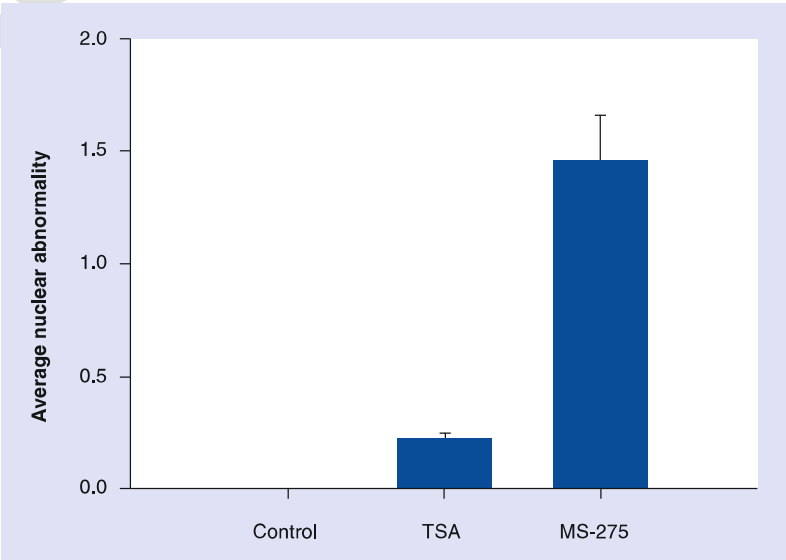


Figure 2. Average nuclear abnormality index extracted from conventional cytometry analysis. Deviation indexes signatures for nuclei from Jurkat cells treated either with TSA or MS-275 were calculated according to Bartels [22], and the average nuclear abnormality was computed as described by Montironi [23]. Indexes (mean ± standard error of mean) are expressed as units of standard deviation in the control untreated Jurkat data set.
TSA: Trichostatin A.

Author Proof

Table 2. Classification of wavenumbers according to their score obtained after random features analysis.

Wavenumbers	Score†
984	184
986	157
983	134
987	124
981	93
982	93
988	85
979	76
978	64
848	50
839	43
989	43
846	38
1358	37
1355	35
1413	35

The more discriminant wavenumbers are represented in the table.
†The values of the score reflect the discriminating power of the corresponding wavenumber.

need of techniques able to evaluate the efficiency of these therapies *in vivo*. Besides molecular biological approaches, these drug effects were also monitored through morphological methods such as nuclear texture analysis. For instance, conventional image cytometry showed significant morphometric signature changes after cell treatments with epigenetic modulators, and similar phenomena have already been observed in human fibrosarcoma [5] or HeLa cells [28]. Interestingly, MS-275 only induced strong variations in chromatin distribution, whereas TSA seemed ineffective on these parameters. In previous studies, we showed that TSA could induce modulation of cytometric features in cells whose DNA was hypomethylated such as OV1 tumor cells [29]. In HT1080, whose DNA was less hypomethylated, TSA appeared effective only after previous treatment with the demethylating agent 5-azadC [5]. This suggests different mechanisms involved by this HDI treatment in chromatin supraorganization changes evaluated by image cytometry. However, this technique could only be applied to fixed-cell specimens and could hardly be correlated with specific molecular events. Moreover, this technique requires cell staining. Raman microspectroscopy could offer a new noninvasive way to follow dynamic epigenetic changes in living, unstained and unfixed tumor cells. We previously reported the use of Raman microspectroscopy to discriminate healthy epidermis from normal and peritumoral stroma in basal cell carcinoma [30]. Moreover, such an approach has been recently applied to the identification of stem cell states [8,31] and to the follow-up of their differentiation pathways by the detection of intracellular molecular markers [32]. A previous study from Chan *et al.* [33] also revealed the beneficial use of LTRS to discriminate human T-lymphocytes from Jurkat cells, hence demonstrating the potential of this technique to discriminate malignant from normal lymphoid cells. Concerning the study of epigenetic

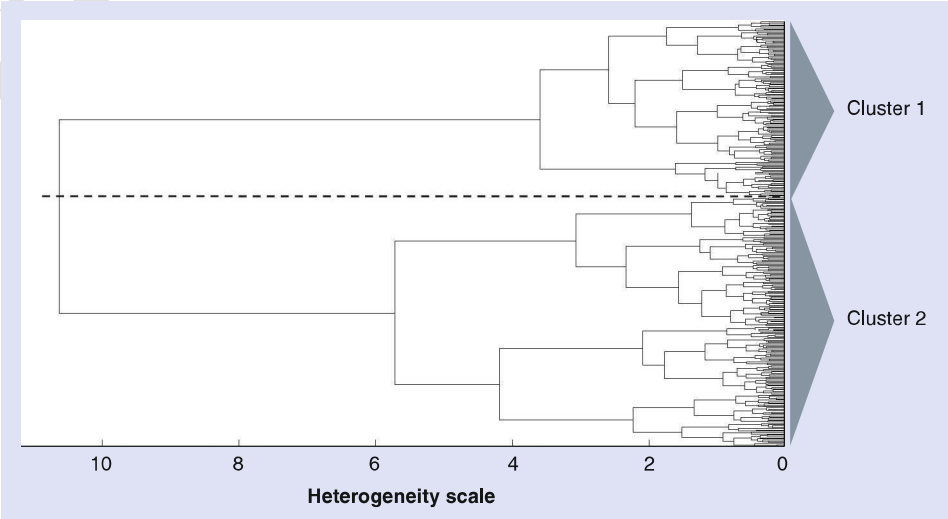


Figure 3. Hierarchical cluster analysis of spectra issued from control and histone deacetylase inhibitor-treated cells. Hierarchical cluster analysis was performed using the discriminant wavenumbers selected by random features analysis listed in TABLE 2.

Author Proof

Table 3. Distribution of spectra from the different cell groups among the clusters defined after random feature analysis and hierarchical cluster analysis of control and histone deacetylase inhibitor-treated cell populations.

Cell group	n†	Spectra classified in clusters (%)	
		Cluster 1	*Cluster 2*
Control	107	77.5	22.5
HDI	226	23.9	76.1
TSA	119	26.9	73.1
MS-275	107	19.6	80.4

Overall classification accuracy: 76.8
†Number of spectra analyzed in each experimental group.
HCA: Hierarchical cluster analysis; HDI: histone deacetylase inhibitor; TSA: Trichostatin A.

phenomena, Raman-based techniques seemed to be restricted, up to now, to the analysis of acellular *in vitro* molecular changes observed on DNA or histones [9,34] and on miRNA identification [35], and to our knowledge no study at the level of the whole-living cell has been performed. We demonstrated here that LTRS, coupled to a statistical data processing, allowed detecting epigenetic changes in whole-living cells.

Epigenetic modulators induce finely tuned molecular changes that can be monitored by Raman microspectroscopy. Using Jurkat cells, we demonstrated that HDI treatments induced changes in the spectral signatures issued from single-cell LTRS. Before treatment, Jurkat cells displayed spectral signatures similar to those previously described by Chan and colleagues [33,36]. Compared with these control cells, HDI-treated cells displayed spectral changes mostly located in the 978–989 cm^{-1} range and around 845 cm^{-1}. With respect to epigenetic changes, these spectral signals have been associated with phosphate bonds and deoxyribose ring vibrations [37]. These changes could be related to modifications of nucleosome conformation induced by the addition of acetyl residues at the N-terminal tail of histones. This was in line with the significant decrease in the intensity of several peaks (720–725; 785; 1093–1094; 1315–1320; and 1575 cm^{-1}), which could be attributed to a decrease of DNA signal [38]. Indeed, these peaks have been attributed to ring breathing modes in the DNA bases, (725, 785, 1318 and 1575 cm^{-1}) or to the symmetric PO_2 stretching vibration of the DNA backbone (1093 cm^{-1}) [38]. Since Raman signals intensity of these vibrations is linked to the concentration of molecular bonds [39] or structural modifications of the related compounds, this decrease in intensity of DNA specific peaks could reflect the chromatin decondensation induced by HDI treatment and already observed by image cytometry in MS-275-treated cells.

Interestingly, some of the DNA relative features allowing discrimination between Jurkat cells and healthy T-lymphocytes (785, 1093 and 1575 cm^{-1}) [33] were also observed in our study when we compared spectra of HDI-treated Jurkat cells to spectra of control cells. As stated above, these spectral modifications suggested an overall DNA decondensation in HDI-treated cells as it was observed in T-lymphocytes compared with their leukemic malignant counterpart. Nevertheless, the differences observed between TSA- and MS-275-treated cells both at the AUT-PAGE and the image cytometry levels were not detected by Raman microspectroscopy analysis. However, the enlarging of the number of discriminant wavenumbers in the random

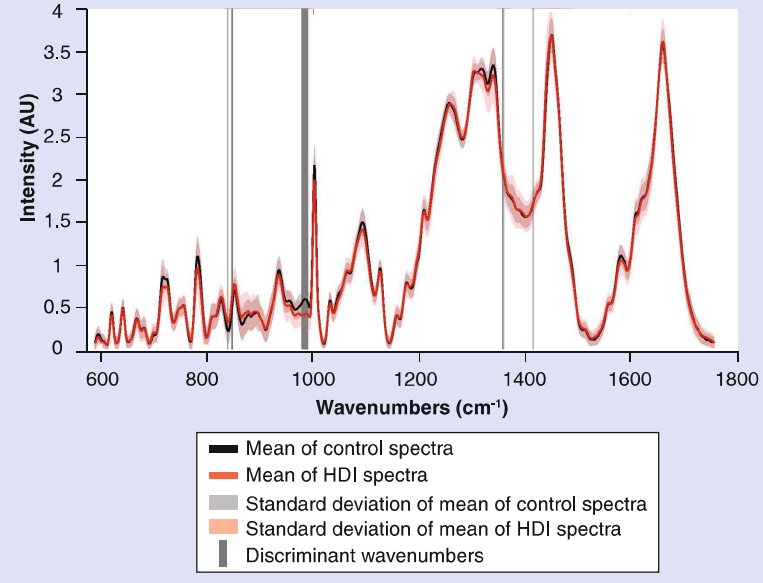


Figure 4. Raman spectra representative of control and histone deacetylase inhibitor-treated cells. Data correspond to 107 individual spectra from control cells and 226 spectra from histone deacetylase inhibitor-treated cells (107 MS-275- and 119 trichostatin A-treated cells). Discriminant wavenumbers were selected after random features analysis.
AU: Arbitrary units; HDI: HDAC inhibitor.

Author Proof

features analysis improved the spectral discrimination of MS-275-treated cells from control cells. This improvement could not be observed for TSA-treated cells (data not shown), suggesting differential biological changes induced by these two drugs as already observed.

Thus, LTRS seems to be a noninvasive powerful technique to assess epigenetic changes in living cells. More advanced LTRS developments could be set up to probe more selectively cellular compartments. In the results described above, the Raman analysis was performed with the simplest configuration in that the same laser beam was used as optical tweezers and simultaneously as excitation source for Raman scattering. However, different experimental schemes could be set up, and improve in the future, the quality of interpretation of the biological alterations observed either in pathological conditions or upon treatment [40–44]. In more advanced configurations, dual wavelength optical tweezers could be coupled to confocal Raman microscopes; when one of the beams excites the Raman spectrum, a second beam for tweezing enables manipulation of the environment of the trapped objects, for example, a single living cell. Raman microspectroscopy could also be combined with dual beam optical traps which allow distinguishing Raman spectra from the nucleus, cytoplasm and cell membrane. In forthcoming studies, these technological improvements could allow for deeper investigations of the molecular mechanisms associated with the observed epigenetic-induced spectral changes.

Acknowledgements

The authors wish to thank Valérie Untereiner for her expert technical advice.

Financial & competing interests disclosure

This work was supported by grants from the Comité Départemental de l'Aube de la Ligue Nationale Contre le Cancer. The authors would like to thank CNRS Projets Exploratoires Pluridisciplinaires for financial support. The authors have no other relevant affiliations or financial involvement with any organization or entity with a financial interest in or financial conflict with the subject matter or materials discussed in the manuscript apart from those disclosed.

No writing assistance was utilized in the production of this manuscript.

Ethical conduct of research

The authors state that they have obtained appropriate institutional review board approval or have followed the principles outlined in the Declaration of Helsinki for all human or animal experimental investigations. In addition, for investigations involving human subjects, informed consent has been obtained from the participants involved.

UNCORRECTED Proof

Executive summary

Histone acetylation in human leukemic Jurkat cells can be modulated by histone deacetylase inhibitors

- Using biochemical methods, we showed the modulation of histone acetylation level in Jurkat cells treated or not by two different histone deacetylase inhibitors (HDIs).
- The MS-275 effects were more pronounced than those of trichostatin A in terms of histone acetylation.
- Biochemical variations of histone acetylation were confirmed by conventional image cytometry on fixed cells.

Laser tweezer Raman spectroscopy can be used to perceive epigenetic modifications in living Jurkat cells

- Molecular and biochemical conventional approaches of epigenetic alteration are biologically perturbing or even destructive.
- Up to now and concerning epigenetics, Raman microspectroscopy was restricted to the analysis of the acellular *in vitro* modeling of the molecular changes observed on DNA or histones.
- We demonstrated here that Raman microspectroscopy, thanks to its trapping capability of the excitation laser, could also be used to probe isolated living cells and allow epigenetic discrimination.

Laser tweezer Raman spectroscopy detects cellular biological modifications following histone deacetylase inhibitors treatment

- The random features selection associated with hierarchical cluster analysis discriminated histone deacetylase inhibitor-treated cells from control cells on the basis of their Raman spectral signatures.

Conclusion

- Laser tweezer Raman spectroscopy is a rapid label-free tool to identify living cells that underwent epigenetic changes.
- In the future, laser tweezer Raman spectroscopy technique improvements should allow for the fine detection of the molecular mechanisms associated with the observed epigenetic-induced changes in living cells.

References

Papers of special note have been highlighted as:
- ▪ of interest
- ▪▪ of considerable interest

1 El-Khoury V, Gomez D, Liautaud-Roger F, Trussardi-Regnier A, Dufer J. Effects of the histone deacetylase inhibitor trichostatin A on nuclear texture and *c-jun* gene expression in drug-sensitive and drug-resistant human H69 lung carcinoma cells. *Cytometry A* 62(2), 109–117 (2004).

2 Watson JA, Mckenna DJ, Maxwell P *et al.* Hyperacetylation in prostate cancer induces cell cycle aberrations, chromatin reorganization and altered gene expression profiles. *J. Cell. Mol. Med.* 14(6B), 1668–1682 (2010).

3 Yatouji S, Trussardi-Regnier A, Trentesaux C, Liautaud-Roger F, Dufer J. Nuclear texture and chromatin structure in OV1/VCR human multidrug-resistant cell line. *Int. J. Oncol.* 23(4), 1225–1230 (2003).

4 Mampumbu AR, Mello ML. DNA methylation in stingless bees with low and high heterochromatin contents as assessed by restriction enzyme digestion and image analysis. *Cytometry A* 69(9), 986–991 (2006).

5 Poplineau M, Dufer J, Antonicelli F, Trussardi-Regnier A. Epigenetic regulation of proMMP-1 expression in the HT1080 human fibrosarcoma cell line. *Int. J. Oncol.* 38(6), 1713–1718 (2011).

6 Stein Gs, Stein JL, Van Wijnen AJ *et al.* An architectural genetic and epigenetic perspective. *Integr. Biol. (Cambridge)* 3(4), 297–303 (2011).

▪ **Review on the nuclear regulatory machinery and architectural organization.**

7 Huser T, Orme CA, Hollars CW, Corzett MH, Balhorn R. Raman spectroscopy of DNA packaging in individual human sperm cells distinguishes normal from abnormal cells. *J. Biophotonics* 2(5), 322–332 (2009).

8 Schulze HG, Konorov SO, Caron NJ, Piret JM, Blades MW, Turner RF. Assessing differentiation status of human embryonic stem cells noninvasively using Raman microspectroscopy. *Anal. Chem.* 82(12), 5020–5027 (2010).

9 Sundararajan N, Mao D, Chan S *et al.* Ultrasensitive detection and characterization of post-translational modifications using surface-enhanced Raman spectroscopy. *Anal. Chem.* 78(11), 3543–3550 (2006).

10 Chan JW, Esposito AP, Talley CE, Hollars CW, Lane SM, Huser T. Reagentless identification of single bacterial spores in aqueous solution by confocal laser tweezers Raman spectroscopy. *Anal. Chem.* 76(3), 599–603 (2004).

11 Chan JW, Winhold H, Corzett MH *et al.* Monitoring dynamic protein expression in living *E. coli*. Bacterial cells by laser tweezers Raman spectroscopy. *Cytometry A* 71(7), 468–474 (2007).

12 Singh GP, Creely CM, Volpe G, Grotsch H, Petrov D. Real-time detection of hyperosmotic stress response in optically trapped single yeast cells using Raman microspectroscopy. *Anal. Chem.* 77(8), 2564–2568 (2005).

13 Chen D, Shelenkova L, Li Y, Kempf CR, Sabelnikov A. Laser tweezers Raman spectroscopy potential for studies of complex dynamic cellular processes: single cell bacterial lysis. *Anal. Chem.* 81(9), 3227–3238 (2009).

14 Kouzarides T. SnapShot: histone-modifying enzymes. *Cell* 131(4), 822 (2007).

15 Kouzarides T. SnapShot: histone-modifying enzymes. *Cell* 128(4), 802 (2007).

16 Kouzarides T. Chromatin modifications and their function. *Cell* 128(4), 693–705 (2007).

17 Mizzen CA, Allis CD. Linking histone acetylation to transcriptional regulation. *Cell. Mol. Life Sci.* 54(1), 6–20 (1998).

18 Wolffe AP. Centromeric chromatin. Histone deviants. *Curr. Biol.* 5(5), 452–454 (1995).

19 Rosato RR, Almenara JA, Grant S. The histone deacetylase inhibitor MS-275 promotes differentiation or apoptosis in human leukemia cells through a process regulated by generation of reactive oxygen species and induction of p21CIP1/WAF1 1. *Cancer Res.* 63(13), 3637–3645 (2003).

20 Weiss A, Wiskocil RL, Stobo JD. The role of T3 surface molecules in the activation of human T-cells: a two-stimulus requirement for IL-2 production reflects events occurring at a pre-translational level. *J. Immunol.* 133(1), 123–128 (1984).

21 Khan N, Jeffers M, Kumar S *et al.* Determination of the class and isoform selectivity of small-molecule histone deacetylase inhibitors. *Biochem. J.* 409(2), 581–589 (2008).

22 Bartels PH, Da Silva VD, Montironi R *et al.* Chromatin texture signatures in nuclei from prostate lesions. *Anal. Quant. Cytol. Histol.* 20(5), 407–416 (1998).

23 Montironi R, Scarpelli M, Mazzucchelli R *et al.* Subvisual changes in chromatin organization state are detected by karyometry in the histologically normal urothelium in patients with synchronous papillary carcinoma. *Hum. Pathol.* 34(9), 893–901 (2003).

24 Trussardi-Regnier A, Lavenus S, Gorisse MC, Dufer J. Thalidomide alters nuclear architecture without *ABCB1* gene modulation in drug-resistant myeloma cells. *Int. J. Oncol.* 35(3), 641–647 (2009).

25 Draux F, Jeannesson P, Beljebbar A *et al.* Raman spectral imaging of single living cancer cells: a preliminary study. *Analyst* 134(3), 542–548 (2009).

26 Draux F, Gobinet C, Sule-Suso J *et al.* Raman spectral imaging of single cancer cells: probing the impact of sample fixation methods. *Anal. BioAnal. Chem.* 397(7), 2727–2737 (2010).

▪ **Technical aspect of Raman spectroscopic analyses on single cells.**

27 Alvelo-Ceron D, Niu L, Collart DG. Growth regulation of human variant histone genes and acetylation of the encoded proteins. *Mol. Biol. Rep.* 27(2), 61–71 (2000).

28 Toth KF, Knoch TA, Wachsmuth M *et al.* Trichostatin A-induced histone acetylation causes decondensation of interphase chromatin. *J. Cell. Sci.* 117(Pt 18), 4277–4287 (2004).

29 Yatouji S, El-Khoury V, Trentesaux C *et al.* Differential modulation of nuclear texture, histone acetylation, and *MDR1* gene expression in human drug-sensitive and -resistant OV1 cell lines. *Int. J. Oncol.* 30(4), 1003–1009 (2007).

30 Ly E, Durlach A, Antonicelli F, Bernard P, Manfait M, Piot O. Probing tumor and peritumoral tissues in superficial and nodular basal cell carcinoma using polarized Raman microspectroscopy. *Exp. Dermatol.* 19(1), 68–73 (2010).

31 Pijanka JK, Kumar D, Dale T *et al.* Vibrational spectroscopy differentiates between multipotent and pluripotent stem cells. *Analyst* 135(12), 3126–3132 (2010).

32 Pascut FC, Goh HT, Welch N, Buttery LD, Denning C, Notingher I. Noninvasive detection and imaging of molecular markers in live cardiomyocytes derived from human embryonic stem cells. *Biophys. J.* 100(1), 251–259 (2011).

33 Chan JW, Taylor DS, Zwerdling T, Lane SM, Ihara K, Huser T. Micro-Raman spectroscopy detects individual neoplastic and normal hematopoietic cells. *Biophys. J.* 90(2), 648–656 (2006).

▪▪ **Analysis of leukemic cell Raman spectroscopic signatures.**

34 Kelly JG, Najand GM, Martin FL. Characterization of DNA methylation status using spectroscopy (mid-IR versus Raman) with multivariate analysis. *J. Biophotonics* 4(5), 345–354 (2010).

35 Driskell JD, Seto AG, Jones LP *et al.* Rapid microRNA (miRNA) detection and classification via surface-enhanced Raman spectroscopy (SERS). *Biosens. Bioelectron* 24(4), 923–928 (2008).

36 Chan JW, Taylor DS, Lane SM, Zwerdling T, Tuscano J, Huser T. Nondestructive identification of individual leukemia cells by laser trapping Raman spectroscopy. *Anal. Chem.* 80(6), 2180–2187 (2008).

37 Parker F. *Applications of Infrared, Raman and Raman Resonance, Spectroscopy in Biochemistry.* Plenum Press, New York, NY, USA (1983).

38 Movasaghi Z, Rehman S, Rehman IU. Raman spectroscopy of biological tissues. *Appl. Spectrosc. Rev.* 42, 493–541 (2007).

39 Moritz TJ, Taylor DS, Krol DM, Fritch J, Chan JW. Detection of doxorubicin-induced apoptosis of leukemic T-lymphocytes by laser tweezers Raman spectroscopy. *Biomed. Opt. Express* 1(4), 1138–1147 (2010).

Author Proof

40 Bankapur A, Zachariah E, Chidangil S, Valiathan M, Mathur D. Raman tweezers spectroscopy of live, single red and white blood cells. *PLoS One* 5(4), e10427 (2010).

41 Dochow S, Krafft C, Neugebauer U *et al.* Tumor cell identification by means of Raman spectroscopy in combination with optical traps and microfluidic environments. *Lab. Chip* 11(8), 1484–1490 (2011).

▪ **Application of Raman spectroscopy in a microfluidic environment enabling high-throughput of cells.**

42 Liu R, Taylor DS, Matthews DL, Chan JW. Parallel analysis of individual biological cells using multifocal laser tweezers Raman spectroscopy. *Appl. Spectrosc.* 64(11), 1308–1310 (2010).

43 Snook RD, Harvey TJ, Correia Faria E, Gardner P. Raman tweezers and their application to the study of singly trapped eukaryotic cells. *Integr. Biol. (Cambridge)* 1(1), 43–52 (2009).

▪▪ **Reviews Raman tweezers and their biological applications.**

44 Zhang P, Kong L, Setlow P, Li YQ. Multiple-trap laser tweezers Raman spectroscopy for simultaneous monitoring of the biological dynamics of multiple individual cells. *Opt. Lett.* 35(20), 3321–3323 (2010).

Author Proof

Research article

Journal of RAMAN SPECTROSCOPY

Received: 1 December 2012 Revised: 13 June 2013 Accepted: 23 June 2013 Published online in Wiley Online Library

(wileyonlinelibrary.com) DOI 10.1002/jrs.4355

Raman comparison of skin dermis of different ages: focus on spectral markers of collagen hydration

The Thuong Nguyen,[a] Teddy Happillon,[a] Jezabel Feru,[b] Sylvie Brassart-Passco,[b] Jean-François Angiboust,[a] Michel Manfait[a] and Olivier Piot[a]*

Skin aging is the most visible sign of aging of the body. This complex process involves molecular and structural alterations of the main skin constituents. The major cutaneous constituent is type I collagen that gives strength and resilience to the skin. This macromolecule possesses a particular triple helix structure and is arranged in the form of a fibrous network. Water plays a crucial role for maintaining this molecular assembly which is altered during intrinsic chronological aging. To investigate some of these structural alterations, Raman microspectroscopy was employed since this biophotonic approach permits to probe the biomolecular composition of the skin in a non-destructive manner. In addition, type I collagen gives specific Raman vibrations; some of which being sensitive to the molecule conformation and to the water environment. In this purpose, Raman spectra of four skin samples of different ages (two samples of 40 year old and two samples of 70 year old) were collected by varying the relative environmental humidity. Our experiments highlighted spectral differences between the four samples. The analysis of the Raman data permitted to suggest a model for the interactions between collagen and water molecules and a decrease in collagen fibers compactness with chronological aging. Our study demonstrates that Raman spectroscopy can be a useful tool to investigate skin aging, with innovative applications in dermatology as well as in cosmetics. Copyright © 2013 John Wiley & Sons, Ltd.

Supporting information may be found in the online version of this article.

Keywords: Raman microspectroscopy; type I collagen; hydration; skin aging; amide I fitting

Introduction

Skin is a multifunctional particular organ. Its main function is a protective role against external aggression due to its three-layered structure (epidermis, dermis and hypodermis). The dermis is particularly important; it is composed of dense connective tissue and contains the vascular system which provides vital nutrients to living cells in the epidermis, as well as nerve fibers and sensory receptors. Its main molecular constituent is type I collagen which represents about 80% of dry weight; this protein participates to ensure the strength and elasticity of the skin. During the chronological skin aging, also called intrinsic aging, the dermis is strongly altered resulting in molecular and morphological modifications. Indeed, skin aging is the most visible sign of the aging of the body with the appearance of wrinkles, sagging and loss of elasticity. The chronological skin aging is a very complex biological process during which protein compounds of the dermis such as collagen, elastin or glycosaminoglycans (GAGs) are structurally altered. Concerning collagen alterations, it has been shown that aging induces a decrease of collagen I density[1] and an increase in the collagen III/collagen I ratio.[2] In addition, the three-dimensional arrangements of collagen and elastic fibers showed marked alterations with age. Collagen bundles appear to unravel.[3] Collagen becomes less soluble, thinner in intrinsically aged skin,[4] but thicker, fragmented and more soluble by photoaging.[5] Moreover, the rate of collagen synthesis decreases in chronologically aged skin.[6]

As mentioned above, type I collagen is the main molecular constituent of the skin. This macromolecule consists of a set of three polypeptide chains arranged in a specific triple-helical conformation. The fundamental unit called tropocollagen is about of 300 nm in length.[7] Each polypeptide chain is characterized by a tri-aminoacid sequence of Glycine-X-Y type where very frequently X and Y correspond to proline and hydroxyproline, respectively.[8] The tropocollagen units are arranged into fibrils forming in turn collagen fibers. The fiber thickness varies between the two dermis layers, papillary and reticular. The uppermost papillary dermis represents about 10% of the full dermal thickness and is composed of thin collagen fibrils of 20–40 nm in diameter that are packed into collagen fibers of 0.3–3.0 μm in diameter.[9] Below the papillary dermis, the reticular dermis is constituted of thicker collagen fibrils of 60–100 nm in diameter forming 10–40 μm diameter fibers.[10] Bunches of collagen

* *Correspondence to: Olivier Piot, MéDIAN Biophotonique et Technologies pour la Santé, CNRS FRE 3481 MEDyC, SFR Cap-Santé, UFR Pharmacie, Université de Reims Champagne-Ardenne. 51 rue Cognacq-Jay, 51096 Reims Cedex, France. E-mail: olivier.piot@univ-reims.fr*

a *MéDIAN Biophotonique et Technologies pour la Santé, CNRS FRE 3481 MEDyC, SFR Cap-Santé, UFR Pharmacie, Université de Reims Champagne-Ardenne, 51 rue Cognacq-Jay, 51096 Reims, Cedex, France*

b *CNRS FRE 3481 MEDyC, SFR Cap-Santé, UFR Médecine, Université de Reims Champagne-Ardenne, 51 rue Cognacq-Jay, 51096 Reims, Cedex, France*

Copyright © 2013 John Wiley & Sons, Ltd.

fibers can form fiber bundles giving an organized network with noteworthy mechanical properties. The particular triple-helical conformation of collagen and its arrangement into a fiber form are stabilized by means of two different types of hydrogen bonds. More precisely, hydrogen bonds can be direct (between the N–H group of glycine of one chain and the C = O group of X residue of another chain) or more frequently indirect by involving water molecules.[11] Thus, water plays a crucial role in maintaining the collagen structure. Indeed, water molecules are implicated in intra- and inter-chain(s) links by forming hydrogen bonds with carbonyl groups of every aminoacid and/or hydroxyl groups of hydroxyproline residues.[11] Figure 1 extracted from Brodsky *et al.*[11] displays schematically how the water molecules can be linked to collagen chains. It is also important to mention that two water molecules can be hydrogen bonded to the same carbonyl group as shown by Kamer *et al.*[12] in collagen-like peptide model. In addition, it is known that water content changes with age.[13,14] A visible sign is the appearance of dry and flaky skin,[15] which is associated with a degradation of the viscoelastic properties of the skin. For understanding the structural alteration of skin during aging, it is therefore essential to consider the role of water and to analyze the collagen behavior towards water.

Raman spectroscopy has been used on skin tissue for different purposes. Some of the important applications include, in investigations on skin cancer, Ly *et al.*[16] showed the interest of polarized Raman measurements to access some structural information on collagen, in which they highlighted a structural alteration of collagen in peritumoral dermis of carcinoma samples. In addition, Raman spectroscopy was proven to be a sensitive technique to probe the degradation of the stroma by matrix metalloproteinases in basal cell carcinoma.[17,18] The potential of confocal Raman microspectroscopy for monitoring skin hydration has also been demonstrated.[19,20] This non-invasive technique allows in-depth measurements of the water concentration from the surface to several tens of microns below the skin surface. By analyzing Raman vibrations specific of collagen, Zhang *et al.*[21] identified spectral markers of collagen hydration in *ex vivo* experiments led on excised pig skin samples. In this study, Raman microspectroscopy was associated with dynamic vapor sorption measurements for monitoring structural changes of type I collagen. Concerning skin aging, Gniadecka *et al.*[22] carried out a study on photo-aged and chronologically aged skins by using Fourier transform Raman spectroscopy. The authors stated, in photo-aged skin specimens, an increase of water in the tetrahedron structures (using the specific peak at 180 cm^{-1}) and structural alterations of collagen which were associated to a decrease of the hydroxyproline proportion in comparison with non photo-exposed skins. For chronologically aged skin, only minor spectral changes associated to protein conformation were observed. But, in this study, the experiments were performed on the epidermis of skin, mainly constituted of keratin rather than collagen. Finally, *in vivo* cutaneous measurements of water content were compared between young and old populations by using confocal Raman spectroscopy.[23] In this study, the spectral data in the high wavenumber region was exploited and permitted to assess higher water content for old skin than for young skin. The analysis was related to the water content in the papillary dermis since the measurement depths were between 70 and 196 μm below the skin surface.

Our study aims to investigate structural modification of dermal collagen as function of the chronological skin aging by analyzing collagen/water interactions. For this, we use a biophysical approach based on Raman microspectroscopy, and we collected spectral data at the level of the reticular dermis on samples of two different ages (40- and 70-year-old skin specimens) by varying the environmental relative humidity (RH). The analysis was realized on thin tissue sections of skin originating from human donors and using a humidity-controlled chamber implemented on the Raman setup.

Materials and methods

Lyophilized type I collagen and rat tail tendon

Two reference samples, lyophilized type I collagen and rat tail tendon, were used to assess the effect of various RH (0–95%)

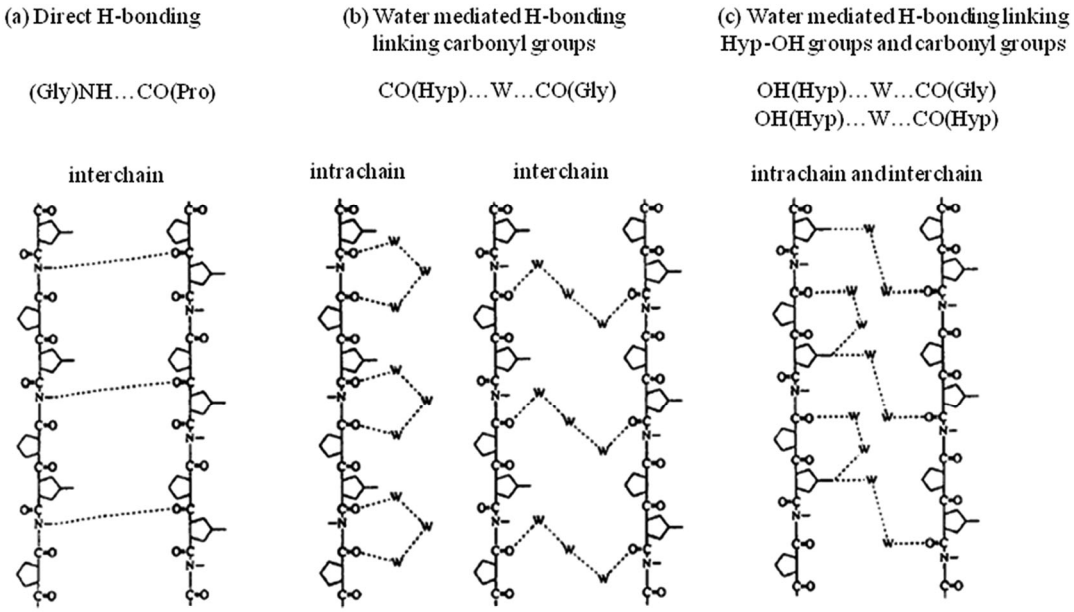


Figure 1. Schematic illustration of the hydrogen bonding patterns stabilizing the collagen structure: (a) direct hydrogen bonding; (b) water mediated hydrogen bonding linking carbonyl groups and (c) water mediated hydrogen bonding linking hydroxyproline OH and carbonyl groups.

Copyright © 2013 John Wiley & Sons, Ltd.

Journal of RAMAN SPECTROSCOPY

values on type I collagen Raman signal. Five μm thin sections of the tail tendon of 2-month-old male rats were deposited on CaF_2 windows (Crystran Ltd., Dorset, UK), which permit to record tissue spectra with reduced parasitic spectral background. Lyophilized type I collagen was prepared from the same rat tail tendons by 0.5M acetic acid extraction, dialysed against distilled water and lyophilization.

Human skin samples

Four skin samples originated from 40 (n = 2) and 70-year-old (n = 2) patients operated for breast cancer resection. The skin specimens taken at the level of breast were considered as normal (no skin pathology) by conventional histology. Thin tissue sections (5 μm) were deposited on CaF_2 windows and analyzed by Raman spectroscopy in imaging mode. For each tissue section, regions of interest (ROI) were defined by selecting areas of about 50×50 μm^2 in the reticular dermis. In each case, ROI were localized at a distance of about 500 μm from the *stratum corneum*, the superficial layer of the skin.

Raman instrumentation

Raman data were collected with a Labram microspectrometer (Horiba Jobin Yvon, Villeneuve d'Ascq, France) and using a 660 nm laser diode as excitation source. A 100X objective (NA = 0.9, Leica) permitted to focus the incident laser beam and to collect the Raman scattering. The laser power on samples was fixed at 30 mW for a laser spot diameter assessed to 2 μm. The Raman scattering was dispersed by a holographic grating of 950 g/mm and analyzed by a Peltier-cooled CCD detector (Andor, Belfast, UK), which permitted to measure simultaneously the intensity of different wavelengths of the spectrum with a spectral resolution of 4 cm^{-1}. The data was collected on a spectral range from 700 to 1720 cm^{-1}. For the recording of spectral images, a motorized stage was used to map defined ROI in point-by-point mode using a step size of 5 μm in the two XY dimensions. The stage temperature was maintained at 20°C during the experiment.

RH controlled setup

A measuring chamber was mounted around the microscope to control the humidity of environmental atmosphere during the acquisition. The RH was adjusted simply by mixing high-humidity air (water vapor saturated air) and dry air and by adjusting the two respective flows (Fig. S1 (Supporting Information)). The system permitted to obtain RH from 0% to 95% with a precision of 2%. Raman analyses were carried out at different RH values of 0, 20, 40, 70 and 95%.

Curve-fitting method

The curve-fitting method was previously applied on Raman data to determine the secondary structure of proteins.[16,24] In our case, curve fitting of the amide I band (1590–1720 cm^{-1}) was performed with Labspec 4 software (Horiba Jobin Yvon, France). Spectra were smoothed using a Savitzky–Golay filter (third order, 11 points), baseline corrected using line segments between the minima points (700, 800, 834, 903, 991, 1018, 1145, 1216, 1362, 1510, 1590 and 1720 cm^{-1}) and normalized on the amide I band (from 1590 cm^{-1} to 1720 cm^{-1}). The six constituent peaks (1605, 1615, 1635, 1658, 1668 and 1688 cm^{-1}) of the amide I band were selected from the second-order derivative of the spectra. Mixed Gaussian/Lorentzian mathematical functions were employed to fit the experimental data; the fitting procedure ran iteratively until the residual error between experimental and modeled curves reached a value ($\chi 2$ as indicated in Labspec) of 10^{-6}. The peak widths at half-height of the elementary peaks were initially set at 20 cm^{-1} and were adjusted during curve fitting within 10 and 40 cm^{-1} as authorized range.

The Matlab randfeatures function

The Matlab randfeatures function (available in MatLab Statistics Toolbox and MatLab Bioinformatics Toolbox) was used to identify the most discriminant Raman vibrations (more precisely, discriminant wavenumbers) between two categories of spectra.[25–27] The randfeatures function randomly selects a subset of 15 wavenumbers and reduces all the spectra from 2479 wavenumbers (in our case) to the 15 selected. The mean spectrum of each category was subtracted from the corresponding spectra. The resulting spectra were then classified using a linear discriminant analysis (LDA). After LDA classification, each spectrum gets percentages of belonging to the two categories. It is considered as well predicted if the percentage of belonging to the correct category is over a Confidence Threshold fixed to 50%. Then, a Performance Threshold, fixed here to 95%, is used to calculate the number of spectra correctly predicted. If this threshold is exceeded, the subset of 15 wavenumbers is retained, and a new iteration starts with 15 other wavenumbers. Note that a wavenumber can be chosen several times. The randfeatures algorithm stops when 1000 subsets of 15 wavenumbers are retained. The cardinality of each wavenumber, also called score, permits to determine its relevancy in the classification. The wavenumbers are sorted following their score in order to determine the most discriminant ones.

Hierarchical clustering analysis

To assess the relevancy of the discriminant wavenumbers identified using the Randfeatures function, hierarchical clustering analysis (HCA) (MatLab Statistics Toolbox) was performed to classify the spectral data on the selected wavenumbers. Inter-spectra distance was calculated using the Euclidean distance, and Ward's algorithm was used for linkage of the spectra into a dendrogram, visualized in the form of a tree-like diagram.

Results

Raman analysis of reference collagens (lyophilized type I collagen and rat tail tendon)

Figure 2 and Figure S2 (Supporting Information) show the average and standard deviation spectra (n = 100 spectra) of lyophilized type I collagen and rat tail tendon, respectively, as function of various RH values from 0% to 95%. Spectra were baseline corrected and normalized on the amide I band (from 1590 cm^{-1} to 1720 cm^{-1}) for comparison. On these spectra, characteristic peaks of type I collagen were found. In particular, the peaks at 856 cm^{-1} and 922 cm^{-1} correspond to the stretching vibration of C–C bond (ν_{C-C}) specific of proline residue ring while the peak at 875 cm^{-1} is assigned to the same vibration mode but for the hydroxyproline residue ring.[21] The peak at 1003 cm^{-1} corresponding to the phenylalanine ring breathing mode[21] is also visible. For these peaks, no variation in intensity as function of RH was observed. In contrast,

Copyright © 2013 John Wiley & Sons, Ltd.

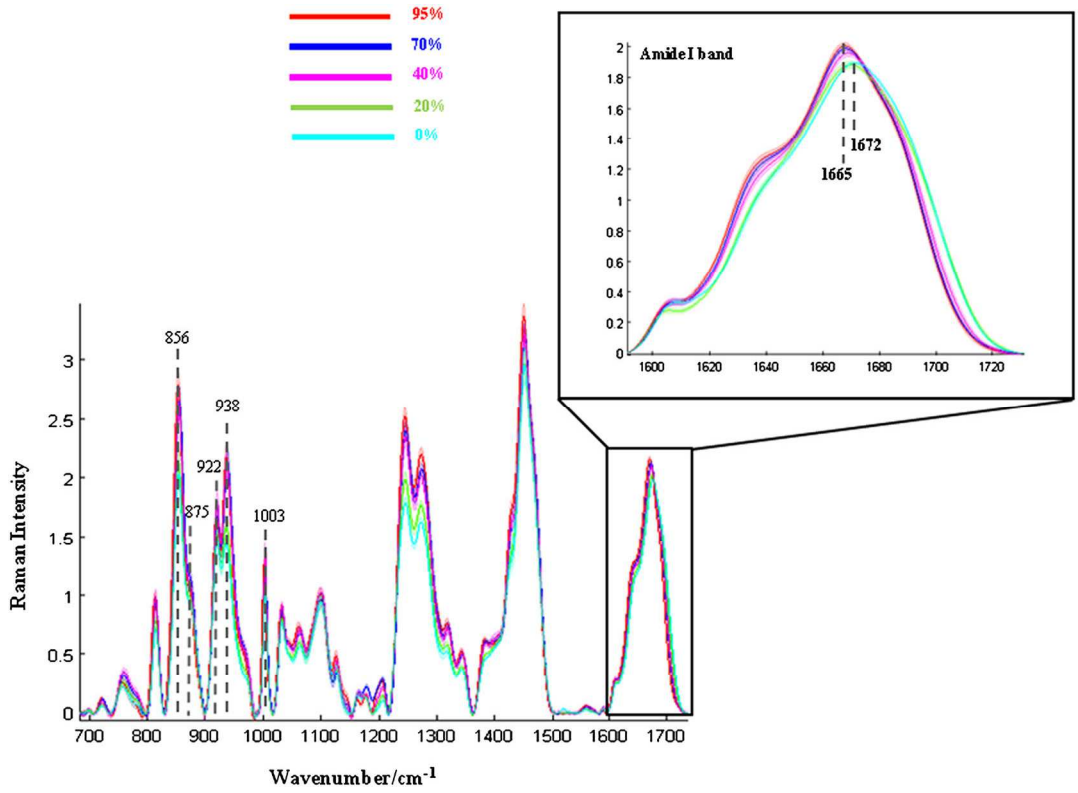


Figure 2. Average spectra and standard deviation (n=100 spectra) of lyophilized type I collagen as function of RH values.

the intensity of the 938 cm^{-1} vibration assigned to skeletal C–C stretching mode of collagen chains according to Zhang *et al.*,[21] increased strongly with the RH (Fig. 2). This vibration was stated to be sensitive to the formation of hydrogen bonds between carbonyl bonds of proline and hydroxyproline residues and water molecules. Thus, this peak can be considered as a potential spectroscopic marker of collagen hydration. To better visualize this particular spectral feature as in Zhang *et al.*,[21] we calculated the ratio of integrated intensities I_{938}/I_{922} as function of RH for both reference samples (Fig. 3). The plot depicts that the ratios increased almost linearly as function of RH, but within the error, no difference appeared between lyophilized collagen and rat tail tendon.

Simultaneously, at the level of amide I and III bands, informative of the protein secondary structure, the spectral profiles appeared to alter when the RH increased. More precisely for the amide I band, a wavenumber shift of the intensity maximum from 1672 cm^{-1} to 1665 cm^{-1} was observed with RH increasing, as also shown by Zhang *et al.*[21] To investigate more precisely this spectral change, curve-fitting procedure was performed for the amide I band. Figure 4 and Figure S3 (Supporting Information) show curve-fitting results as function of RH from 0% to 95% for lyophilized type I collagen and rat tail tendon, respectively. Among the six peaks used for modeling the amide I band, the peaks at 1605 and 1615 cm^{-1} were assigned to vibrations of phenylalanine and tyrosine residues, respectively.[24] The other peaks could correspond to collagen C=O stretching vibrations in different chemical environments. Indeed, the peak around 1685 cm^{-1} can be assigned to non-hydrogen bonded C=O stretching. Weakly hydrogen bonded C=O stretching (i.e. carbonyl group in which only one water molecule is likely to be involved in) could correspond to the component at ~1668 cm^{-1}. With RH increasing, this vibration

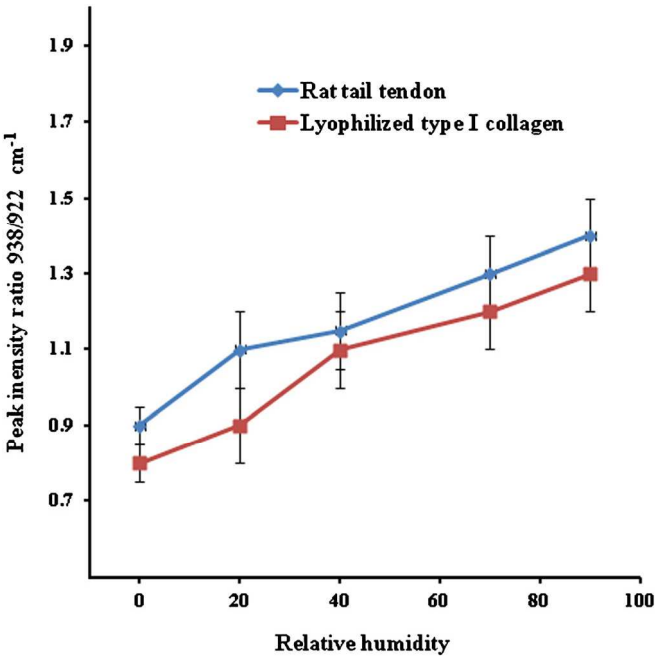


Figure 3. Ratio 938/922 cm^{-1} as function of RH values for rat tail tendon and lyophilized type I collagen.

displays a downshift in wavenumber from 1672 to 1665 cm^{-1} as also shown by Zhang *et al.*,[21] and its intensity decreases strongly. Besides the previous peaks, the component centered on 1658 cm^{-1} could be attributed to strongly hydrogen bonded C=O stretching

Copyright © 2013 John Wiley & Sons, Ltd.

Journal of RAMAN SPECTROSCOPY

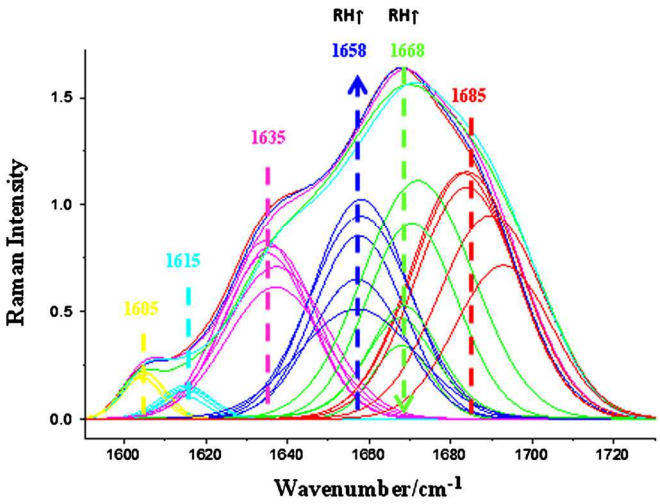


Figure 4. Curve fitting of the amide I band as function of RH values for lyophilized type I collagen. Colour coding of the experimental amide I band: red, blue, pink, green and cyan correspond to the RH values of 95%, 70%, 40%, 20% and 0% respectively.

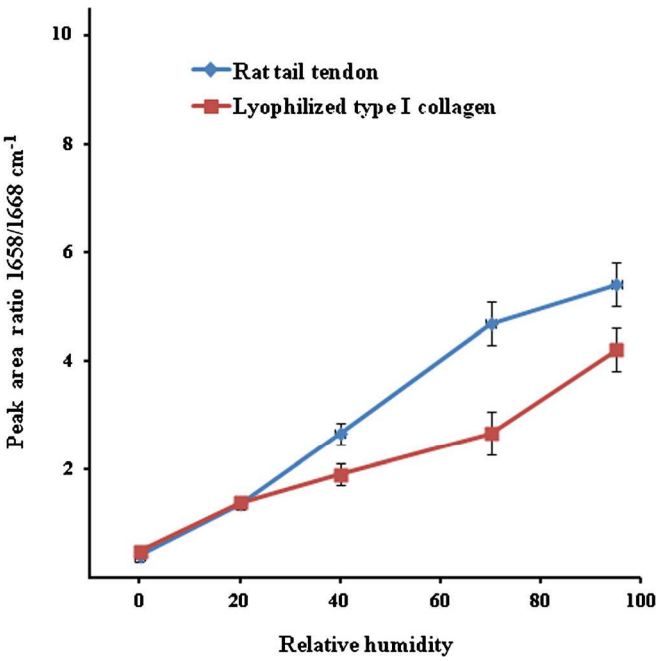


Figure 5. Ratio 1658/1668 cm^{-1} as function of RH values for rat tail tendon and lyophilized type I collagen.

(i.e. carbonyl group in which two water molecules can be involved in). With RH increasing, this peak intensity increases strongly because the carbonyl groups are likely to be involved in hydrogen bonds with two water molecules rather than only one, according the schematic water/peptide interactions depicted by Kramer *et al.*[12] Finally, the C=O group of proline can be linked by direct inter-chain bonding to the NH group of glycine, without involving water molecules.[11] In this case, the C=O stretching would be around 1637 cm^{-1}. At this wavenumber, a minor contribution of O–H bending vibration of water is also possible. With RH increasing, there is no visible intensity change for this last vibration.

Following this assignment, the 1658/1668 cm^{-1} would be a direct marker of water/collagen interactions since it reflects the transition from weak to strong hydrogen bonding of the carbonyl groups when RH increases. Inside the collagen triple helix, the more important spacing between the polypeptide chains, the more water molecules can enter and interact with carbonyl groups via the formation of hydrogen bonds. Consequently, the comparison of this ratio between different skin samples could reflect differences of compactness of the collagen fiber bundles. The ratios of the integrated intensities 1658/1668 cm^{-1} as a function of RH values were calculated from Raman spectra of the lyophilized type I collagen and of the rat tail tendon (Fig. 5). The ratios increase with the RH values, and contrary to the 938/922 cm^{-1} ratio previously calculated (Fig. 3), it appeared that the 1658/1668 cm^{-1} values are significantly higher for the rat tail tendon than for the lyophilized collagen.

From these observations revealing that 938/922 cm^{-1} and 1658/1668 cm^{-1} ratios constitute Raman markers of collagen hydration, we focused our investigation on skin samples by comparing the evolution of the Raman signal as function of RH, for specimens of two different ages.

Raman analysis of skin samples (40- and 70-year-old dermal skin samples)

Figures 6 and Fig. S4 (Supporting Information) show the average and standard deviation spectra (n=100 spectra) of 40- ((sample #1 in Fig. 6a); (sample #2 in Fig. S4a)); and 70- ((sample #1 in Fig. 6b); (sample #2 in Fig. S4b)) year-old reticular dermal skin samples as function of varying RH from 0% to 95%.

The peaks at 856 cm^{-1} and 922 cm^{-1}, and 875 cm^{-1} corresponding to the proline and hydroxyproline residues, respectively, were clearly observable. In addition, as for the reference specimens of type I collagen, no intensity variation of these peaks as function of RH was visible, and no difference between the two ages appeared. Concerning the other peaks specific of the collagen studied above, the intensity variation of the ν_{C-C} peak at 938 cm^{-1} (Fig. 6) and the wavenumber downshift of the amide I maximum from 1672 cm^{-1} to 1665 cm^{-1} were observable when RH increased, as well as for the lyophilized type I collagen and the rat tail tendon. Besides these collagen vibrations, from a visual comparison of spectra collected on skin and those of reference collagen, additional signal appeared for skin. Indeed, weak intensity peaks at 725 cm^{-1} and 1160 cm^{-1} were detected and attributed to elastin and GAG, respectively.[28,29]

Figure S5 (Supporting information) shows curve-fitting results of the amide I band as a function of RH for 40- (Fig. S5a and b) and 70- (Fig. S5c and d) year-old reticular dermis sections, respectively. Six constitutive vibration-modeled bands were recovered at 1605, 1615, 1635, 1658, 1668 and 1685 cm^{-1}, similar to the curve fitting of reference collagen spectra.

In order to compare the behavior of the skin samples towards water, the 938/922 cm^{-1} and 1658/1668 cm^{-1} ratios were calculated and plotted as function of RH (Figs. 7 and 8, respectively). As expected, their values increased with RH. For the 938/922 cm^{-1} intensity ratio, within the error, no significant difference between the skin samples of different ages appeared due to the signal variability. However, the ratio values are higher than for the reference collagen products. For example, at an RH of 95%, the ratio achieved a value of around 1.8 for skin samples and only 1.4 for the rat tail tendon. This difference could be due to cutaneous constituents such as GAG, known to be involved in tissue hydration, or/and to structural difference of type I collagen between human skin and

Copyright © 2013 John Wiley & Sons, Ltd.

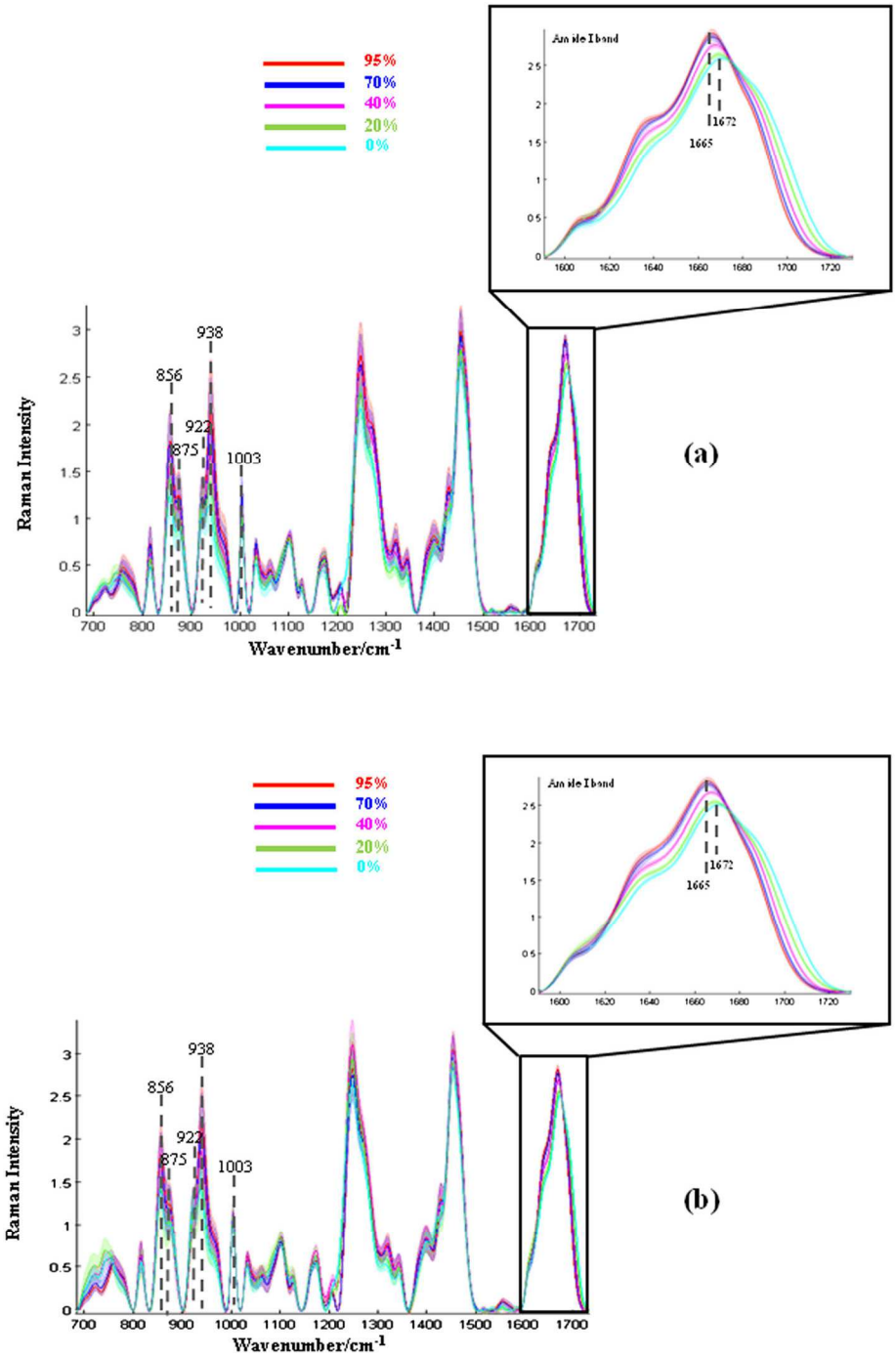


Figure 6. Average spectra and standard deviation (n=100 spectra) of 40- (a: sample #1 of this age) and 70- (b: sample #1 of this age) year-old reticular dermal skin as function of various RH values.

reference products. More interestingly, the ratio of integrated intensities 1658/1668 cm^{-1}, calculated from curve fitting, permitted to distinguish the 40- and 70-year-old skins with much higher values at elevated RH for the oldest skins ($I_{1658}/I_{1668} \approx 10$ at 95% RH) compared to the youngest ($I_{1658}/I_{1668} \approx 4$) and to the reference products ($I_{1658}/I_{1668} \approx 6$ for the tendon). Figure S6 (Supporting information) displays a superimposition of the average spectra of 40- and 70-year-old reticular dermal skin samples for 95% RH. Differences of profile at the level of the amide I band are clearly visible between the samples of different ages, as more strongly emphasized by the curve-fitting procedure. By searching a biological meaning to the Raman data, this result could reflect a difference of compactness of the collagen fiber bundles of the reticular dermis between the 40-and 70-year-old dermis samples, as explained later in the discussion.

As visible on the Raman spectra (mean and standard variance spectra in Figs. 6 and S4), or on ratio plots (standard deviation in Figs. 7 and 8), the studied tissue samples presented a variability that cannot be overlooked. In order to find statistically discriminant vibrations between the four skin samples, randfeatures algorithm

Copyright © 2013 John Wiley & Sons, Ltd.

Journal of RAMAN SPECTROSCOPY

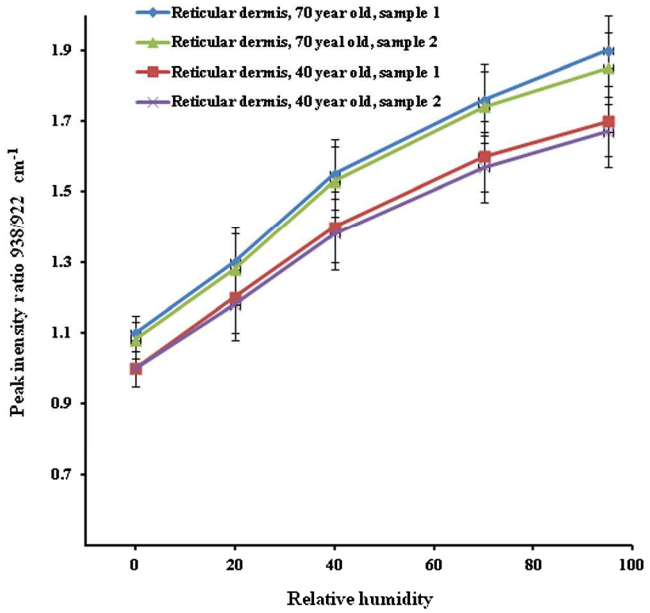


Figure 7. Ratio 938/922 cm^{-1} as function of RH values for 40- and 70-year-old reticular dermal skin samples.

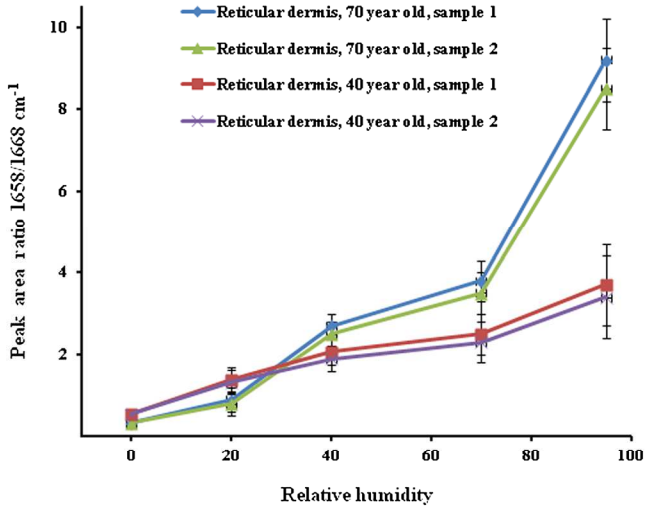


Figure 8. Ratio 1658/1668 cm^{-1} as function of RH values for 40- and 70-year-old reticular dermal skin samples.

was applied to skin data collected with the high RH value (95%). A total of 400 spectra, equally distributed between the 40- and 70-year-old skin samples, were randomly selected for this processing. The six most discriminating wavenumbers were retained. These wavenumbers appeared in the amide I band, which is consistent with the curve-fitting results (Fig. 9). The HCA dendrogram in Fig. S7 (Supporting Information) showed the high discriminant value of these vibrations since the two groups of spectra assigned to the 40- and 70-year-old skin samples were clearly separated. The same procedure was also performed for the other RH values (data not shown), and the same wavenumbers in the amide I band were systematically identified as discriminant.

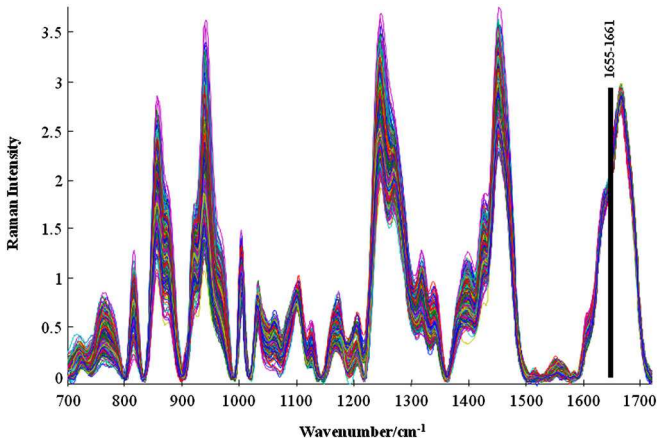


Figure 9. Discriminant Raman wavenumbers, identified using randfeatures algorithm, between 40- and 70-year-old dermal skin samples for the RH value of 95%.

Discussion and conclusion

First, our study shows that Raman microspectroscopy is a potential tool to investigate some bio-physiological mechanisms involved in skin aging and to assess their effects or extents. In the present study, the Raman technique was used to highlight differences of behavior towards water, of dermal collagen in four skin samples of different age. This non-destructive biophotonic technique has the advantage to reveal *in situ* information on molecular structures or ordering modified with the chronological aging. The biological or molecular meaning of the spectral data is a hard task and remains hypothetic mainly because of the lack of equivalent technique allowing to access non-destructively structural data at a microscopic scale.

In our approach, two spectral ratios likely to be sensitive to collagen/water interactions were computed in order to assess their discriminant potential between skin samples of two different ages. The first ratio (I_{938}/I_{922}) did not permit to distinguish between the 40- and 70-year-old skin samples. Although Zhang *et al.*[21] demonstrated that the 938 cm^{-1} vibration, assigned to skeletal C–C stretching mode of collagen chains, was sensitive to the formation of hydrogen bonds between carbonyl groups of proline and hydroxyproline residues and water molecules, other reports[30,31] mentioned also the assignment of this vibration to the C–C stretching of proline ring. Thus, this first ratio appears not to be specific enough for our comparison. On the contrary, the second ratio (I_{1658}/I_{1668}) permitted to clearly differentiate the skin samples of 40 and 70 years. As explained previously in results, this ratio can be used as marker of water/collagen interactions. And differences of values between skin samples could reflect differences of compactness of the collagen fiber bundles. According to this hypothesis, our data are in agreement with other studies that have demonstrated thinner diameters of collagen fiber bundles in young skins,[32] presenting consequently a more compact collagen structure. The evolution of the collagen network structure with aging also involves several complex mechanisms such as modification of covalent cross-linking bonds or glycation process. Multiphoton microscopy appears as an innovative imaging modality to assess the extent of these phenomena[33,34] due to the specific collagen signal in second harmonic generation (SHG). In the future, combining Raman and SHG approaches could permit to investigate further the molecular alterations of collagen with aging.

Copyright © 2013 John Wiley & Sons, Ltd.

Among Raman markers of skin hydration, the ratio of the integrated intensities $I_{(3350–3550)}/I_{(2910–2965)}$ (OH stretching vibration of water (ν_{OH})/CH_3 stretching vibration attributed to skin protein ($\nu_{C–H\ of\ CH3}$)) has been considered in several renowned publications as a reference indicator to measure water in skin and also to assess the moisturizing power of dermo-cosmetic agents at the level of the *stratum corneum* and the epidermis.[19,20] We tested this method to determine the water concentration in our dermal skin samples. Figure S8 (Supporting Information) depicts the spectra, in the high wavenumbers region, of 40- (Figs. S8a and b) and 70- (Figs. S8c and d) year-old skin samples as function of RH. As expected, the ratio $\nu_{OH}/\nu_{C–H\ of\ CH3}$ increased with RH, but no difference between the two ages at RH of 95% ((Fig. S9 (Supporting Information)) was visible contrary to our results based on the I_{1658}/I_{1668} ratio. The integrated band between 3350 and 3550 cm^{-1} includes O–H vibrations of water molecules in both free or bound forms indiscriminately. So, for a precise assessment of collagen hydration, it is important to consider the content in bound water molecules as probed by the vibrations at 1658 and 1668 cm^{-1}, specific of collagen C=O stretching when hydrogen bonds form between water and carbonyl groups. However, the high wavenumbers spectral range is also informative since the interaction between water and protein can be monitored by a wavenumber shift of the $\nu_{C–H\ of\ CH3}$ band from 2937 to 2942 cm^{-1} with RH increasing, which represents unfolding of the protein and exposure of its aliphatic amino acid chains to the surrounding water.[35] In addition, this spectral region is commonly used to assess the water content, which was calculated from the intensity ratio of water vibration (OH stretching: 3350–3550 cm^{-1}) and protein vibration (CH stretching: 2910–2965 cm^{-1}).[19,23] Comparison with the *in vivo* study of Nakagawa *et al.*[23] shows that the dermal tissue samples present a much lower content water in our study. Indeed, the *in vivo* water content was calculated around 75% (mass%), at a RH of 50%,[23] whereas in our study, the water content in the reticular dermis ranged from 10% to 28% for RH values from 0% to 95%. But, it is important to note that different parts of the dermis were probed. We focused our investigation on the reticular dermis, whereas Nakagawa *et al.*[23] collected data at the level of the upper papillary dermis. This difference of hydration between the papillary dermis and the reticular dermis originates mainly from a difference of GAGs content. Indeed, the reticular dermal layer is essentially constituted of thicker collagen with a very small content in GAGs; thus, tissue hydration is principally due to collagen hydration. Contrary, the papillary dermis presents a high content in GAGs that will bind water molecules.[36,37]

This feasibility study shows that the Raman comparison of skin collagen hydration permits to distinguish between 40- and 70-year-old skin samples. Similarly photo-aging effects can be studied using the same approach. Development for in routine applications requires analyzing more cutaneous samples in order to take into account intra- and inter-individual(s) variability. The ability of Raman microspectroscopy to investigate collagen/water interactions and to shed light on structural modifications of collagen in reticular dermis with aging makes this biophotonic technique particularly exciting and promising for applications in dermatology and cosmetics.

Acknowledgements

T.T. Nguyen is a recipient of a doctoral fellowship and CIRGE Région Champagne-Ardenne. The authors would like to thank Ligue contre le Cancer (Comité de l'Aisne), INSERM PNR Imagerie, CNRS Projets Exploratoires Pluridisciplinaires and CNRS PIR Longévité et Vieillissement for financial support. We also thank Dr Ali Tfayli for the equipement permitting to adjust the environmental relative humidity.

Supporting information

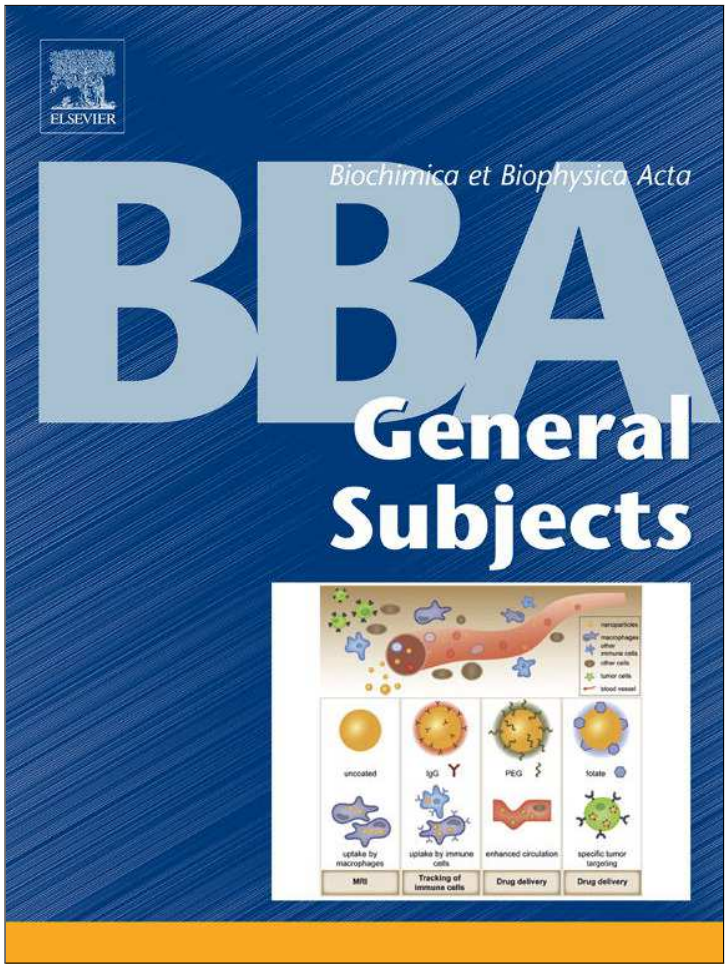
Supporting information may be found in the online version of this article.

References

[1] S. Shuster, M. M. Black, E. McVitie, *Br. J. Dermatol.* **1975**, *93*, 639.
[2] C. R. Lovell, K. A. Smolenski, V. C. Duance, N. D. Light, S. Young, M. Dyson, *Br. J. Dermatol.* **1987**, *117*, 419.
[3] R. M. Lavker, P. S. Zheng, G. Dong, *J. Invest. Dermatol.* **1987**, *88*, 44.
[4] J. Uitto, *Clin. Geriatr. Med.* **1989**, *5*, 127.
[5] M. Gniadecka, R. Gniadecki, J. Serup, J. Søndergaard, *J. Invest. Dermatol.* **1994**, *102*, 362.
[6] J. Varani, M. K. Dame, L. Rittie, S. E. G. Fligiel, S. Kang, G. J. Fisher, J. J. Voorhees, *Am. J. Pathol.* **2006**, *168*, 1861.
[7] J. P. Borel, *Eurobiologiste* **1991**, *25*, 247.
[8] K. Beck, B. Brodsky, *J. Struct. Biol.* **1998**, *122*, 17.
[9] I. A. Brown, *J. Anat.* **1972**, *113*, 159.
[10] F. H. Silver, J. W. Freeman, D. DeVore, *Skin Res. Technol.* **2001**, *7*, 18.
[11] J. Bella, B. Brodsky, H. M. Berman, *Structure* **1995**, *3*, 893.
[12] R. Z. Kramer, M. G. Venugopal, J. Bella, P. Mayville, B. Brodsky, H. M. Berman, *J. Mol. Biol.* **2000**, *301*, 1191.
[13] R. O. Potts, E. M. Buras, D. A. Chrisman, *J. Invest. Dermatol.* **1984**, *82*, 97.
[14] S. Richard, B. Querleux, J. Bittoun, O. Jolivet, I. Idy-Peretti, O. de Lacharriere, J. L. Leveque, *J. Invest. Dermatol.* **1993**, *100*, 705.
[15] K. P. Wilhelm, *Arch. Dermatol.* **1991**, *127*, 1806.
[16] E. Ly, O. Piot, A. Durlach, P. Bernard, M. Manfait, *Appl. Spectrosc.* **2008**, *62*, 1088.
[17] J. Varani, Y. Hattori, Y. Chi, T. Schmidt, P. Perone, M. E. Zeigler, D. J. Fader, T. M. Johnson, *Br. J. Cancer* **2000**, *82*, 657.
[18] T. Yucel, A. Mutnal, K. Fay, S. E. Fligiel, T. Wang, T. Johnson, S. R. Baker, J. Varani, *Exp. Mol. Pathol.* **2005**, *79*, 151.
[19] P. J. Caspers, G. W. Lucassen, E. A. Carter, H. A. Bruining, G. J. Puppels, *J. Invest. Dermatol.* **2001**, *116*, 434.
[20] L. Chrit, P. Bastien, B. Biatry, J.-T.Simonnet, A. Potter, A. M. Minondo, F. Flament, R. Bazin, G. D. Sockalingum, F. Leroy, M. Manfait, C. Hadjur, *Biopolymers* **2007**, *85*, 359.
[21] Q. Zhang, K. L. Andrew Chan, G. Zhang, T. Gillece, L. Senak, D. J. Moore, R. Mendelsohn, C. R. Flach, *Biopolymers* **2011**, *95*, 607.
[22] M. Gniadecka, O. F. Nielsen, S. Wessel, M. Heidenheim, D. H. Christensen, H. C. Wulf, *J. Invest. Dermatol.* **1998**, *111*, 1129.
[23] N. Nakagawa, M. Matsumoto, S. Sakai, *Skin Res. Technol.* **2010**, *16*, 137.
[24] N. C. Maiti, M. M. Apetri, M. G. Zagorski, P. R. Carey, V. E. Anderson, *J. Am. Chem. Soc.* **2004**, *126*, 2399.
[25] L. Li, D. M. Umbach, P. Terry, J. A. Taylor, *Bioinformatics* **2004**, *20*, 1638.
[26] H. Liu, H. Motoda, Feature Selection for Knowledge Discovery and Data Mining, Kluwer Academic Publishers: Norwell, MA, USA, **1998** .
[27] D. T. Ross, U. Scherf, M. B. Eisen, C. M. Perou, C. Rees, P. Spellman, V. Iyer, S. S. Jeffrey, M. V. de Rijn, M. Waltham, A. Pergamenschikov, J. C. F. Lee, D. Lashkari, D. Shalon, T. G. Myers, J. N. Weinstein, D. Botstein, P. O. Brown, *Nat. Genet.* **2000**, *24*, 227.
[28] N. Mainreck, S. Brézillon, G. D. Sockalingum, F.-X. Maquart, M. Manfait, Y. Wegrowski, *J. Pharm. Sci.* **2011**, *100*, 441.
[29] L. Debelle, A. J. Alix, S. M. Wei, M. P. Jacob, J. P. Huvenne, M. Berjot, P. Legrand, *Eur. J. Biochem.* **1998**, *258*, 533.
[30] M. Diem, R. S. Bhatnagar, M. E. Druyan, V. Renugopalakrishnan, *Biopolymers* **1984**, *23*, 2955.
[31] Y. Xiao, M. Guo, P. Zhang, G. Shanmugam, P. L. Polavarapu, M. S. Hutson, *Biophys. J.* **2008**, *94*, 1359.
[32] M. C. Branchet, S. Boisnic, C. Frances, C. Lesty, Robert, *Arch. Gerontol. Geriatr.* **1991**, *13*, 1.
[33] A. A. Ghazaryan, P.-S. Hu, S.-J. Chen, H.-Y. Tan, C.-Y. Dong, *J. Dermatol. Sci.* **2012**, *65*, 189.
[34] V. Lutz, M. Sattler, S. Gallinat, H. Wenck, R. Poertner, F. Fischer, *Skin Res. Technol.* **2012**, *18*, 168.
[35] M. Gniadecka, O. Faurskov Nielsen, D. H. Christensen, H. C. Wulf, *J. Invest. Dermatol.* **1998**, *110*, 393.
[36] R. J. Beyth, L. A. Culp, *Mech. Ageing Dev.* **1985**, *29*, 151.
[37] M. D. Willen, J. M. Sorrell, C. C. Lekan, B. R. Davis, A. I. Caplan, *J. Invest. Dermatol.* **1991**, *96*, 968.

Copyright © 2013 John Wiley & Sons, Ltd.

Provided for non-commercial research and education use.
Not for reproduction, distribution or commercial use.



This article appeared in a journal published by Elsevier. The attached copy is furnished to the author for internal non-commercial research and education use, including for instruction at the authors institution and sharing with colleagues.

Other uses, including reproduction and distribution, or selling or licensing copies, or posting to personal, institutional or third party websites are prohibited.

In most cases authors are permitted to post their version of the article (e.g. in Word or Tex form) to their personal website or institutional repository. Authors requiring further information regarding Elsevier's archiving and manuscript policies are encouraged to visit:

http://www.elsevier.com/authorsrights

Author's personal copy

Biochimica et Biophysica Acta 1830 (2013) 3525–3531

Contents lists available at SciVerse ScienceDirect

Biochimica et Biophysica Acta

journal homepage: www.elsevier.com/locate/bbagen

Probing non-enzymatic glycation of type I collagen: A novel approach using Raman and infrared biophotonic methods

Marie Guilbert [a], Georges Said [a], Teddy Happillon [a], Valérie Untereiner [a], Roselyne Garnotel [a], Pierre Jeannesson [a], Ganesh D. Sockalingum [a,*]

[a] Equipe MéDIAN-Biophotonique et Technologies pour la Santé, Université de Reims Champagne-Ardenne, Unité MEDyC FRE CNRS/URCA 3481, UFR de Pharmacie, 51 rue Cognacq-Jay, 51096 Reims, France

ARTICLE INFO

Article history:
Received 9 October 2012
Received in revised form 8 January 2013
Accepted 13 January 2013
Available online 1 February 2013

Keywords:
Type I collagen
Glycation
Advanced Glycation End products
Fourier-transform infrared microspectroscopy
Raman microspectroscopy

ABSTRACT

Background: Non-enzymatic glycation is the main post-translational modification of long-life proteins observed during aging and physiopathological processes such as diabetes and atherosclerosis. Type I collagen, the major component in matrices and tissues, represents a key target of this spontaneous reaction which leads to changes in collagen biomechanical properties and by this way to tissue damages.
Methods: The current study was performed on *in vitro* glycated type I collagens using vibrational microspectroscopies, FT-IR and Raman, to highlight spectral features related to glycation effect.
Results and conclusions: We report a conservation of the triple-helical structure of type I collagen and noticeable variations in the exposure of proline upon glycation. Our data also show that the carbohydrate band can be a good spectroscopic marker of the glycation level, correlating well with the fluorescent AGEs formation with sugar addition.
General significance: These non-invasive and label-free methods can shed new light on the spectral features of glycated collagens and represent an effective tool to study changes in the extracellular matrix observed *in vivo* during aging or on the advent of a pathological situation.

© 2013 Elsevier B.V. All rights reserved.

1. Introduction

Type I collagen is the most abundant extracellular matrix protein in the human body. It is assembled in triple-helical structure, and cross-linking between triple helices allows it to form a fibrillar network [1]. Because of its long lifespan which can be variable, about 15-years in skin, collagen I undergoes post-translational modifications during aging or pathological conditions such as diabetes mellitus [2] and more recently in cancer processes [3,4]. The main modification is a non-enzymatic glycation [5,6], resulting in the fixation of reducing agents, such as aldose sugars, essentially on lysine residues of type I collagen. In the field of cancer, scaffolds of glycated collagens have been used to study the impact of this post-translational modification on tumor cell proliferation [7] and migration [8]. Further, in tissue engineering strategies, glycated collagen has also been shown to be of great interest for preparing cartilage constructs [9,10], mimicking diabetic wound healing [11] or evaluating the effect of mechanical constraints on cellular behavior [12].

At the mechanistic level, glycation leads to the formation of Schiff bases that are transformed into Amadori products. After complex rearrangements, the Amadori compounds give rise to irreversible products called advanced glycation end products (AGEs) [3]. Among these end products, N^{ε}-carboxymethyllysine (CML) and pyrraline have been found to be non-fluorescent and non-cross-linking, while pentosidine and crossline have been described as fluorescent and cross-linking compounds [13–15]. AGEs contribute to changes in the collagen properties such as loss of the triple helix solubility and flexibility, resulting in an increase of its rigidity [1]. Consequently, *via* these structural and molecular modifications of collagen triple helix, its enzymatic digestivity is less efficient with aging and in pathological situations like diabetes [16]. Glucose is the major blood circulating sugar in the human body but it exhibits less reducing properties than agents such as ribose, glyceraldehydes or fructose [4,17]. Although these latter compounds are less involved in *in vivo* protein glycation processes, they are able to generate elevated AGEs levels and are therefore used for *in vitro* collagen glycation [1].

Protein glycation evaluation is currently assessed by conventional spectrometric, chromatographic or immunohistochemical methods. To detect post-translational modifications of the collagen and to provide a quantitative determination of resulting cross-links, mass spectrometry and high performance liquid chromatography have

* Corresponding author at: FRE CNRS/URCA no. 3481, UFR Pharmacie, 51 rue Cognacq-Jay, 51096 Reims Cedex, France. Tel.: +33 3 26 91 35 53; fax: +33 3 26 91 35 50.
E-mail addresses: marie.guilbert@univ-reims.fr (M. Guilbert), georgessaid@gmail.com (G. Said), teddy.happillon@gmail.com (T. Happillon), valerie.untereiner@univ-reims.fr (V. Untereiner), roselyne.garnotel@univ-reims.fr (R. Garnotel), pierre.jeannesson@univ-reims.fr (P. Jeannesson), ganesh.sockalingum@univ-reims.fr (G.D. Sockalingum).

0304-4165/$ – see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.bbagen.2013.01.016

Author's personal copy

been largely used [14,16]. However, these methods need sample reduction and protein hydrolysis into amino-acids [18] and as such cannot be used directly on *ex vivo* or *in vivo* tissues. For immunostaining of tissues, AGEs' accumulation can be also quantified but it gives the AGEs' content from all proteins present in the tissue section, but not from a single protein like collagen [19,20]. For these reasons, the most current technique used to quantify glycation rate of collagen is the fluorimetric assay that permits to detect fluorescent AGEs such as pentosidine [21]. However, even if this technique provides AGEs' quantification, it does not inform on the molecular changes associated to the overall AGEs.

In the present study, we investigated the non-enzymatic glycation process of type I collagen using a biophotonic complementary approach based on infrared and Raman vibrational microspectroscopies. These robust and non-invasive methods can provide information on both the composition and the structure, and are sensitive to the environment of proteins [22,23]. They are well adapted to highlight the impact of non-enzymatic glycation directly on type I collagen. In particular, Roy *et al.* [24] have previously demonstrated that Fourier-transform infrared (FT-IR) used in the Attenuated Total Reflection (ATR) mode allowed to quantify *in vitro* glycation of collagen gels, which was clearly correlated with the fluorescent-AGEs' accumulation. In our study, in addition to the FT-IR microspectroscopic approach, we have used Raman microspectroscopy to complement the analysis of native, fibrillar, and non-pepsined type I collagen *in vitro*-glycated with glucose or ribose. These vibrational spectroscopic techniques are label-free, direct, rapid, and non-destructive approaches that are applied here to explore to what extent *in vitro* glycation of type I collagen impact on protein structural features. Both methodologies are sensitive to molecular and conformational structural information and are complementary.

2. Material and methods

2.1. Extraction of type I collagen and in vitro glycation

Fibrillar type I collagen was extracted as previously described [25]. Briefly, acid-soluble type I collagen was obtained from tail tendons of Sprague-Dawley rats by 0.5 M acetic acid extraction then purified by dialysis against distilled water and lyophilized. This native, purified, and non-pepsined collagen was stored at −80 °C until further use. For preparation of *in vitro* glycated collagens, the lyophilized extracted native type I collagen was incubated, under sterile conditions, at 37 °C for 14 days in 0.15 M of phosphate buffer pH 7.4 containing 5 or 50 mM of glucose (Sigma-Aldrich, L'isle d'Abeau Chesnes, France) or for 5 days in the same buffer containing 50 mM of ribose (Sigma-Aldrich). Both types of glycated collagens were extensively dialyzed against distilled water and then lyophilized.

2.2. Estimation of electrophoretic properties

Collagens were solubilized at 2 mg/ml in 0.018 M of acetic acid (v/v) and denatured by heating for 2 min at 90 °C. Electrophoretic properties of native and *in vitro*-glycated collagens were estimated by 5% sodium dodecyl sulfate-polyacrylamide gel electrophoresis (SDS-PAGE). Gels were stained with Coomassie Brilliant Blue R250 for revealing characteristic type I collagen chains.

2.3. Fluorescent AGEs detection

To quantify the glycation level in the *in vitro*-glycated collagens, AGEs-specific fluorescence was measured using a Shimadzu RF-5000 fluorescence spectrophotometer (Shimadzu, Marne la Vallée, France) at 380 nm excitation and 440 nm emission, after solubilizing collagens at 2 mg/ml in 0.018 M of acetic acid (v/v). Data are presented as mean ± SEM. The values were analyzed with Kruskall–Wallis followed by Mann–Whitney tests. Statistical significance was set at $p<0.05$.

2.4. FT-IR microspectroscopy

Micro-FT-IR acquisitions of lyophilized native and *in vitro*-glycated type I collagens were performed on thin pellets, prepared by mixing 1 mg of collagen with 99 mg of potassium bromide (Sigma-Aldrich) and pressed in the form of thin discs at 10,000–15,000 psi during 10 min. Infrared transmission spectra were recorded between 4000 and 700 cm^{-1} in the point mode using a 100×100 μm^2 aperture with a Spotlight 300 microscope coupled to a Spectrum 100 FT-IR spectrometer (both from Perkin Elmer, Courtaboeuf, France). For each spectrum, 128 scans were averaged at 4 cm^{-1} resolution and data were pre-processed by baseline correction with the elastic correction method, vector normalized on the total spectral range, and offset corrected to bring the minimum absorbance value to 0, using the Opus 6.5 software (Bruker Optics, Marne la Vallée, France).

2.5. Raman microspectroscopy

Micro-Raman spectra were recorded directly on lyophilized native and *in vitro*-glycated type I collagens without further preparation using a Labram microspectrometer (Horiba Jobin-Yvon SAS, Lille, France) operating with a 785 nm diode laser (Spectra-Physics, Les Ulis, France) and equipped with a 100× long working distance objective (numerical aperture of 0.8). For each measurement, two acquisitions of 80 s were averaged at 4 cm^{-1} spectral resolution over the spectral window 350 to 1800 cm^{-1}. All spectra were pre-processed by smoothing when necessary (7 points Savitzky–Golay smoothing), and then baseline corrected, vector normalized, and offset corrected, using the LabSpec 4.18 software (Horiba Jobin-Yvon).

2.6. Spectral data processing

Spectral data were compared using Hierarchical Clustering Analysis (HCA) that allows regrouping spectra on a minimal distance criterion and so, according to their homogeneity. For this purpose, Euclidian distances and Ward's algorithm implemented in the Opus 6.5 software (Bruker Optics) were used and the results were displayed *via* a dendrogram.

3. Results

3.1. Evaluation of in vitro glycation efficiency of type I collagen: electrophoretic mobility and fluorescent AGEs accumulation

In order to validate the accumulation of AGEs, we investigated the electrophoretic mobility of *in vitro*-glycated collagens and the fluorescence intensity associated with AGEs production after incubation of extracted type I collagen with glucose (5 or 50 mM) or ribose (50 mM). As shown in Fig. 1A, all *in vitro*-glycated collagens exhibit characteristic bands of native type I collagen, showing α_1 and α_2 chains. However, a decrease of the electrophoretic mobility can be observed in glycated collagens compared to the native one. This effect is more marked with ribose-glycated collagen, confirming the higher reducing effect of ribose on collagen I compared to glucose. Fig. 1B displays the results of spectrofluorimetric assay of non-glycated and *in vitro*-glycated type I collagens. A proportional increase in the fluorescence of collagens as a function of glycation level can be observed, indicating that the content of fluorescent AGEs increases with the concentration and nature of the reducing sugar. For instance, a two-fold increase of AGEs accumulation is observed in 50 mM ribose-glycated collagen compared to 50 mM glucose-glycated collagen, confirming the higher reducing effect of ribose.

Author's personal copy

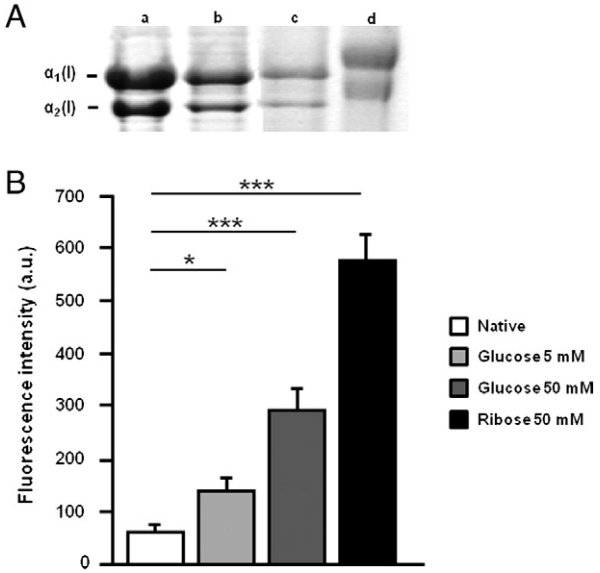


Fig. 1. Biochemical properties of *in vitro*-glycated collagens. (A) SDS-PAGE of collagen samples, 25 μg of either native non-glycated (a) or *in vitro* glycated 5 mM glucose (b), 50 mM glucose (c), and 50 mM ribose (d) type I collagens were analyzed on 5% polyacrylamide gels under reducing conditions. (B) Spectrofluorimetric analysis was performed on these different collagens solubilized at 2 mg/ml in 18 mM acetic acid (v/v) to detect AGEs-specific fluorescence using a spectrofluorimeter (Shimadzu model RF-5000, France) at $\lambda_{ex} = 380$ nm and $\lambda_{em} = 440$ nm. Values are the mean of three independent experiments (*$p<0.05$ and ***$p<0.001$ are compared to native non-glycated collagen, mean ± SEM).

These two experimental results confirm the efficiency of the *in vitro* glycation process used in this study.

3.2. FT-IR microspectroscopic analysis of in vitro glycated collagens

Table 1 summarizes the FT-IR frequency positions and a tentative assignment of the major functional groups of native type I collagen with the changes observed with *in vitro* glycation in the characteristic peaks. It can be observed that the mean FT-IR spectrum of native non-glycated type I collagen exhibits the characteristic spectral features previously reported [26]. Comparison of the mean FT-IR spectra of *in vitro* glycated collagens with the control (Fig. 2) shows no significant change in the position of the Amide I band, representative of the protein secondary structure. This indicates that the triple helix structure of the molecule is well-conserved upon glycation. The up-shift observed in the Amide II band (1552 cm^{-1} to 1546 cm^{-1}) for ribose-glycated sample may reflect discrete conformational changes in tertiary structure of collagen such as changes of amino-acid exposure. Upon glycation, the most important changes are observed in the 1100–900 cm^{-1} spectral range, which corresponds to the characteristic absorption region of carbohydrates. In the native type I collagen, two distinct peaks are observed corresponding to total protein and carbohydrate signals (1082 and 1032 cm^{-1} respectively). The signal of carbohydrates increases further until it becomes predominant over the total protein signal. In order to verify if these spectral modifications observed in this region can be a pertinent spectroscopic marker of the glycation degree, we have computed the intensity ratio of the 1032 cm^{-1} band with respect to the Amide I band at 1660 cm^{-1}. Our results clearly show that the I_{1030}/I_{1660} ratio increases (Fig. 3) of the same order as that observed for AGEs' detection by spectrofluorimetry.

Table 1
Changes in vibrational band frequencies (cm^{-1}) of characteristic FT-IR bands due to *in vitro* glycation of type I collagen.

	Native (control)	Glucose 5 mM	Glucose 50 mM	Ribose 50 mM
νNH	3326	3326	3326	3326
νNH	3084	3084	3083	3083
νCH_2	2955	2955	2955	2955
νCH_2	2880	2880	2880	2880
Amide I, νC═O	1660	1660	1660	1661
Amide II, δNH_2	1552	1551	1550	1546
δCH_2	1453	1453	1452	1451
δCH_2, CH_3	1403	1403	1403	1403
δCH_2	1338	1338	1338	1338
Amide III, δNH_2	1240	1240	1239	1237
C – N	1205	1205	1205	1205
C – OH	1082	1082	1082	1082
C – OH	1032	1032	1032	1032

An FT-IR spectrum contains at least a thousand of variables and another way of comparing spectral data is to use multivariate statistical analysis. Here, for instance, we applied HCA based on Ward's algorithm and Euclidean distance calculation, to determine spectral similarity *via* a plot of the heterogeneity values between spectra. The HCA of mean infrared spectra of native and glycated collagens using the total spectral range is displayed in Fig. 4A. As can be seen, it shows a clear discrimination between native and glycated collagens. For glycated samples, collagens are clearly differentiated as a function of the reducing sugar or the concentration used. The HCA performed on the carbohydrate region (Fig. 4B) confirms a clear-cut discrimination of the different collagens.

3.3. Raman microspectroscopic analysis of in vitro glycated collagens

In a complementary way, samples were also analyzed by Raman microspectroscopy in order to study molecular groups which may be active in Raman but not in FT-IR or active in both. The rationale is that while FT-IR spectra give information on the Amide I, corresponding to the stretching coordinate of C═O bonds (νC═O) and Amide II corresponding to the deformation coordinate of NH_2 groups (δNH_2) bands Raman spectra exhibit Amide I (νC═O) and Amide III (δNH_2) features as well as side chain amino-acids.

The Raman mean spectra of both native type I collagen and *in vitro* glycated forms, shown in Fig. 5, exhibit the characteristic spectral features previously reported [27]. Spectra show that *in vitro* glycation does not alter the band positions of all functional groups since no frequency shift is observed. The Amide I (1673 cm^{-1}) and Amide III (1244 cm^{-1}) bands do not exhibit change, confirming the conservation of the triple-helix structure with *in vitro* glycation as yet demonstrated with FT-IR data. However, at the 533 cm^{-1} frequency position, it appears that the intensity of this Raman band increases upon glycation, suggesting conformational changes such as a skeletal deformation which may be due to the presence of cross-linking. Furthermore, other important changes in peak intensities are revealed upon glycation, especially for residues of prolines (Pro) and hydroxyprolines (HyPro) which are respectively the second and the third amino-acids in the type I collagen triple-helix composition. For each peak of interest, the ratio intensities ($I_{peak\ of\ interest}/I_{1451}$) were calculated, the peak at 1451 cm^{-1} (assigned to $-CH_2$ groups) remaining highly constant upon glycation. The observed changes are represented in Fig. 6, showing an increase of both the 394 cm^{-1} (HyPro) and 875 cm^{-1} (HyPro) peaks and a decrease of both the 470 cm^{-1} (Pro) and 1556 cm^{-1} (HyPro/Pro) peaks with glycation. These data suggest changes in the amino-acid exposure and protein backbone with type I collagen glycation. An HCA of Raman mean spectra (Fig. 7) using the total spectral range (350–1800 cm^{-1}) confirms a clear separation of glycated type I collagens from the native one and also a good discrimination as a function of the sugar or concentration used for *in vitro* glycation conditions.

Author's personal copy

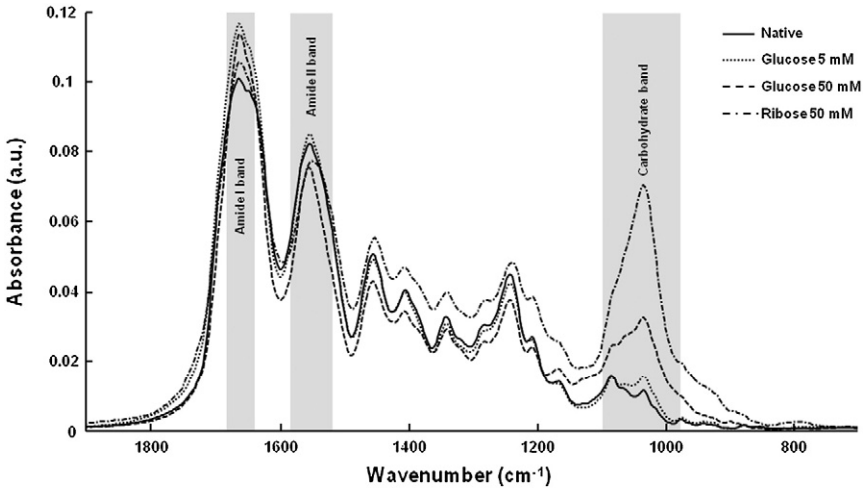


Fig. 2. FT-IR analysis of *in vitro*-glycated collagens. Comparison between FT-IR spectra of native non-glycated and *in vitro*-glycated (glucose 5 mM, glucose 50 mM and ribose 50 mM) collagens. Spectra represent means of three independent measurements and were normalized in the 1900–700 cm^{-1} range. Gray zones indicate interest spectral regions to discriminate samples according to *in vitro* glycation.

4. Discussion

In this study, IR and Raman microspectroscopies have been applied as probing non-destructive biophotonic techniques for studying native and glycated type I collagens. Results obtained demonstrate that upon *in vitro* glycation i) the conservation of the triple-helical structure of type I collagen; ii) noticeable variations in the exposure of proline, the second mostly abundant amino-acid in type I collagen chains; and iii) the level of glycation can be appreciated in a semi-quantitative manner. Concerning the third point, our results obtained at the microspectroscopic level largely confirm the previous observations highlighted by Roy et al. [24] using ATR-FTIR showing a direct proportionality between glycation level *via* the carbohydrate signal increase and the fluorescent-AGEs' formation.

This spectral analysis of non-enzymatic glycation is highly representative of the physiopathological range of glycation which impacts type I collagen *in vivo*. Indeed, the glucose concentrations used here are representative of the sugar content in mammalian circulating blood (5 mM) [28] and in severe diabetes (50 mM) [29,30]. In this pathology, it has been demonstrated that high glucose level leads to the formation of AGEs on proteins such as collagen or hemoglobin [31,32]. In this study, ribose being the highest reducing sugar, it permits a faster and higher glycation process compared to glucose. Even if it is not a physiological sugar, because of its higher reducing effect, it has been used for convenience reasons to glycate faster and therefore represents a positive control of protein glycation *in vitro* [16].

In this spectroscopic study, spectra of native and *in vitro* glycated type I collagens were recorded from thin pellets of the freeze-dried proteins. The rationale for using freeze-dried form is that this procedure concentrates the proteins and provides well-resolved data for the spectroscopic analysis. Indeed, we have previously analyzed the native and *in vitro* glycated collagens in solution, but spectra were less informative compared to spectra from freeze-dried collagens since they exhibited only the Amide I band (data not shown). Furthermore, FT-IR spectra on the same solutions were completely dominated by water bands. Thus, the approach on freeze-dried collagens allows a comparative study in both infrared and Raman microspectroscopies,

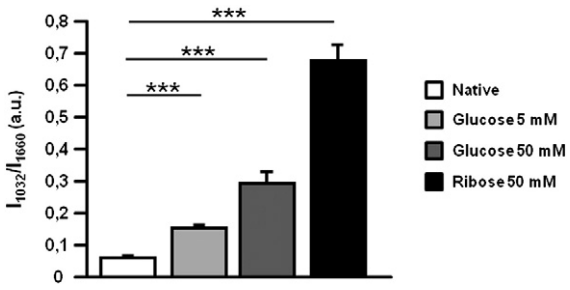


Fig. 3. Ratio of the FT-IR intensities (I) of the carbohydrate and Amide I bands. The 1032 cm^{-1} and 1660 cm^{-1} peaks correspond at the FT-IR signal of C–OH bonds specific to carbohydrates and the FT-IR signal of Amide I band respectively. The ratio of the intensities of these two peaks (I_{1032}/I_{1660}) was calculated.

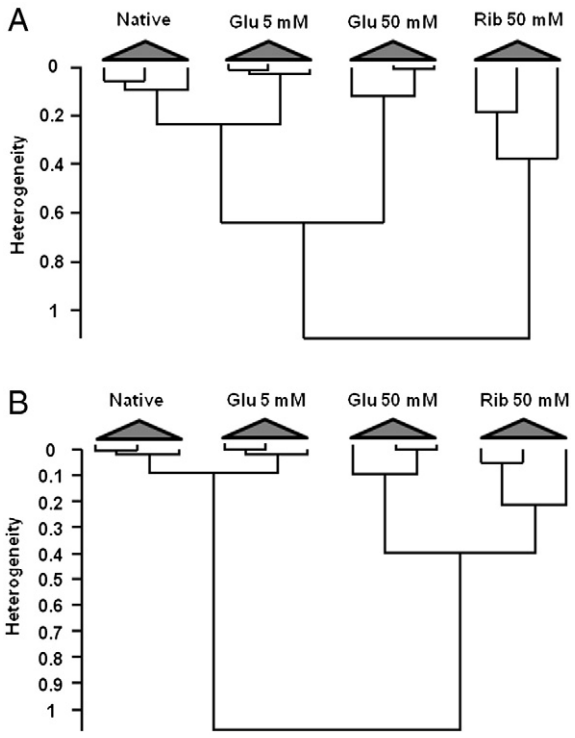


Fig. 4. Hierarchical cluster analysis (HCA) on FT-IR spectra from native and *in vitro*-glycated collagens. HCA was calculated with the mean spectra of three independent experiments for each condition, on the 4000–700 cm^{-1} total spectral range (A) and on the spectral window 1100 to 900 cm^{-1} corresponding to the specific fingerprint of carbohydrates (B). All mean spectra with a high homogeneity degree are clusterized together.

Author's personal copy

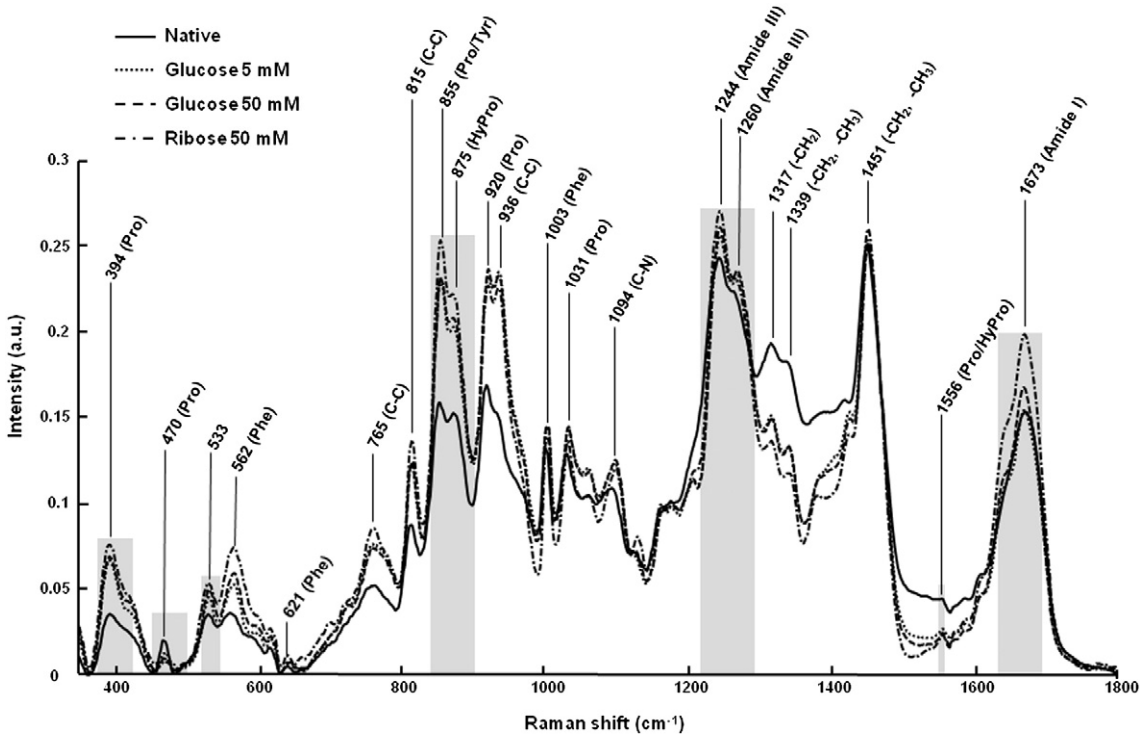


Fig. 5. Raman spectral analysis of *in vitro*-glycated collagens. Comparison between Raman spectra of native non-glycated and *in vitro*-glycated (glucose 5 mM, glucose 50 mM and ribose 50 mM) collagens. Spectra represent means of three independent measurements and were normalized in the 350–1800 cm^{-1} range. Gray zones indicate interest spectral regions to discriminate samples according to *in vitro* glycation.

allowing the detection of minor spectral changes that occur upon *in vitro* glycation. In addition, it has been demonstrated by another biophotonic approach using pulsed terahertz spectroscopy, that the freeze-dried form is representative of proteins close to the hydrated form in terms of signal and can be considered as native protein [33].

In molecules with different elements of symmetry (such as in the triple-helical structure of collagen), certain bands may be active in IR, Raman, both or neither. That is why Raman and FT-IR micro-spectroscopies are complementary in the analysis of complex proteins such as collagen. In order to obtain a full vibrational fingerprint, samples were analyzed by both FT-IR and Raman microspectroscopies. For example, in macromolecules or complex biomolecular systems, FT-IR gives a very high signal of carbohydrates, which is less discernable on a Raman spectrum. In contrast, the Raman spectrum of proteins better highlights amino-acid signatures and reveals the characteristic peaks of prolines and hydroxyprolines. These two residues are among the most important amino-acids present in the composition of both α_1 and α_2 chains of human collagens. Indeed, proline and hydroxyproline residues represent respectively the second (10%) and the third (8%) most abundant amino-acids in the protein sequence of type I collagen [34]. Their Raman activities exhibit important intensity differences between the native non-glycated (control) and glucose-glycated collagens and to a lesser extent for the ribose-glycated counterpart. These spectral modifications suggest that upon *in vitro* glycation, the molecule undergoes a slight supra-molecular rearrangement, resulting in variations

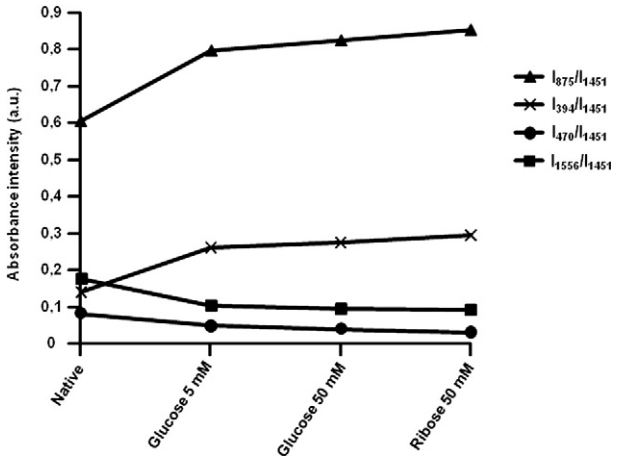


Fig. 6. Analysis of Raman signals from proline and hydroxyproline residues from native and *in vitro*-glycated collagens. The 1451 cm^{-1} peak corresponds to the Raman signal of $-CH_2$ and $-CH_3$ groups, the 470 cm^{-1} peak corresponds to the main Raman signal of prolines, the 394 cm^{-1} and the 875 cm^{-1} peaks correspond to the main Raman signal of hydroxyprolines, and the 1556 cm^{-1} peak corresponds to the Raman signal of both prolines and hydroxyprolines. The ratio of the intensities of each of these peaks (I_{394}/I_{1451}; I_{470}/I_{1451}; I_{875}/I_{1451} and I_{1556}/I_{1451}) was calculated.

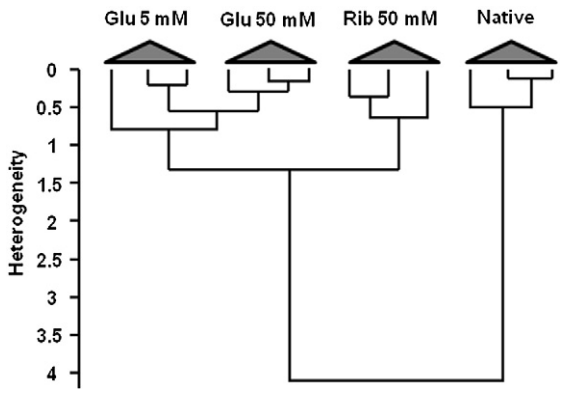


Fig. 7. Hierarchical cluster analysis (HCA) on Raman spectra from native and *in vitro*-glycated collagens. HCA was calculated with the mean spectra of three independent experiments for each condition, on the total spectral range (350–1800 cm^{-1}). All mean spectra with a high homogeneity degree are clusterized together.

Author's personal copy

in the proline and hydroxyproline residue exposures. These observations, that can be associated with the increased AGEs formation induced by the glycation process, corroborate with the up-shift in the IR Amide II band, especially for ribose. The profiles of both the IR and Raman Amide I bands, well identified as representative of the secondary structure of proteins, exhibit neither intensity change nor band shift upon glucose or ribose exposure. This indicates that the triple-helical structure of type I collagen is not perturbed by the *in vitro* glycation. A recent study has confirmed similar observations with fructose, another reducing sugar, using circular dichroism and X-ray diffraction [35].

By integrating the area of the carbohydrate band (1100–900 cm^{-1}), a semi-quantitative appreciation of the glycation level was possible. This spectral region therefore appears like a good IR spectroscopic marker of collagen I *in vitro* glycation. Recently, Roy et al. [24] have identified this spectral zone to characterize glycation levels between pre-glycated and post-glycated collagen gels. Further, we show that this spectroscopic marker is well correlated with the fluorescent-AGEs' quantification by conventional spectrofluorimetric assay. In parallel, at the tissue level, we recently reported that the intensity variations of this marker band can be highly indicative of changes in carbohydrate contents in glycoproteins such as human secreted mucins [36].

This study demonstrates the potential of vibrational microspectroscopies for detecting changes due to post-translational modifications of proteins. In addition, FT-IR and Raman microspectroscopies are rapid, sensitive, non-destructive and straight-forward techniques for analyzing sample composition and structure of proteins [37]. These biophotonic approaches are under constant progress for developing innovative tools for biomaterial molecular characterization. They are robust and well appropriate for evaluating structural changes in proteins and can be alternative and complementary tools to conventional techniques. In physiopathological conditions such as aging or diabetes, a major foreseen interest could be the direct evaluation of collagen glycation in tissues, as a complement to classical immunohistological detection assays.

5. Conclusions

The present study demonstrates that vibrational microspectroscopies allow detecting the effects of *in vitro* glycation on type I collagen, the major component in human tissues. We compared native type I collagen and *in vitro* glycated forms using reducing sugars and this model helped to highlight spectroscopic markers related to changes in the molecule upon glycation. Used in a complementary manner, FT-IR and Raman techniques can provide characteristic and specific spectral information of this non-enzymatic reaction. The spectral assignment for the triple-helical structure, that remained unchanged upon glycation, the proline and hydroxyproline attributions, and the carbohydrate band, which appear like a robust marker for a semi-quantitative appreciation of the glycation effect, are all very informative spectral ranges for studying post-translational modifications of matrix proteins. These techniques could be powerful tools for monitoring collagen changes occurring during aging or during the advent of pathological situations such as diabetes in a rapid, direct, quantitative, and label-free manner.

Acknowledgements

This study was supported by grants of the Institut National du Cancer (INCa), the Cancéropôle Grand-Est and FEDER/CPER. M.G. is a recipient of a doctoral fellowship from Région Champagne-Ardenne and G.S. from Cancéropôle Grand-Est. We are thankful to Pr Hamid Morjani for his kind assistance with the spectrofluorimetric analysis. We thank the Plateforme IBiSA "Imagerie Cellulaire et Tissulaire" for providing the equipments.

References

[1] N.C. Avery, A.J. Bailey, The effects of the Maillard reaction on the physical properties and cell interactions of collagen, Pathol. Biol. 54 (2006) 387–395.
[2] D.G. Dyer, J.A. Dunn, K.E. Bailie, T.J. Lyons, D.R. McCance, J.W. Baynes, Accumulation of Maillard reaction products in skin collagen in diabetes and aging, J. Clin. Invest. 91 (1993) 2463–2469.
[3] N.A. Ansari, Z. Rasheed, Non-enzymatic glycation of proteins: from diabetes to cancer, Biomed. Khim. 56 (2010) 168–178.
[4] J. Takino, S. Yamagishi, M. Takeuchi, Cancer malignancy is enhanced by glyceraldehyde-derived advanced glycation end-products, J. Oncol. (2010) 739852.
[5] R.G. Paul, A.J. Bailey, Glycation of collagen: the basis of its central role in the late complications of ageing and diabetes, Int. J. Biochem. Cell Biol. 28 (1996) 1297–1310.
[6] M.E. Francis-Sedlak, S. Uriel, J.C. Larson, H.P. Greisler, D.C. Venerus, E.M. Brey, Characterization of type I collagen gels modified by glycation, Biomaterials 30 (2009) 1851–1856.
[7] B. Bartling, M. Desole, S. Rohrbach, R.E. Silber, A. Simm, Age-associated changes of extracellular matrix collagen impair lung cancer cell migration, FASEB J. 23 (2009) 1510–1520.
[8] G. Said, M. Guilbert, E. Millerot-Serrurot, L. Van Gulick, C. Terryn, R. Garnotel, P. Jeannesson, Impact of carbamylation and glycation of collagen type I on migration of HT1080 human fibrosarcoma cells, Int. J. Oncol. 40 (2012) 1797–1804.
[9] R. Roy, A.L. Boskey, L.J. Bonassar, Non-enzymatic glycation of chondrocyte-seeded collagen gels for cartilage tissue engineering, J. Orthop. Res. 26 (2008) 1434–1439.
[10] S.V. Eleswarapu, J.A. Chen, K.A. Athanasiou, Temporal assessment of ribose treatment on self-assembled articular cartilage constructs, Biochem. Biophys. Res. Commun. 414 (2011) 431–436.
[11] H. Liao, J. Zakhaleva, W. Chen, Cells and tissue interactions with glycated collagen and their relevance to delayed diabetic wound healing, Biomaterials 30 (2009) 1689–1696.
[12] B.N. Mason, A. Starchenko, R.M. Williams, L.J. Bonassar, C.A. Reinhart-King, Tuning three-dimensional collagen matrix stiffness independently of collagen concentration modulates endothelial cell behavior, Acta Biomater. 9 (2013) 4635–4644.
[13] N. Ahmed, Advanced glycation endproducts—role in pathology of diabetic complications, Diabetes Res. Clin. Pract. 67 (2005) 3–21.
[14] K. Mikulíková, A. Eckhardt, S. Pataridis, I. Miksík, Study of posttranslational non-enzymatic modifications of collagen using capillary electrophoresis/mass spectrometry and high performance liquid chromatography/mass spectrometry, J. Chromatogr. A 1155 (2007) 125–133.
[15] F.J. Tessier, The Maillard reaction in the human body. The main discoveries and factors that affect glycation, Pathol. Biol. 58 (2010) 214–219.
[16] S.F. Tian, S. Toda, H. Higashino, S. Matsumura, Glycation decreases the stability of the triple-helical strands of fibrous collagen against proteolytic degradation by pepsin in a specific temperature range, J. Biochem. 120 (1996) 1153–1162.
[17] M.J. Kent, N.D. Light, A.J. Bailey, Evidence for glucose-mediated covalent cross-linking of collagen after glycosylation in vitro, Biochem. J. 225 (1985) 745–752.
[18] T.J. Sims, N.C. Avery, A.J. Bailey, Quantitative determination of collagen crosslinks, Methods Mol. Biol. 139 (2000) 11–26.
[19] F.A. Shamsi, A. Partal, C. Sady, M.A. Glomb, R.H. Nagaraj, Immunological evidence for methylglyoxal-derived modifications in vivo. Determination of antigenic epitopes, J. Biol. Chem. 273 (1998) 6928–6936.
[20] M. Bendayan, Immunocytochemical detection of advanced glycated end products in rat renal tissue as a function of age and diabetes, Kidney Int. 54 (1998) 438–447.
[21] K. Mikulíková, A. Eckhardt, J. Kunes, J. Zicha, I. Miksík, Advanced glycation end-product pentosidine accumulates in various tissues of rats with high fructose intake, Physiol. Res. 57 (2008) 89–94.
[22] F.S. Parker, Applications of Infrared, Raman and Resonance Raman Spectroscopy in Biochemistry, Plenum Press, NewYork, 1983.
[23] N. Mainreck, S. Brézillon, G.D. Sockalingum, F.X. Maquart, M. Manfait, Y. Wegrowski, Rapid characterization of glycosaminoglycans using a combined approach by infrared and Raman microspectroscopies, J. Pharm. Sci. 100 (2011) 441–450.
[24] R. Roy, A. Boskey, L.J. Bonassar, Processing of type I collagen gels using nonenzymatic glycation, J. Biomed. Mater. Res. A 93 (2010) 843–851.
[25] E. Millerot-Serrurot, M. Guilbert, N. Fourré, W. Witkowski, G. Said, L. Van Gulick, C. Terryn, J.M. Zahm, R. Garnotel, P. Jeannesson, 3D collagen type I matrix inhibits the antimigratory effect of doxorubicin, Cancer Cell Int. 13 (2010) 10–26.
[26] C. Petibois, G. Gouspillou, K. Wehbe, J.P. Delage, G. Délérisn, Analysis of type I and IV collagens by FT-IR spectroscopy and imaging for a molecular investigation of skeletal muscle connective tissue, Anal. Bioanal. Chem. 386 (2006) 1961–1966.
[27] S. Jaisson, S. Lorimier, S. Ricard-Blum, G.D. Sockalingum, C. Delevallée-Forte, G. Kegelaer, M. Manfait, R. Garnotel, P. Gillery, Impact of carbamylation on type I collagen conformational structure and its ability to activate human polymorphonuclear neutrophils, Chem. Biol. 13 (2006) 149–159.
[28] R.J. Clark, P.M. McDonough, E. Swanson, S.U. Trost, M. Suzuki, M. Fukuda, W.H. Dillmann, Diabetes and the accompanying hyperglycemia impairs cardiomyocyte calcium cycling through increased nuclear O-GlcNAcylation, J. Biol. Chem. 278 (2003) 44230–44237.
[29] S.V. Pizzo, M.A. Lehrman, M.J. Imber, C.E. Guthrow, The clearance of glycoproteins in diabetic mice, Biochem. Biophys. Res. Commun. 101 (1981) 704–708.

Author's personal copy

[30] A.K. Death, E.J. Fisher, K.C. McGrath, D.K. Yue, High glucose alters matrix metalloproteinase expression in two key vascular cells: potential impact on atherosclerosis in diabetes, Atherosclerosis 168 (2003) 263–269.
[31] R. Meerwaldt, T. Links, C. Zeebregts, R. Tio, J.L. Hillebrands, A. Smit, The clinical relevance of assessing advanced glycation endproducts accumulation in diabetes, Cardiovasc. Diabetol. 7 (2008) 29.
[32] A. Lapolla, A. Mosca, D. Fedele, The general use of glycated haemoglobin for the diagnosis of diabetes and other categories of glucose intolerance: still a long way to go, Nutr. Metab. Cardiovasc. Dis. 21 (2011) 467–475.
[33] A.G. Markelz, A. Roitberg, E.J. Heilweil, Pulsed terahertz spectroscopy of DNA, bovine serum albumin and collagen between 0.1 and 2.0 THz, Chem. Phys. Lett. 320 (2000) 42–48.
[34] M. Nassa, P. Anand, A. Jain, A. Chhabra, A. Jaiswal, U. Malhotra, V. Rani, Analysis of human collagen sequences, Bioinformation 8 (2012) 26–33.
[35] R. Usha, S.M. Jaimohan, A. Rajaram, A.B. Mandal, Aggregation and self assembly of non-enzymatic glycation of collagen in the presence of amino guanidine and aspirin: an in vitro study, Int. J. Biol. Macromol. 47 (2010) 402–409.
[36] A. Travo, O. Piot, R. Wolthuis, C. Gobinet, M. Manfait, J. Bara, M.E. Forgue-Lafitte, P. Jeannesson, IR spectral imaging of secreted mucus: a promising new tool for the histopathological recognition of human colonic adenocarcinomas, Histopathology 56 (2010) 921–931.
[37] K. Belbachir, R. Noreen, G. Gouspillou, C. Petibois, Collagen types analysis and differentiation by FTIR spectroscopy, Anal. Bioanal. Chem. 395 (2009) 829–837.

RESUME en français

La spectroscopie vibrationnelle est une technologie permettant de générer une grande quantité de données très informatives quant à la composition moléculaire des échantillons analysés. Lorsqu'elle est couplée à des méthodes chimiométriques de traitement et de classification de données, elle devient un outil très performant pour l'identification de structures et sous-structures des échantillons. Appliqué dans le domaine du biomédical, cet outil présente alors un fort potentiel pour le diagnostic de maladie. C'est dans ce cadre qu'ont été réalisés les travaux de ce manuscrit. Dans une première étude relevant du développement algorithmique, un algorithme automatique de classification non supervisée (basé sur les Fuzzy C-Means) et récemment implémenté au sein du laboratoire pour apporter une aide au diagnostic de cancers cutanés par imagerie infrarouge, a été amélioré afin de i) considérablement réduire le temps nécessaire à son exécution ii) augmenter la qualité des résultats obtenus sur les données infrarouge et iii) étendre son champs d'application à des données réelles et simulées, habituellement employées dans la littérature. Cet outil a été testé sur des données infrarouge acquises sur 16 échantillons de cancers cutanés (BCC, SCC, maladie de Bowen et mélanomes), et sur 49 jeux de données réels et simulés. Les résultats obtenus ont montré la capacité de ce nouvel algorithme à estimer des partitions proches de la réalité quelque soit le type de données étudié. La seconde étude de ce manuscrit avait pour but de mettre au point un outil chimiométrique autonome d'aide au diagnostic de la leucémie lymphoïde chronique par spectroscopie Raman. Dans ce travail, des traitements numériques et l'algorithme de classification supervisée Support Vector Machines, ont été appliqués à des données acquises sur des cellules sanguine de 27 témoins et 49 patients présentant une leucémie lymphoïde chronique. Les résultats de classification obtenus ont montré une sensibilité de 80% et une spécificité de 100% dans la détection de la maladie.

TITRE en anglais

Vibrational spectroscopies coupled with numerical multivariate analyzes as an aid to diagnose skin cancers and chronic lymphocytic leukemia

RESUME en anglais

Vibrational spectroscopy is a technology able to record a large amount of molecular information from studied samples. Coupled with chemometrics and classification methods, vibrational spectroscopy is an efficient tool to identify sample structures and substructures. When applied to the biomedical field, this tool shows a high potential for disease diagnosis. It is in this context that the works presented in this thesis have been realized. In a first study, dealing with algorithmic development, an automatic and unsupervised classification algorithm (based on the Fuzzy C-Means) and developed by our laboratory in order to help for skin cancer diagnosis using IR spectroscopy, was improved in order to i) reduce the computational time needed to realize clustering, ii) increase results quality obtained on infrared data, iii) and extend its application fields to simulated and real datasets, commonly used in the literature. This tool has been tested on 16 infrared spectral images of skin cancers (BCC, SCC, Bowen's disease and melanoma), and 49 real and simulated datasets. The obtained results showed the ability of this new algorithm to estimate realistic data partitions regardless the considered dataset. The second study of this work aimed at developing an independent chemometric tool to assist for chronic lymphocytic leukemia diagnosis by Raman spectroscopy. In this second work, different numerical preprocessing steps and a supervised classification algorithm, Support Vector Machines, have been applied on data recorded on blood cells coming from 27 healthy persons and 49 patients with chronic lymphocytic leukemia. The classification results showed a sensitivity of 80% and a specificity of 100% in the disease diagnosis.

DISCIPLINE

Informatique Biophysique

MOTS-CLES

Classification supervisée, Support Vector Machines, classification non supervisée, Fuzzy C-Means, spectroscopie infrarouge, spectroscopie Raman, cancers cutanés, leucémie lymphoïde chronique.

KEYWORDS

Supervised classification, Support Vector Machines, unsupervised classification, Fuzzy C-Means, infrared spectroscopy, Raman spectroscopy, skin cancers, chronic lymphocytic leukemia

INTITULE ET ADRESSE DE L'UNITE DE RECHERCHE

MéDIAN-Biophotonique et Technologies pour la Santé, Université de Reims Champagne-Ardenne, FRE CNRS 3481 MEDyC, UFR de Pharmacie, 51 rue Cognacq-Jay, 51096 REIMS cedex, France.