

# THÈSE DE DOCTORAT

Pour obtenir le grade de

**DOCTEUR DE L'UNIVERSITÉ DE REIMS CHAMPAGNE-ARDENNE**

*Discipline* : **SCIENCES DE LA VIE ET DE LA SANTE**

*Spécialité* : Biophysique

Présentée et soutenue publiquement par

**MICHAEL FERÉ**

Le 18 décembre 2018

---

**M3S – Développement de la spectroscopie Raman en cytopathologie : Application au diagnostic de la leucémie lymphoïde chronique**

---

Thèse dirigée par **OLIVIER PIOT**

## JURY

M. Olivier PIOT	Professeur	UFR Pharmacie, Reims	<b>Directeur de thèse</b>
M. Erik GOORMAGHTIGH	Professeur	Campus de la Plaine, Bruxelles	<b>Rapporteur</b>
M. Ludovic DUPONCHEL	Professeur	Université de Lille1, Villeneuve d'Ascq Cedex	<b>Président</b>
M. Jacques KLOSSA	Docteur	Société TRIBVN, Châtillon	<b>Examineur</b>
Mme Anne QUINQUENEL	Docteur	Centre hospitalier universitaire, Reims	<b>Examineur</b>
M. Cyril GOBINET	Docteur	UFR Pharmacie, Reims	<b>Examineur</b>



---

## **M3S – Développement de la spectroscopie Raman en cytopathologie : Application au diagnostic de la leucémie lymphoïde chronique**

---

---

En cytopathologie, le diagnostic de routine repose sur l'utilisation de coloration ou d'immunomarquage. Le développement des techniques optiques vibrationnelles, telles que la diffusion Raman, offre l'opportunité d'accéder à des informations sur la composition moléculaire globale de l'échantillon, sans en altérer l'intégrité. L'objectif de cette thèse est d'évaluer les capacités diagnostiques de la spectroscopie Raman dans le contexte de la leucémie lymphoïde chronique. Au-delà de la preuve de concept, nous avons cherché à intégrer les caractéristiques d'une application clinique pour construire un classifieur prédictif robuste à partir de données Raman collectées sur frottis sanguins non colorés. Dans une première étape, nous avons montré qu'il était possible d'homogénéiser les données acquises dans différentes conditions expérimentales. Ensuite, nous avons établi une stratégie d'entraînement et de validation permettant de créer des modèles stables et insensibles au sur-apprentissage, et adaptés aux exigences cliniques.

---

spectroscopie Raman ; leucémie lymphoïde chronique ; analyse des données ; diagnostic de la leucémie lymphoïde chronique ; du laboratoire au chevet du patient

---

### **M3S - Development of Raman spectroscopy in cytopathology: Application to the diagnosis of chronic lymphocytic leukaemia**

---

In cytopathology, routine diagnosis is based on the use of staining or immunolabelling. The development of optical vibration techniques, such as Raman scattering, provides the opportunity to access information on the global molecular composition of the sample, without altering its integrity. The objective of this thesis is to evaluate the diagnostic capabilities of Raman spectroscopy in the context of chronic lymphocytic leukaemia. Beyond proof of concept, we sought to integrate the characteristics of a clinical application to build a robust predictive classifier from Raman data collected on uncolored blood smears. In a first step, we showed that it was possible to homogenize the data acquired under different experimental conditions. Secondly, we established a training and validation strategy to create stable models that are insensitive to overlearning and adapted to clinical requirements.

---

Raman spectroscopy; chronic lymphocytic leukaemia; data analysis; diagnosis of chronic lymphocytic leukaemia; from bench to bedside

---

**Discipline : SCIENCES DE LA VIE ET DE LA SANTE**

---

**Spécialité : Biophysique**

---



Université de Reims Champagne-Ardenne  
Biospectroscopie Translationnelle : BioSpecT  
51 rue Cognacq-Jay, 51096 Reims Cedex, France

*A mes parents...*

*A mes amis de toujours...*

# Remerciements

Cette merveilleuse aventure dans le monde de la recherche a commencé il y a maintenant six ans lorsque j'ai réalisé mon stage de Master 2 dans le laboratoire BioSpecT, au sein de l'équipe dirigé par le professeur Olivier Piot. Avant d'entamer ce manuscrit, je remercie profondément tous ceux qui m'ont aidé à accomplir ce projet de recherche.

Mes premières pensées vont à mon directeur de thèse, M. Olivier Piot, Professeur des Universités et directeur du laboratoire BioSpecT pour m'avoir accueilli au sein de son équipe. Je le remercie pour le temps conséquent qu'il m'a accordé, ses qualités pédagogiques et scientifiques, sa franchise et sa sympathie. J'ai beaucoup appris à ses côtés et je lui adresse ma gratitude pour tout cela.

J'adresse de chaleureux remerciements à mon co-encadrant de thèse, M. Cyril Gobinet, pour son attention de tout instant sur mes travaux, pour ses conseils avisés et son écoute qui ont été prépondérants pour la bonne réussite de cette thèse. Son énergie et sa confiance ont été des éléments moteurs pour moi. J'ai pris un grand plaisir à travailler avec lui. Cyril, je ne saurais combien te remercier pour tout ce que tu as pu m'apprendre. Je te remercie pour ta gentillesse, ta patience et ta sympathie quotidienne.

Je tiens également à remercier très sincèrement les membres de mon jury de thèse. Je remercie le professeur Ludovic Duponchel 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.

Je suis très reconnaissant envers le professeur Erik Goormaghtich et je le 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.

Docteur Anne Quinquenel, 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.

Monsieur KLOSSA, je vous suis très reconnaissant de faire partie de mon jury de thèse. Au-delà de cela, je tenais à vous remercier pour toute l'énergie que vous avez mise dans le projet

M3S et qui m'a permis de mener à bien ces travaux de thèse qui ont été un vrai bonheur pour moi. Vous êtes une Je vous remercie également de conjuguer professionnalisme, sympathie et bonne humeur à chaque fois que nous nous rencontrons.

Je remercie Abdelilah Beljebbar, pour m'avoir apporté ton savoir en spectroscopie, pour tes suggestions, je te remercie. Et comment ne pas te remercier pour ta sympathie, qui ont su rendre nos échanges très agréables, merci Abdel.

Valérie Untereiner, je te remercie pour toute ton aide depuis que je suis au laboratoire, pour ces connaissances que tu m'as apportées pendant ma thèse, mais également pour toutes ces discussions et nombreux éclats de rires que nous avons partagés dans le bureau, en salle de traitement, la cuisine et les couloirs. Je te remercie pour ta gentillesse et ton aide précieuse.

Je remercie mon pote Liu pour toute l'aide que tu m'as apporté durant cette thèse. Je te remercie pour ces moments de délires et de bonnes humeurs qu'on a partagés au labo et en dehors du labo.

Monsieur le Docteur Bernard Chatelain, je vous adresse mes plus sincères remerciements pour votre expertise en hématologie. J'ai pu apprécier vos qualités tant pédagogiques, scientifiques et humaines. Je vous remercie de votre disponibilité, de votre aide et vos conseils précieux que vous m'avez donné.

Je tiens à remercier également tout particulièrement Martin CHOLLAT, un des éléments important de ce projet. Je te remercie pour toute l'aide et l'énergie que tu as pu mettre afin que ce projet puisse aboutir. Tu as été je pense un élément important dans ce projet et je tiens vraiment à te remercier pour toute ta patience.

Je tiens également à remercier Damien GHELDOF et Morgane EMMERECHEITS pour leur aide dans l'acquisition des données et leurs sympathies.

Vincent Vuiblet, Christine Terry, Ganesh, D. Sockalingum, Hamid Morjani, Jean-François Angiboust, Laurence Van-Gulick, Pierre Jeannesson, merci à tous pour votre aide, pour vos apports scientifiques et pour votre sympathie au cours de ces 4 années de doctorat.

Je remercie également tous mes amis, en particulier, Thomas, Patrice, Flo, Nicolas, Thibault, Momo et Romain. Je remercie aussi mes copains de labo, Christophe, Fatima, Nicolas, Angela, Lila, Adrian, Caroline, Christophe, David, Elodie, Georges, Goutam, Imane, Jaya, Joan, Julien, Laura, Lucie, Mathieu, Nathalie, Qué, The Thuong et Vincent, avec qui j'ai passé des moments

formidables au sein du labo mais aussi et surtout en dehors. Ces dernières années n'auraient pas été les mêmes sans vous.

Je remercie profondément toute ma famille pour leur soutien.

Je remercie du fond du cœur, mes parents et tout particulièrement ma maman qui toute sa vie s'est démenée pour que je puisse faire des études et que je puisse aller le plus loin possible. Sans vous, je n'aurai jamais pu arriver jusqu'ici ! Merci beaucoup.



---

*« La science est un projet coopératif qui se transmet entre les générations. C'est le relais d'une torche du professeur à l'étudiant au professeur. Une communauté d'esprits prenant racine dans l'Antiquité et se dirigeant vers les étoiles. »*

---



# Table des matières

<b>LISTE DES TABLEAUX ET DES ILLUSTRATIONS .....</b>	<b>10</b>
1. Liste des tableaux .....	10
2. Table des illustrations .....	10
<b>LISTE DES ABRÉVIATIONS.....</b>	<b>12</b>
<b>PRÉAMBULE .....</b>	<b>13</b>
<b>CHAPITRE 1 : INTRODUCTION GÉNÉRALE .....</b>	<b>16</b>
1. Fonctionnalité et composition du sang .....	17
1.1. <i>Fonctions principales</i> .....	17
1.2. <i>Composition détaillée</i> .....	17
2. Principales hémopathies .....	23
2.1. <i>Classification des lymphomes</i> .....	24
2.2. <i>Classification des leucémies</i> .....	24
3. La leucémie lymphoïde chronique .....	25
3.1. <i>Le diagnostic</i> .....	25
3.2. <i>Le pronostic</i> .....	30
3.3. <i>Approches thérapeutiques</i> .....	33
4. Les techniques de spectroscopie vibrationnelle .....	37
4.1. <i>Introduction</i> .....	37
4.2. <i>Spectroscopie d'absorption Infrarouge</i> .....	45
4.3. <i>Spectroscopie de diffusion Raman</i> .....	47
4.4. <i>Application des spectroscopies vibrationnelles en clinique</i> .....	52
4.5. <i>La spectroscopie Raman pour l'aide au diagnostic</i> .....	55
5. Analyse des données .....	56
5.1. <i>Prétraitement des données</i> .....	58
5.2. <i>Réduction de données</i> .....	61
5.3. <i>Algorithmes de classification automatiques</i> .....	63
<b>CHAPITRE 2 : MATÉRIELS ET MÉTHODES .....</b>	<b>67</b>
1. Préparation des échantillons sanguins.....	68
1.1. <i>Groupes d'étude</i> .....	68
1.2. <i>Frottis sanguin</i> .....	68
1.3. <i>Cytométrie en flux (MFC)</i> .....	69
1.4. <i>Vérité terrain</i> .....	69
2. Instrumentation et workflow des acquisitions .....	70

2.1.	<i>Dispositif Raman</i> .....	70
2.2.	<i>Acquisition des spectres Raman</i> .....	74
3.	Tests qualités.....	75
3.1.	<i>Rapport Signal/Bruit</i> .....	75
3.2.	<i>Présence de signal de l'hémoglobine</i> .....	76
3.3.	<i>Détection des rayonnements cosmiques</i> .....	77
3.4.	<i>Saturation des spectres</i> .....	78
4.	Prétraitement des spectres Raman .....	78
4.1.	<i>Méthode de prétraitement « stepwise »</i> .....	78
4.2.	<i>Méthode de prétraitement « conjointe »</i> .....	78
5.	Analyses statistiques des spectres.....	80
5.1.	<i>Réduction de la dimension des données</i> .....	80
5.2.	<i>Algorithmes de classifications supervisées</i> .....	80
5.3.	<i>Validation croisée</i> .....	86
5.4.	<i>Test en aveugle des classificateurs optimisés : vote majoritaire</i> .....	91
<b>CHAPITRE 3 : CONSTRUCTION D'UNE BASE DE DONNEES MULTIMODALES .....</b>		<b>92</b>
1.	Les différentes étapes de production de la base de données .....	94
1.1.	<i>Formulaire clinique électronique : e-CRF</i> .....	94
1.2.	<i>Scanneur de lame et identification des cellules d'intérêt</i> .....	95
1.3.	<i>Imagerie à phase quantitative : QPI</i> .....	95
1.4.	<i>Spectroscopie Raman et imagerie monochrome à haute résolution</i> .....	95
1.5.	<i>Nomenclature</i> .....	96
1.6.	<i>Identification de la classe des cellules analysées</i> .....	97
2.	Problèmes rencontrés.....	97
<b>CHAPITRE 4 : CONSTRUCTION DE MODELES RAMAN POUR LE DIAGNOSTIC AU NIVEAU PATIENT .....</b>		<b>98</b>
Article #1 : Privilégier l'étape de prétraitement pour assurer la transférabilité clinique de données Raman multicentriques acquises sur les lymphocytes dans différentes conditions expérimentales et instrumentales.....		99
Article #2 : Mise en œuvre d'une stratégie de classification de données Raman collectées dans différentes conditions cliniques : application au diagnostic de la leucémie lymphocytaire chronique.....		125
<b>CHAPITRE 5 : APPLICATION À L'ECHELLE CELLULAIRE .....</b>		<b>147</b>
1.	Stratégie de labellisation par immunohistochimie .....	149
1.1.	<i>Mesures des spectres Raman</i> .....	149
1.2.	<i>Prétraitement des mesures Raman</i> .....	150
1.3.	<i>Similarité des signatures Raman entre les sous-groupes lymphocytaires</i> .....	150

1.4. Développement de modèles de reconnaissance lymphocytaire et paramètres d'optimisation .....	152
2. Résultats complémentaires : études pilotes .....	162
2.1. Tri cellulaire magnétique .....	162
2.2. Numération spectrale semi-supervisée par algorithme mémétique .....	164
<b>CONCLUSIONS ET PERSPECTIVES .....</b>	<b>167</b>
1. Conclusions .....	168
2. Perspectives .....	169
2.1. Construction d'une base de données multimodale .....	169
2.2. Enrichissement de la base de données spectrale et utilisation d'autres stratégies de classification .....	170
<b>BIBLIOGRAPHIE .....</b>	<b>172</b>
<b>ANNEXES .....</b>	<b>181</b>

# LISTE DES TABLEAUX ET DES ILLUSTRATIONS

## 1. Liste des tableaux

Tableau 1 - Les principaux constituants du sang, leurs fonctions et leurs temps de vie [27] .....	19
Tableau 2 - Résumé de différents facteurs pronostiques en fonction de leur issue bonne ou mauvaise.....	32
Tableau 3 - Résumé des principales thérapies existantes pour la LLC, les molécules courantes utilisées et leur action.....	34
Tableau 4 - Nombre de patient et de cellules en entraînement, validation et test par tour .....	89
Tableau 5 - Tentative d'attribution des bandes spectrales permettant de séparer les lymphocytes NK+T et B .....	155
Tableau 6 - Nombre total de cellules dans le jeu d'entraînement et le jeu de validation .....	157
Tableau 7 - Nombre de cellules NK+T, B et B sains et B LLC en entraînement et validation externe .....	159
Tableau 8 - Performances des différentes pour chaque paire de lymphocytes .....	164
Tableau 9 - Résultats issus de l'algorithme génétique .....	166

## 2. Table des illustrations

Figure 1 - Schéma représentatif des étapes de séparation des constituants du sang – (1) Prise de sang (2) centrifugation (3) séparation des constituants sanguins .....	17
Figure 2 - Représentation schématique des différents compartiments du système hématopoïétique .....	18
Figure 3 - Image morphologique d'un frottis sanguin .....	21
Figure 4 - Morphologies de différents lymphocytes, de gauche à droite : lymphome folliculaire sain, réactif, leucémie à cellules velues, ombre de Gumprecht, lymphocytes à chromatine mottée (LLC) .....	28
Figure 5 - Isolement des cellules B ou T/NK saines/LLC par MFC .....	29
Figure 6 - Le stade de Binet permet au clinicien selon différents critères d'évaluer le stade de la leucémie lymphoïde chronique [36].....	31
Figure 7 - Schéma présentant l'effet des analogues aux purines sur un lymphocyte B LLC - Les purines ont pour effet de bloquer la réplication de l'ADN du lymphocyte, ceci va engendrer l'apoptose du lymphocyte.....	35
Figure 8 - Schéma représentatif des étapes d'action du Rituximab sur un lymphocyte B - Le rituximab s'accroche spécifiquement à l'antigène CD20 des lymphocytes pré-B et B matures. Cet antigène s'exprime dans plus de 95 % des cellules B [55] .....	36
Figure 9 - Représentation schématique des différents phénomènes lors d'une interaction d'une lumière/matière .....	37
Figure 10 - Ensemble des domaines électromagnétiques .....	38
Figure 11 - Représentation schématique de vibration de valence .....	39
Figure 12 - Oscillateur mécanique .....	40
Figure 13 - Niveaux vibrationnels d'un système harmonique et d'un système réel .....	43
Figure 14 - Spectre infrarouge et Raman d'ester méthylique d'acide oléique. [60].....	44
Figure 15 - Exemples de modes fondamentaux de vibration moléculaire et de leurs pics Raman associés .....	45
Figure 16 - Appareillage Spotlight 2 FTIR, le système est constitué d'un microscope Spotlight 300 de Perkin Elmer associé à un spectromètre infrarouge à transformée de Fourier (Spectrum One, Perkin Elmer) ..	47
Figure 17 - Schéma montrant les principaux processus de diffusion Rayleigh, Raman Stokes et Raman anti-Stokes .....	51
Figure 18 - Représentation d'un spectre de diffusion Raman à température ambiante .....	52
Figure 19 - Diagramme des principales étapes de traitement numérique pour des spectres Raman. ....	58
Figure 20 - Exemple de dendrogramme de clustering hiérarchique .....	64
Figure 21 - Étaleur de frottis automatisés Hemaprep .....	68
Figure 22 - Plateforme d'acquisition de spectres Raman M3S - Le système est composé d'un microscope conventionnel, d'un module Raman, d'un laser 532nm, d'un éclairage par LED et d'une platine motorisée XYZ.....	70

Figure 23 - Schéma optique d'un spectromètre Raman .....	71
Figure 24 - Principe de fonctionnement d'un réseau holographique .....	73
Figure 25 - Tâche de diffraction de la lumière [132] .....	74
Figure 26 - Interface d'acquisition du système M3S .....	75
Figure 27 - Spectre Raman ayant un mauvais rapport signal sur bruit .....	76
Figure 28 - Spectre Raman de la superposition d'un lymphocyte et d'un globule rouge .....	77
Figure 29 - Spectre de lymphocyte contenant un Spike (flèche noire) .....	77
Figure 30 - Effet du prétraitement par EMSC. (a) Spectres Raman de lymphocytes avant l'EMSC, (b) spectres prétraités par EMSC .....	80
Figure 31 - Principe de l'algorithme PLS-R. X et Y sont les données de départ et les réponses (classes) du modèle, respectivement. T et U sont les scores des variables latentes issues de la modélisation. P et Q correspondent aux loadings [139].....	82
Figure 32 - Schéma présentant le calcul d'un hyperplan pour des données linéaire et séparable par l'algorithme SVM linéaire .....	83
Figure 33 - Représentation en deux dimensions de données imbriquées de façon concentrique et non linéairement séparable par un SVM linéaire. ....	84
Figure 34 - a) Exemple d'une projection de données non linéairement séparables vers un espace des caractéristiques de dimension supérieure. b) Représentation en deux dimensions du résultat d'une SVM à noyau Rbf. Ici il est possible de voir l'hyperplan optimal choisi par l'algorithme.....	85
Figure 35 - Schéma de fonctionnement d'une forêt aléatoire décisionnelle .....	86
Figure 36 - Diagramme du principe de fonctionnement de la double validation croisée répétée (rdCV) .....	90
Figure 37 - Principe du vote majoritaire pour le diagnostic d'un nouveau patient .....	91
Figure 38 - Représentation des différentes étapes du flux d'acquisition d'un frottis sanguin sur le système M3S .....	94
Figure 39 - Exemple de nomenclature type utilisée pour identifier chaque fichier de données .....	96
Figure 40 - Moyenne et écart-type des différents sous-type lymphocytaire .....	151
Figure 41 - Projection 3D de l'ensemble des scores provenant de l'ACP pour les trois sous-groupes de lymphocytes .....	152
Figure 42 - Spectre moyen pour les lymphocytes NK+T (bleu), B sain (vert) et B LLC. Région d'intérêt, composée de 50 nombres d'onde les plus caractéristiques, permettant de différencier les sous-groupes lymphocytaires. ....	154
Figure 43 - Spectre moyen pour les lymphocytes NK+T (rouge) et B (vert). Région d'intérêt, composée de 50 nombres d'onde les plus caractéristiques, permettant de différencier les sous-groupes lymphocytaires .....	154
Figure 44 - Spectre moyen pour les lymphocytes B sains (rouge) et LLC (vert). Région d'intérêt, composée de 50 longueurs d'onde les plus caractéristiques, permettant de différencier les sous-groupes lymphocytaires. ....	155
Figure 45 - Comparaison des différents algorithmes de classification pour le modèle tri-classe .....	158
Figure 46 - Comparaison des différents algorithmes de classification pour le modèle bi-classe NK+T vs B ...	160
Figure 47 - Comparaison des différents algorithmes de classification pour le modèle bi-classe B sains vs B LLC .....	160
Figure 48 - Protocol du tri magnétique des différents sous-types cellulaires .....	163

## LISTE DES ABRÉVIATIONS

<b>LLC</b>	<b>Leucémie lymphoïde chronique</b>
<b>rdCV</b>	<b>Double validation croisée répétée</b>
<b>CV</b>	<b>Validation croisée</b>
<b>GRB</b>	<b>Globules rouges</b>
<b>LNH</b>	<b>Lymphomes non hodgkiniens</b>
<b>MDH</b>	<b>Lymphomes d'Hodgkin</b>
<b>LA</b>	<b>Leucémies aiguës</b>
<b>LAL</b>	<b>Leucémies aiguës lymphoblastiques</b>
<b>LAM</b>	<b>Leucémies aiguës myéloblastiques</b>
<b>SR</b>	<b>Syndrome de Richter</b>
<b>NFS</b>	<b>Hémogramme ou numération sanguine</b>
<b>MFC</b>	<b>Cytométrie en flux</b>
<b>TDL</b>	<b>Temps de doublements des lymphocytes</b>
<b>ACP</b>	<b>Analyse en composantes principales</b>
<b>FM</b>	<b>Méthodes par factorisation matricielle</b>
<b>ACI</b>	<b>Analyse en composantes indépendantes</b>
<b>NMF</b>	<b>Factorisation en matrice non-négative</b>
<b>KM</b>	<b>K-Means</b>
<b>KMD</b>	<b>K-Médoïdes</b>
<b>SVM</b>	<b>Support Vector Machine</b>
<b>RF</b>	<b>Random Forest</b>
<b>KW</b>	<b>Kruskall Wallis</b>
<b>ANN</b>	<b>Réseau de neurones</b>
<b>KNN</b>	<b>K plus proche voisin</b>
<b>PLS-DA</b>	<b>Partial Least Squares - Discriminant Analysis</b>

# **PRÉAMBULE**

Les techniques biophotoniques sont des dispositifs d'investigation permettant d'avoir des informations d'ordre moléculaire. Elles reposent sur l'interaction non destructive de la lumière avec la matière. Ces techniques sont très utilisées dans le domaine de la recherche pour développer de nouvelles modalités qui pourraient être implantées en clinique, afin d'aider les cliniciens à améliorer la prise en charge des patients. Elles permettraient d'améliorer le diagnostic, le pronostic, ou encore la surveillance des patients tout en réduisant potentiellement les coûts hospitaliers.

Les spectroscopies vibrationnelles permettent d'avoir un accès direct à l'empreinte moléculaire ou biochimique de l'échantillon. Ainsi la moindre variation au sein de l'entité biologique liée à une physiopathologie ou cytopathologie, se répercute directement sur la signature moléculaire de l'échantillon. Ainsi ces techniques semblent capables d'identifier les anomalies liées aux pathologies sans utiliser de marquage (label-free). C'est ce qui a pu être montré ces dernières années dans différents domaines tels que la caractérisation tissulaire (peau [1-3] et rein [4, 5]). Mais également dans l'approche diagnostique pour différents types de cancer : col de l'utérus [6-9], prostate [10], poumon [11], colon [12], œsophage [12-14] et cerveau [15]. Ces techniques ont pu également montrer tout leur potentiel dans la quantification moléculaire sans réactif pour différent type d'échantillon biologique (sérum humain [16], urine [8]).

Au cours des dernières années, de nombreuses applications, ont pu être rapportées dans le domaine biomédical [17, 18], en particulier au niveau cellulaire. Différents auteurs ont pu par exemple, analyser du sperme humain, montrant qu'il était possible de discriminer des échantillons normaux et altérés au niveau de l'ADN [19, 20]. D'autres études se sont intéressées à l'évaluation au niveau cellulaire des états de différenciation [21-24], mais également à la détection des changements épigénétiques sur des lignées Jurkat [25]. Il a été montré que les spectroscopies vibrationnelles peuvent également servir de descripteurs pour des phénomènes dynamiques par exemple suivre l'évolution de la molécule de cytochrome C dans le cadre de l'apoptose cellulaire [26].

L'objectif de ce travail de thèse est de contribuer au développement d'un système multimodal basé sur la technologie vibrationnelle et implantable en seconde ligne sur la chaîne d'analyse en clinique hématologique. Ce développement est mené dans le contexte du diagnostic de la leucémie lymphoïde chronique (LLC) et de sa caractérisation biochimique. Le travail a consisté dans un premier temps à la construction d'une base de données spectrale. Pour

se faire des acquisitions Raman ont été réalisées sur des frottis sanguin de patients sains et atteints de LLC. Afin d'être dans des conditions cliniques optimales, les frottis ont été préparés en milieu clinique et sur des lames de verre. Une fois les acquisitions réalisées, les données ont subi une préparation et des analyses chimiométriques. La finalité de ce projet a été d'établir un algorithme basé sur des modèles de prédiction améliorés dont la construction a été guidée par une vérité terrain.

Ce mémoire de thèse comprend plusieurs parties. Dans **une première partie**, nous décrirons la composition du sang et de ces éléments figurés (cellules sanguines). Nous verrons également les maladies hématologiques et plus particulièrement les cancers du sang tel que la leucémie lymphoïde chronique. Nous développerons ensuite les principes des spectroscopies vibrationnelles infrarouges et Raman et nous verrons comment ces données vibrationnelles peuvent être traitées et exploitées au moyen de méthodes d'analyses statistiques multivariées.

**La seconde partie** développera le matériel qui a été utilisé pour la préparation des échantillons. Nous verrons également les différentes caractéristiques des microscopes et spectromètres utilisés ainsi que les différentes méthodes chimiométriques pour l'exploitation des données vibrationnelles.

**La troisième partie** présente l'ensemble des expérimentations réalisées et des résultats obtenus. Tout d'abord, il sera question de voir l'effet de l'application d'un prétraitement sur des données acquises dans des conditions expérimentales et instrumentales différentes. Dans un second temps, nous verrons les résultats obtenus suite à la construction d'un classifieur diagnostique pour la leucémie lymphoïde chronique au niveau patient, en utilisant une méthode hybride de double cross-validation (rdCV) combinée à une prise de décision ensembliste. Enfin, nous appliquerons cette méthode hybride dans le cas d'une étude préliminaire, afin d'identifier les sous-types cellulaires, lymphocytes B sains, B LLC, T + NK.

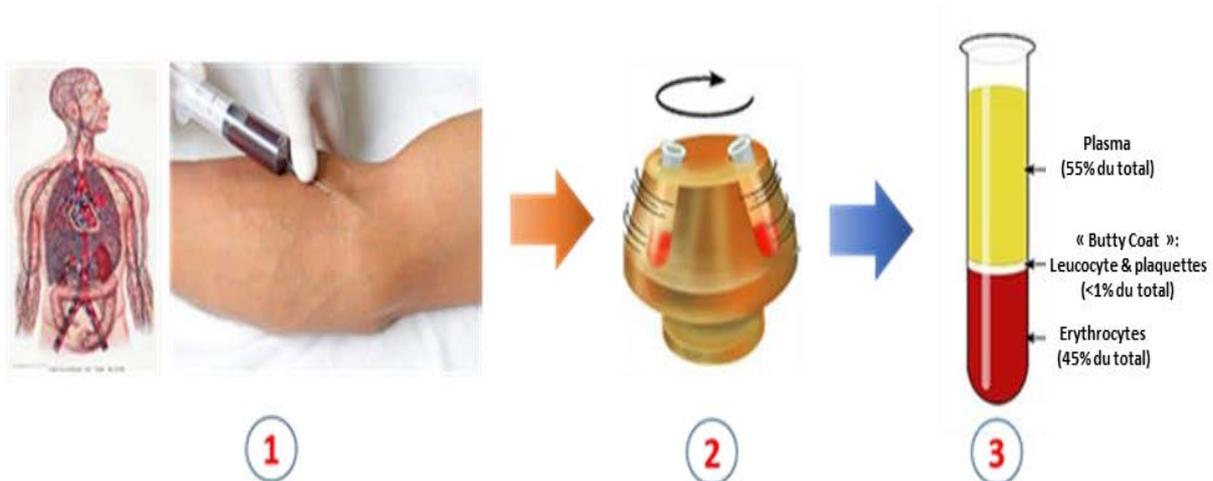
**La dernière partie** expose les principales conclusions de ces travaux ainsi que leurs perspectives.

# **CHAPITRE 1 : INTRODUCTION GÉNÉRALE**

## 1. Fonctionnalité et composition du sang

### 1.1. Fonctions principales

Le sang est un élément extrêmement complexe de par sa composition. Comme il est possible de voir sur la Figure 1, après centrifugation du sang, celui-ci se sépare en différentes phases. La première est la phase porteuse appelée plasma (55% du volume sanguin), la seconde est une phase contenant les érythrocytes ou globules rouges (45% du volume sanguin) et enfin la dernière est une phase contenant les leucocytes et les plaquettes (< 1% du total sanguin). Ces cellules sont en suspension dans le plasma et circulent de façon unilatérale au travers du corps grâce aux contractions cardiaques. Le sang à lui seul représente 7% du poids corporel soit un volume de 5 litres chez un adulte moyen. Sa fonction principale est le transport et la diffusion dans tout le corps des éléments indispensables à la survie et au bon fonctionnement de l'entité biologique : oxygène, éléments nutritifs, hormones et système immunitaire. Le sang permet également d'évacuer les déchets par exemple le dioxyde de carbone ou les composés azotés. Grâce à tous ces constituants sanguins, il est possible de réaliser différents tests permettant de diagnostiquer des pathologies.



*Figure 1 - Schéma représentatif des étapes de séparation des constituants du sang – (1) Prise de sang (2) centrifugation (3) séparation des constituants sanguins*

### 1.2. Composition détaillée

L'hématopoïèse (Figure 2) qui signifie « production du sang » est le processus physiologique de production et de renouvellement des éléments figurés constitutifs du sang : globules rouges, globules blancs et plaquettes à partir des cellules souches hématopoïétiques.

Les cellules du sang ont des durées de vie très variables selon leur type allant de quelques heures à plusieurs années (tableau 1). Le processus d'hématopoïèse comprend 3 compartiments, les progéniteurs hématopoïétiques, les précurseurs et les cellules différenciées.

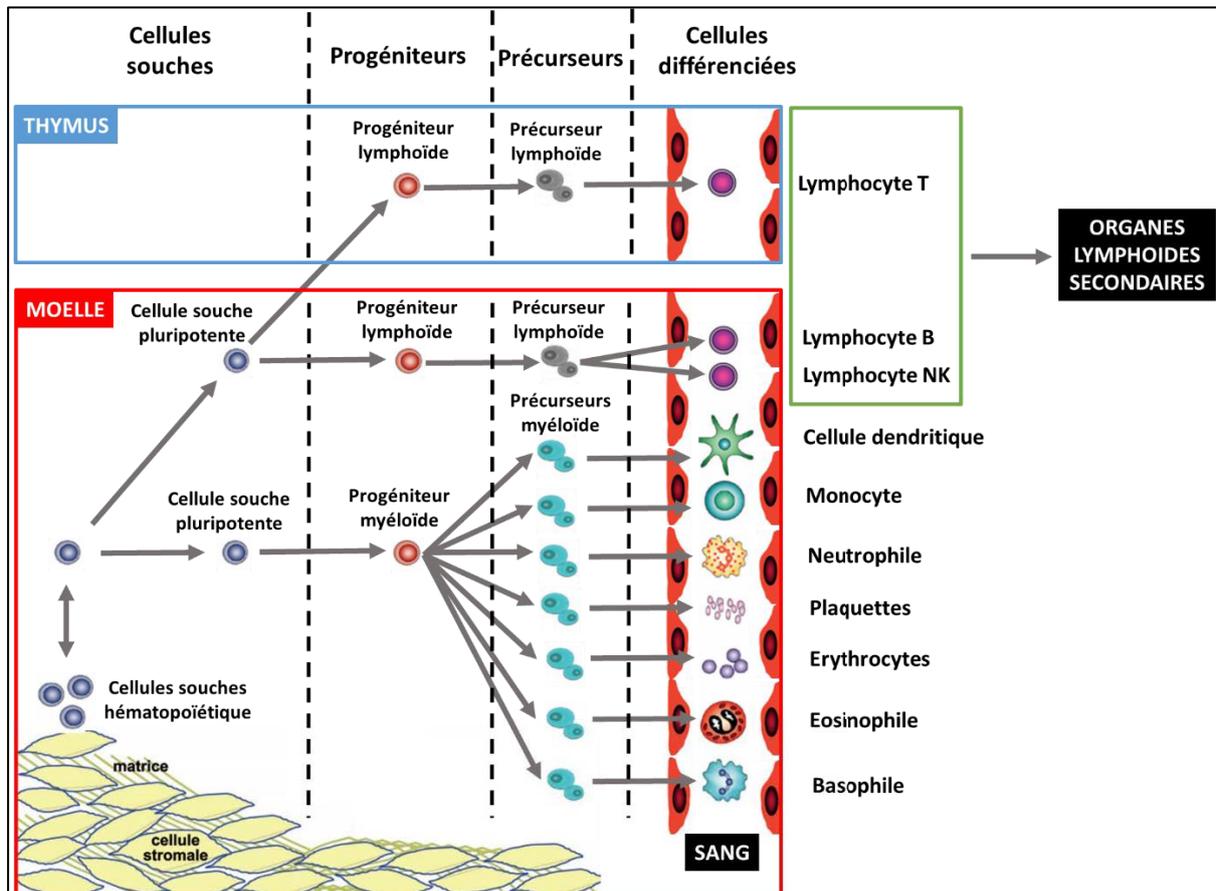


Figure 2 - Représentation schématique des différents compartiments du système hématopoïétique

Les cellules souches sont des cellules immatures totipotentes qui ont le potentiel de devenir à peu près n'importe quel type de cellules du corps, elles sont capables de s'autorenouveler et sont en très faible quantité dans l'organisme. Dans le processus d'hématopoïèse, grâce à une série de transformations, ces cellules souches deviennent dans un premier temps des cellules progénitrices. Après différenciations ces cellules progénitrices donnent naissance à des cellules précurseurs qui deviendront après multiplications et maturations des cellules sanguines matures. Ces cellules matures vont alors migrer vers le sang et les organes lymphoïdes secondaires (ganglions lymphatiques, rate et foie...).

Dans la moelle osseuse, il existe deux grands axes de différenciations de ces cellules souches: les cellules myéloïdes et lymphoïdes. Les cellules lymphoïdes donneront naissance

aux différents lymphocytes B, T et NK, tandis que les cellules myéloïdes sont impliquées dans la production de quatre lignées de cellules sanguines: les globules rouges, les granulocytes, les monocytes et les plaquettes. Au cours de la vie humaine, il est important que le corps produise des cellules sanguines de façon continue et régulée. Un dysfonctionnement dans le système hématopoïétique peut provoquer des maladies graves comme la leucémie ou le lymphome.

*Tableau 1 - Les principaux constituants du sang, leurs fonctions et leurs temps de vie [27]*

Type cellulaire	Fonctions dans le corps	Durée de vie
<b>Plaquettes</b>	Hémostase	10 jours
<b>Polynucléaires et monocytes</b>	<p><b>Les neutrophiles</b> ont un rôle de phagocytose, bactéricide</p> <p><b>Les basophiles</b> jouent un rôle dans le déclenchement des réactions inflammatoires</p> <p><b>Les éosinophiles</b> ont pour fonction la lutte contre les parasites</p> <p><b>Les monocytes jouent un rôle dans l'activation du système lymphocytaire.</b> Ils se différencient en passant dans les tissus en macrophage</p>	<p>Polynucléaires : 24-30 heures</p> <p>Monocytes : Plusieurs jours</p>
<b>Globules rouges</b>	Transport de l'oxygène et du gaz carbonique	120 jours
<b>Lymphocytes</b>	Système immunitaire	Plusieurs jours à plusieurs années (pour les lymphocytes mémoires)

### **1.2.1. Le plasma**

Le plasma est un liquide physiologique composé à 90% d'eau, qui a pour rôle de permettre aux globules rouges, aux globules blancs, plaquettes et autres éléments figurés du sang de pouvoir se mouvoir dans le corps. Le plasma représente environ 55% du volume sanguin. Le plasma est également riche en constituants fonctionnels, en protéines de transport en éléments pour la défense (immunoglobulines ou anticorps), en éléments nutritifs (glucose, acides aminés, acides gras...), en ions, en sels, en hormones et en enzymes.

### **1.2.2. Les éléments figurés**

#### **a) Les plaquettes**

Les plaquettes ou thrombocytes, sont des unités granuleuses sans noyau, circulant dans le sang, formés dans la moelle osseuse par fragmentation de cellules géantes que l'on appelle des mégacaryocytes en phase mature. À l'état normal, elles ont une forme discoïde de diamètre entre 2 et 5  $\mu\text{m}$  pour un volume de 6 à 10  $\mu\text{m}^3$ . Leur fonction biologique principale est d'intervenir dans la première phase de l'hémostase sanguine (coagulation) pour arrêter les saignements : l'hémostase primaire. Ce sont de petits sacs contenant différentes substances actives permettant la formation de caillot sanguin. En cas de lésion endothéliale, les plaquettes s'activent et adhèrent principalement au collagène constituant le sous-endothélium, formant ainsi un agrégat de plaquettes reliées par du fibrinogène, appelé thrombus. L'adhésion plaquettaire s'effectue grâce à différents récepteurs exprimés à la surface des thrombocytes. Chez un sujet en bonne santé, leur nombre varie de 150 000 à 400 000/mm<sup>3</sup> de sang et leur durée de vie est d'environ 7 jours.

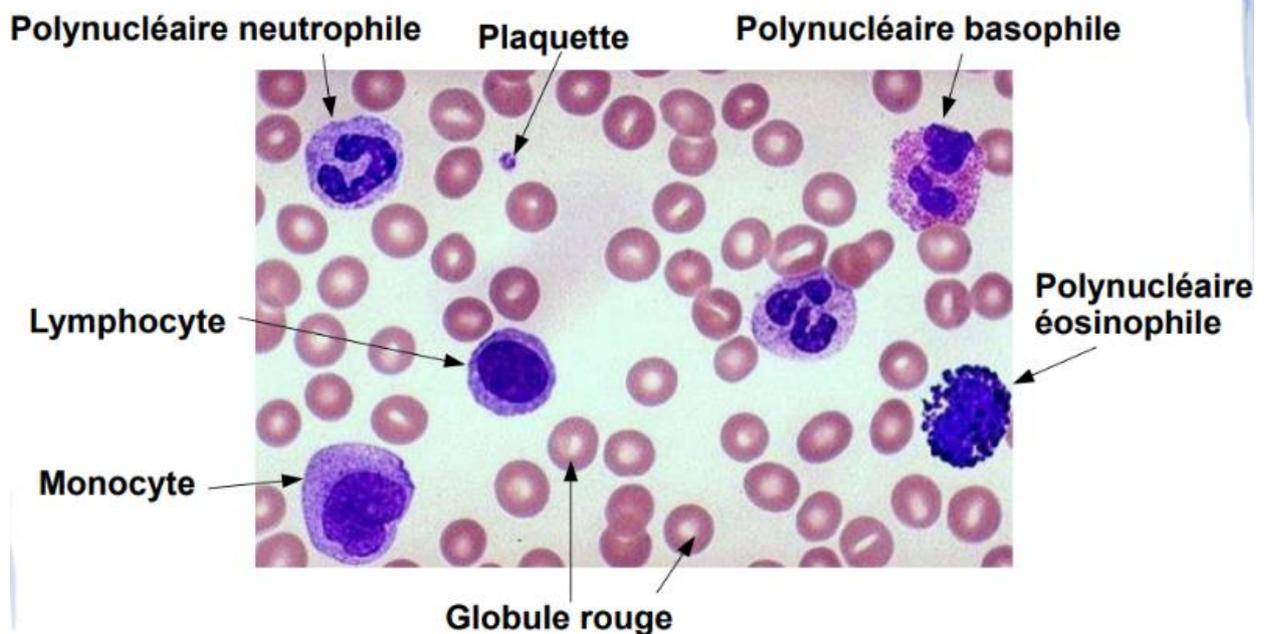
#### **b) Les hématies**

Les globules rouges (GBR), érythrocytes ou hématies sont des cellules sanguines qui sont issues de la lignée des érythroblastes de la moelle osseuse aboutissant aux réticulocytes (cellules précédant le stade des GBR) après éjection de leur noyau. Les GBR sont donc des petits disques biconcaves sans noyau ni organites, d'environ 7 $\mu\text{m}$  de diamètre qui permettent les échanges gazeux dans tout l'organisme. Ils possèdent des propriétés élastiques leur permettant de se déformer pour passer les capillaires sanguins plus petits qu'eux (plasticité globulaire). Schématiquement, on peut représenter l'hématie comme un réservoir contenant de l'hémoglobine (couleur rouge). L'hémoglobine est un hétérotétramère protéique constitué de 4

chaînes polypeptidiques ( $2\alpha$  et  $2\beta$ ) contenant du fer et qui a pour fonction de capter l'oxygène  $O_2$  et le dioxyde de carbone  $CO_2$ . Sa durée de vie est d'environ 120 jours. C'est également grâce aux GBR que l'on peut déterminer les différents groupes sanguins.

### c) **Maintien de l'intégrité de l'organisme : cellules du système immunitaires**

Le système immunitaire se compose de différentes familles de globules blancs ou leucocytes (Figure 3) qui ont pour fonction la défense de l'organisme contre tous les corps étrangers ou endommagés (virus, bactérie, toxines, débris cellulaires...). Ainsi ce système est capable de faire la distinction entre les cellules du soi et du non-soi, c'est-à-dire de différencier les cellules qui lui appartiennent ou non. Les cellules du non-soi ont une double origine, les cellules exogènes et endogènes. En effet, il s'agit soit d'un élément étranger potentiellement pathogène et étant extérieur à l'organisme, soit d'un élément de l'organisme lui-même ayant subi une modification structurelle ou moléculaire (cellules cancéreuses). Les globules blancs sont reconnaissables sur un frottis sanguin coloré au May Grünwald Giemsa (MGG). Ainsi lors de la lecture au microscope, on peut distinguer deux types de globules blancs selon la morphologie de leur noyau, qui peut être soit monolobé (lymphocytes, monocytes et macrophages) ou plurilobé (granulocytes).



*Figure 3 - Image morphologique d'un frottis sanguin*

Les granulocytes ou polynucléaires sont des cellules de type myéloïde, qui sont impliquées dans les réactions immunitaires naturelles ou innées, rapides et non spécifiques. Ces

granulocytes sont dirigés contre les bactéries, les champignons, les parasites et les particules étrangères. Les polynucléaires ou granulocytes peuvent être de plusieurs sous-catégories, neutrophiles, éosinophiles et basophiles.

Les granulocytes neutrophiles sont les plus nombreux des polynucléaires, représentant 50 à 70% de la formule leucocytaire global. Ils possèdent une durée de vie de 24h et une taille d'environ 12  $\mu\text{m}$  de diamètre. Ils se caractérisent par la présence d'un noyau multilobé segmenté de 2 à 5 lobes, réunis par un filament fin de chromatine. Ces polynucléaires neutrophiles sont des cellules phagocytaires car ils sont capables d'ingérer et de digérer par phagocytose des agents pathogènes et des débris cellulaires. Ces cellules sont également capables de libérer des agents antibactériens et des enzymes protéolytiques (qui digèrent l'agent pathogène).

Les granulocytes éosinophiles sont des cellules antiparasitaires, possédant également un pouvoir phagocytaire. Ces cellules représentent environ 2 à 4% des leucocytes. Ces granulocytes se présentent comme des cellules bilobées avec un cytoplasme contenant de grosses granulations. Ces cellules s'attaquent aux parasites en se fixant dessus et libérant des enzymes destinées à les détruire.

Les granulocytes basophiles ont une fonction de messenger et permettent d'attirer d'autres globules blancs en libérant des molécules chimiques telles que l'histamine qui favorisent l'inflammation. Ces cellules ne représentent que <1% des leucocytes. Après coloration le noyau est généralement invisible à cause des denses granulations présentes dans le cytoplasme.

Parmi la famille des globules blancs monolobés, les monocytes contribuent à l'élimination des bactéries, des particules étrangères et des globules rouges en fin de vie par phagocytose. Ils sont généralement de grande taille (15 à 22  $\mu\text{m}$ ) et possèdent un cytoplasme abondant et un noyau en forme de haricot. Ils peuvent rester entre 1 à 3 jours dans le sang, pour ensuite migrer vers différents tissus de l'organisme pour se différencier en macrophages. Ces cellules représentent environ 2 à 8% des leucocytes.

Les lymphocytes sont des cellules de type lymphoïde, qui sont impliquées dans les réactions immunitaires acquises, lentes et spécifiques, ce sont des cellules immunocompétentes. Chez un adulte sain, les lymphocytes sont au nombre de  $5 \times 10^{11} - 2 \times 10^{12}$  qui se répartissent dans

les différents organes du corps. Ces cellules sont composées de trois classes, les lymphocytes B, les lymphocytes T et les lymphocytes NK.

Les cellules tueuses (lymphocytes NK) ont un rôle important dans la défense immunitaire innée. En effet, elles permettent de réguler et d'empêcher la dissémination des cellules cancéreuses ou infectées par un virus en les détruisant non pas en les phagocytant mais en libérant des substances chimiques qui engendrent la destruction de la cellule infectée ou défectueuse. Pour cela elles détectent les arrangements anormaux des protéines membranaires.

Les lymphocytes B et T ont un rôle important dans la défense immunitaire spécifique (adaptative). Cette défense adaptative ou acquise s'active après la défense innée et à un temps de réponse beaucoup plus long que cette dernière. Les lymphocytes B deviennent matures et acquièrent leurs immunocompétences au niveau de la moelle osseuse. Ils ont pour rôle de reconnaître un antigène moléculaire grâce à leurs récepteurs de surface. Après avoir reconnu le corps étranger, le lymphocyte B se différencie en plasmocyte qui va se mettre à sécréter des anticorps contre les antigènes étrangers. Les lymphocytes T se développent au niveau du thymus. On peut distinguer deux sous-types de lymphocytes T : les lymphocytes T CD4<sup>+</sup> et les lymphocytes T CD8<sup>+</sup>. Les CD4<sup>+</sup> ou lymphocytes T « helper » ont pour fonction d'activer les autres lymphocytes en sécrétant des cytokines et ont aussi une activité cytotoxique. Les CD8<sup>+</sup> ont une activité cytolytique et éliminent les cellules étrangères ou les cellules infectées par un virus ou un parasite intracellulaire. Les lymphocytes stimulés possèdent une "mémoire" qui leur permet de répondre rapidement à une rencontre subséquente du même antigène. Le système immunitaire est enfin doté d'une capacité de reconnaissance d'un vaste panel d'antigènes, ce qui lui attribue sa qualité d'adaptation.

## **2. Principales hémopathies**

Une hémopathie est une maladie du sang qui peut toucher les globules rouges, les globules blancs et les plaquettes. Elle peut donc dans le cas d'un cancer du sang, avoir pour origine une prolifération anarchique des cellules immunitaires matures ou immatures à l'origine d'hémopathies chroniques à évolution lente ou rapide, pouvant toucher également les organes d'origine lymphoïde tel que la moelle osseuse, le thymus, la rate ou les ganglions lymphatiques. La dénomination des hémopathies prend en compte l'aspect cytologique, l'évolution de celle-ci, mais aussi le site du développement initial. Ainsi il existe plusieurs types de néoplasies en particulier les lymphomes et les leucémies. Les leucémies se développent à partir de cellules

uniques (clones) dans le sang ou la lymphe, alors que les lymphomes vont se développer et donner une masse tumorale solide dans un tissu lymphoïde.

## **2.1. Classification des lymphomes**

Les néoplasies lymphoïdes ou lymphomes sont des cancers du tissu lymphoïde ganglionnaire qui sont caractérisés par une prolifération monoclonale maligne du système lymphatique, affectant couramment la rate, la moelle osseuse et les ganglions périphériques. Elles se développent pour former une tumeur à partir des cellules lymphoïdes à différentes phases de leur maturation donnant les lignées B ou T. Il existe deux grands groupes bien distincts : les lymphomes non hodgkiniens (LNH) et les lymphomes d'Hodgkin (MDH) nommé d'après Thomas Hodgkin [28].

## **2.2. Classification des leucémies**

On peut distinguer deux grandes familles de leucémies qui se distinguent schématiquement en fonction de deux critères fondamentaux que sont leur évolution et la nature des cellules anormales. Dans les deux cas, la maladie se caractérise par une hyperlymphocytose monoclonale anormale. Selon la vitesse d'évolution et l'origine des globules blancs anormaux, les leucémies prennent le nom de « chronique » ou « aigüe » caractérisant un développement lent ou rapide de la maladie, et de « lymphoïdes » ou « myéloïdes » quand les cellules trouvent leur origine au niveau des cellules souches lymphoïdes ou des cellules souches myéloïdes. Ainsi pour toutes ces leucémies, les symptômes, leur évolution et leurs traitements sont donc différents.

### **2.2.1. Les leucémies aigües**

Les leucémies aiguës (LA) sont des hémopathies à évolution rapide, qui sont caractérisées par une prolifération monoclonale de cellules cancéreuses, issues de précurseurs hématopoïétiques anormales. Ces cellules se retrouvent bloquées dans un processus de maturation à l'état de « blaste » soit dans un stade immature. Ces précurseurs anormaux s'accumulent de façon anarchique dans la moelle, dans le sang et dans d'autres organes sans finir leur maturation et ils ne remplissent pas leurs fonctions habituelles. Cette prolifération anarchique empêche la maturation des autres cellules sanguines normales entraînant ainsi anémie, neutropénie et thrombopénie.

On peut distinguer deux types de LA, selon leur origine cellulaire. Les leucémies aiguës lymphoblastiques (LAL) issues de précurseurs lymphoïdes (donnant les lymphocytes T, B et

NK) et les leucémies aiguës myéloblastiques (LAM) issues des précurseurs myéloïdes (donnant les autres globules blancs : polynucléaires, monocytes...). Sans traitement adéquat, ce type de leucémie peut engager le pronostic vital du patient.

### **2.2.2. Les leucémies chroniques**

Les leucémies chroniques (LC) sont des hémopathies à évolution lente donc généralement diagnostiquées tardivement. Elles sont caractérisées par une prolifération monoclonale de cellules qui ont terminé leur maturation ou d'aspect mature. Tout comme la leucémie aiguë, on retrouve deux types de leucémie chronique: lymphoïde et myéloïde caractérisé selon l'origine des cellules. La leucémie lymphoïde chronique est une des hémopathies les plus fréquentes [29, 30], elle touche essentiellement les adultes et plus particulièrement les hommes (56,5%) que les femmes. L'âge médian pour le diagnostic est de 71 ans pour les hommes et 74 ans pour les femmes. Dans le rapport de l'institut de veille sanitaire de 2013, on estime à 4500 les nouveaux cas en France en 2012. Il existe deux types de LLC touchant soit dans 95% des cas les lymphocytes B, soit dans 5% des cas les lymphocytes T. Généralement, cette hémopathie est caractérisée par la prolifération dans les tissus lymphoïdes, la moelle osseuse et le sang, de lymphocytes B monoclonaux matures. La monoclonalité de ces lymphocytes B LLC est due principalement à l'expression d'un seul type d'immunoglobulines (Ig) de surface, soit kappa, soit lambda [31]. En effet, un patient sain possède les deux types d'Ig. C'est cette maladie que nous avons prise comme sujet d'étude dans ce mémoire.

## **3. La leucémie lymphoïde chronique**

### **3.1. Le diagnostic**

Le diagnostic actuel de la LLC se déroule en quatre étapes, une étape d'examen clinique à l'hôpital ou en cabinet médical et trois étapes d'examens sanguins (hémogramme, morphologie, immunophénotypage) au laboratoire.

#### **3.1.1. Signes de la maladie**

La LLC est une maladie hétérogène en termes de présentation clinique au moment du diagnostic, avec différentes évolutions possibles en termes de gravité de la maladie. Tous les patients atteints de LLC ne sont pas égaux en ce qui concerne leur survie globale mais également en ce qui concerne la durée ou l'efficacité du traitement.

La leucémie lymphoïde chronique étant une maladie à évolution progressive, les patients ne présentent généralement aucun symptôme. Ainsi au début de la maladie, lors d'une simple prise de sang, on ne peut mettre en évidence qu'une augmentation limitée des lymphocytes. Dans la majorité des cas, la leucémie lymphoïde chronique est détectée par hasard. Dans le cas d'une personne non atteinte de la LLC, les anciens lymphocytes du sang entrent en apoptose et sont remplacés par des cellules plus jeunes selon le processus d'hématopoïèse de la Figure 2.

Chez le patient atteint de la LLC, ce processus d'apoptose est perturbé, ainsi une partie des anciens lymphocytes ne meurent plus et une autre partie se multiplie de façon anarchique. Au fur et à mesure de l'évolution de la maladie et de l'accumulation des lymphocytes matures clonales, les ganglions lymphatiques gonflent et peuvent être palpés tant au niveau du cou, que des aisselles ou de l'aîne. Les organes tels que la rate et le foie peuvent également augmenter de volume et prendre plus de place dans l'abdomen avec comme conséquence une sensation de douleur abdominale. Ainsi le patient se rend chez le clinicien soit pour un contrôle de routine, s'il est asymptomatique, soit par ce qu'il ressent des symptômes qui sont caractéristiques mais non spécifiques de la LLC :

- ✓ Fatigue
- ✓ Adénomégalie : inflammation chronique d'un ganglion
- ✓ Splénomégalie : augmentation du volume de la rate
- ✓ Hépatomégalie : augmentation de volume du foie.
- ✓ Complication infectieuse (plus rare)
- ✓ Cytopénie (plus rare)

Le clinicien prescrit alors une analyse sanguine en précisant, le cas échéant s'il suspecte une LLC.

### **3.1.2. Évolution de la maladie**

On distingue deux types d'évolutions possibles :

**La première est la mutation de la LLC vers une forme plus agressive :** La LLC pouvant rester stable pendant plusieurs années, il serait possible de recueillir des données Raman au cours du temps et construire une banque de données spectrales permettant par la suite de faire une étude rétrospective. Cette banque de données pourrait permettre d'identifier des changements biomoléculaires au cours du temps et identifier un ou plusieurs marqueurs

spectroscopiques prédictifs de l'évolution de la maladie et permettre d'établir un pronostic avant même que la maladie évolue vers la forme agressive.

**La deuxième correspond à l'apparition d'un lymphome de haut grade** plus agressif et qui remplace la LLC. Le syndrome de Richter (SR) correspond à l'évolution vers un lymphome diffus à grandes cellules chez un malade porteur d'une LLC. L'évolution vers cette maladie marque un tournant majeur pour le pronostic du malade, modifiant de façon significative sa prise en charge. Ainsi la connaissance du risque du SR pourrait favoriser un diagnostic précoce et donc une meilleure prise en charge du patient. Une étude rétrospective des spectres Raman recueillis sur les frottis sanguins des patients atteints par cette évolution pourrait mettre en évidence une augmentation de la variabilité des lymphocytes B pathologiques.

### **3.1.3. Méthodes de diagnostics hématologiques**

#### **a) Prise de sang, l'hémogramme**

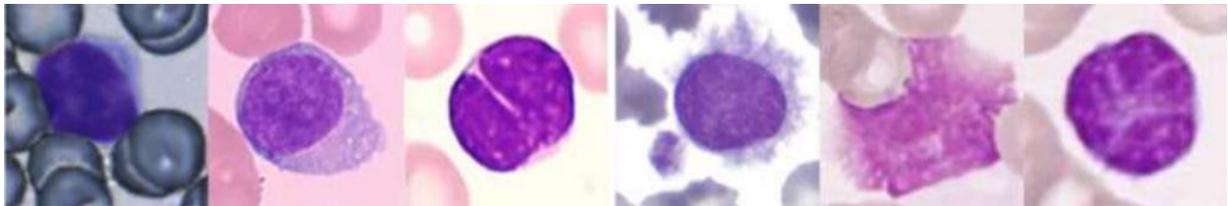
L'hémogramme ou numération sanguine (NFS) est un examen biologique de routine permettant de réaliser une analyse quantitative du sang c'est-à-dire une numération de tous les éléments figurés du sang : hématies, leucocytes, plaquettes ou réticulocytes, mais également une mesure de l'hématocrite, dosage de l'hémoglobine ou mesures des constantes érythrocytaires. Il permet d'évaluer l'état de santé général d'un patient. Il est prescrit pour diverses raisons, telles que la recherche de divers troubles, anémie, infections, lymphocytose, de voir l'état nutritionnel du patient ou de voir si le patient a été exposé ou non à une substance toxique. C'est l'examen de première ligne pour détecter une anomalie éventuelle dans le sang. Ce test ne constitue pas en lui-même un examen diagnostique, il est généralement couplé à d'autres analyses beaucoup plus poussées en cas d'anomalie. Le diagnostic de la LLC est posé par l'hémogramme lorsque le taux de lymphocytes B est supérieur à 5 G/L (gramme de lymphocyte par litre de sang) depuis plus 3 mois [32] et confirmé par d'autres analyses comme l'analyse morphologique par marquage ou coloration microscopique.

#### **b) Frottis sanguin, imagerie morphologique**

L'examen morphologique des lymphocytes sur les frottis sanguins colorés est la première étape clé pour le diagnostic de la leucémie et l'orientation vers un ensemble d'analyses complémentaires. En effet, elle permet d'orienter le clinicien vers une pathologie réactionnelle (comme une virose) ou une hémopathie. Cet examen permet d'observer la forme des cellules

sanguines (Figure 4). Pour ce faire, on étale manuellement ou à l'aide d'un automate une goutte de sang sur une lame de microscope puis on réalise une coloration. Dans le cas de la LLC, les lymphocytes B sont des cellules lymphoïdes matures monomorphes, de petites ou moyennes tailles, rondes ou légèrement irrégulières. Les lymphocytes B LLC possèdent une chromatine décondensée dite « mottée » ou en « écaille de tortue ».

Étant donné que les lymphocytes sont fragilisés lors d'une LLC, ils sont souvent détruits lors de la réalisation du frottis et seuls des résidus typiques sont encore visibles (Figure 4). Les spécialistes les désignent comme les « ombres de Gumprecht » [33-35]. Cependant, cet examen ne permet pas de spécifier exactement le type d'hémopathie ainsi la cytométrie en flux est obligatoire pour obtenir un diagnostic précis.



*Figure 4 - Morphologies de différents lymphocytes, de gauche à droite : lymphome folliculaire sain, réactif, leucémie à cellules velues, ombre de Gumprecht, lymphocytes à chromatine mottée (LLC)*

### **c) La cytométrie en flux (MFC)**

Cette technique est indispensable pour poser le diagnostic de la LLC. Comme le montre la Figure 5, elle consiste à rechercher sur la membrane cellulaire des lymphocytes B, des marqueurs typiques de cette pathologie, elle fournit donc des informations sur l'expression de molécules membranaires et/ou intracytoplasmiques (CD5, CD23, CD22, FMC7). L'utilisation d'anticorps marqués permet de détecter ces molécules, car elles sont spécifiquement reconnues. Les lignées cellulaires lymphocytaires produisent au hasard des immunoglobulines de surface qui sont constituées soit d'une chaîne légère soit  $\lambda$  ou soit  $\kappa$ . Dans le cas des patients non atteints de la LLC, les situations polyclonales conduisent à un équilibre entre les lymphocytes  $\lambda$  et  $\kappa$  dans le flux sanguin. Le taux  $\lambda/\kappa$  est donc étudié par immunocytochimie pour détecter la monoclonalité qui est une caractéristique maligne. L'immunophénotypage est complexe à mettre en œuvre et nécessite des réactifs spécifiques pour chaque pathologie. Cette analyse est donc coûteuse et nécessite une réelle expertise pour interpréter les résultats. Généralement cette technique n'est utilisée que dans les grands centres hospitaliers.

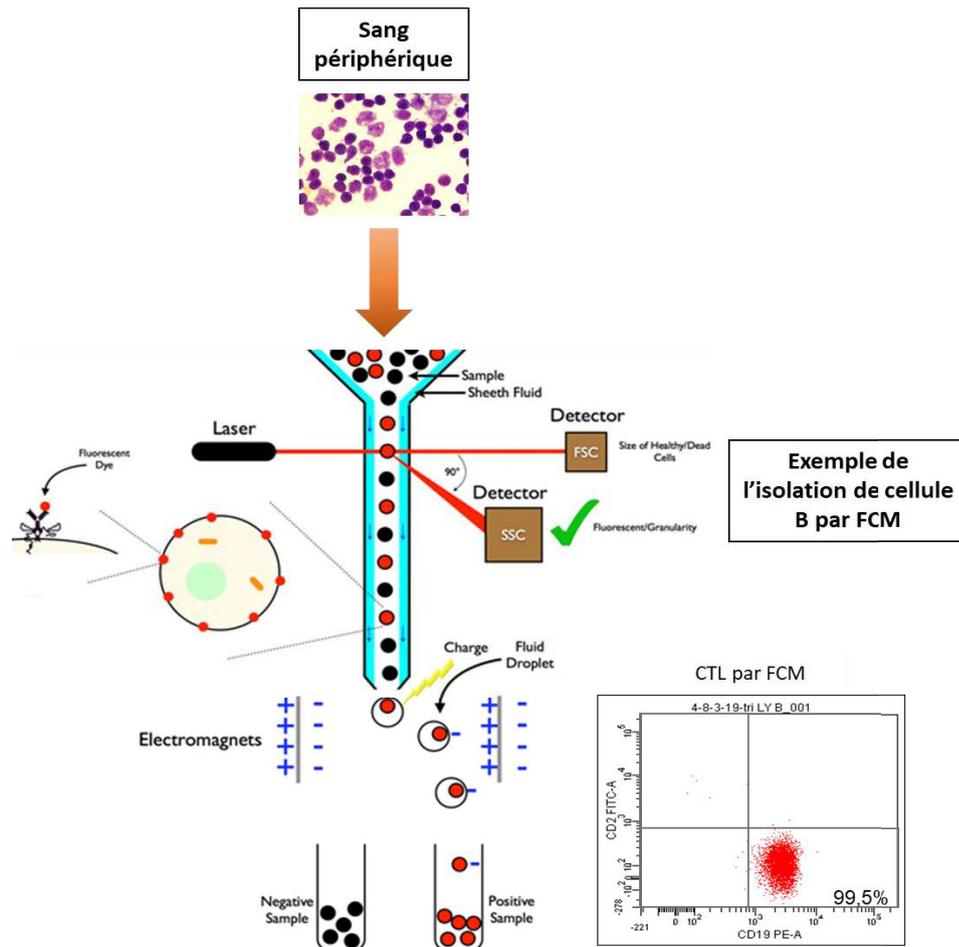


Figure 5 - Isolement des cellules B ou T/NK saines/LLC par MFC

#### d) Caryotype et étude cytogénétique moléculaire

La cytogénétique est une technique qui s'intéresse aux chromosomes. Son but est de détecter des anomalies chromosomiques afin d'appuyer un diagnostic biologique. Chaque hémopathie a ses propres anomalies permettant d'identifier le type de la maladie. Dans le cas d'une LLC, on recherche 4 types d'anomalies, les délétions 11q, 13q et 17 p (délétion = perte d'un fragment de chromosome) et la trisomie 12.

##### 3.1.4. Limites de ces techniques

Les techniques d'analyses actuelles de la LLC, à visée diagnostique ou pronostique sont entièrement automatisées et performantes, mais elles se heurtent à des limites non négligeables. En effet, elles requièrent un plateau technique lourd et du personnel médical et paramédical formé, ce qui demeure une limite dans le système de diagnostic actuel. Leur capacité à détecter une LLC réside essentiellement sur des méthodes de coloration, d'immunohistologie

(immunofluorescence et immunohistochimie) et sur des interprétations humaines, ce qui implique qu'en dépit de la grande expérience de l'hématopathologiste, il existe un pourcentage d'erreur intra et interindividuelle significatif. De même, lors de la suspicion d'une LLC chez un patient, dans 93% des cas celui-ci se révèle souffrir d'une autre hémopathie chronique maligne après analyse. Ceci induit que pour 14 patients sur 15, des tests relatifs au diagnostic et au pronostic de la LLC sont réalisés sans que ceux-ci ne soient nécessaires. En raison de toutes ces limites, il est indispensable de développer des technologies innovantes pouvant compléter ou remplacer les techniques actuelles qui sont chronophages et demandent beaucoup de personnels et de consommables. Les techniques de spectroscopie vibrationnelle comme la diffusion Raman sont des outils potentiellement intéressants dans ce contexte.

## 3.2. Le pronostic

### 3.2.1. Les différents stades

En Europe et plus particulièrement en France, la LLC est classée selon trois stades qui est fonction de la sévérité (**stades de Binet : 1981**) : A, B et C. Cette classification (Figure 6) tient en compte les résultats de l'hémogramme et le nombre d'atteintes d'aires ganglionnaires (ganglions cervicaux, ganglions axillaires, ganglions inguinaux, foie, rate).

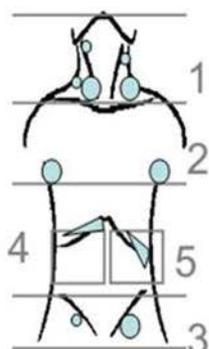
Un patient est dans le **stade A** (plus  $\frac{3}{4}$ ) lorsque moins de trois aires ganglionnaires sont enflées et palpables. Le nombre de globules rouges et de plaquettes est normal donc il n'y a ni présence d'anémie ni de thrombopénie. Ce stade présente un risque faible pour la santé du patient donc aucun traitement n'est nécessaire, mais une surveillance tous les 6 à 12 mois est nécessaire, ainsi qu'un examen clinique et certains paramètres biologiques simples pour vérifier une éventuelle progression de la LLC.

Dans le **stade B**, trois aires ganglionnaires ou plus sont enflées. Le nombre de globule rouge et de plaquettes est normal. Ce stade présente un risque moyen pour la santé du patient. A ce stade une prise en charge thérapeutique est nécessaire s'il y a présence de signes généraux, évolutivité rapide, splénomégalie volumineuse ou une masse ganglionnaire volumineuse. Environ 15% des patients sont en stade B.

Dans le **stade C**, le patient présente une anémie et/ou une thrombopénie. Ce stade présente un risque élevé pour la santé du patient, quel que soit le nombre d'aires lymphoïdes palpables. Une prise en charge thérapeutique est nécessaire. Le taux d'hémoglobine est inférieur

à 10g/100 ml et le nombre de plaquettes est inférieur à 100 000/mm<sup>3</sup>. Environ 5% des patients sont en stade C.

La nature du traitement pour les stades B et C est proportionnelle à la gravité et à la progression de la maladie. Avec les traitements conventionnels, les médianes de survie varient de 2 à plus de 10 ans.



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

70 à 80% de stades A au moment du diagnostic

Figure 6 - Le stade de Binet permet au clinicien selon différents critères d'évaluer le stade de la leucémie lymphoïde chronique [36]

### 3.2.2. Les facteurs pronostiques

La leucémie lymphoïde chronique se caractérise par une évolution clinique variable d'un patient à l'autre, en effet certains patients restent stables pendant des années tandis que d'autres auront une maladie qui évoluera plus rapidement vers une forme plus agressive. Cette pathologie est donc très hétérogène et son niveau d'agressivité dépend directement de plusieurs critères biologiques, cytogénétiques et phénotypiques. Il est donc essentiel de réaliser un diagnostic précoce et individuel pour chaque patient. Malheureusement, tous les facteurs pronostiques ne peuvent être pris en compte en routine clinique lors d'un diagnostic. On peut répertorier deux grandes familles de facteurs pronostiques [37, 38] ; **de bon pronostic ou mauvais pronostic** en fonctions de différents critères (tableau 2).

Tableau 2 - Résumé de différents facteurs pronostiques en fonction de leur issue bonne ou mauvaise

<b>Bon pronostic</b>	<b>Mauvais pronostic</b>
Stade A	Stades B ou C
Pas de signe de prolifération	Signe de prolifération
<ul style="list-style-type: none"> <li>▪ TDL long &gt; 12 mois</li> <li>▪ Thymidine kinase sérique normale</li> <li>▪ <math>\beta</math>2M normale</li> </ul>	<ul style="list-style-type: none"> <li>▪ TDL court (&lt; 12 mois)</li> <li>▪ Thymidine kinase sérique élevée</li> <li>▪ <math>\beta</math>2M élevée</li> </ul>
CD38 (-)	CD38 (+)
Présence de mutations VH	Absence de mutation VH
ZAP70 (-)	ZAP70 (+)
Délétion 13q isolée ou absence d'anomalie cytogénétique	Délétion 11q, délétion 17p, translocation (14;19), caryotype complexe

Il existe 3 facteurs pronostiques principaux :

- **Le temps de doublements des lymphocytes (TDL)** permet de calculer le temps nécessaire pour que le nombre de lymphocytes soit multiplié par deux. Si un patient possède un TDL inférieur à 12 mois, il est dans une phase évolutive de sa maladie, avec un très mauvais pronostic et vice versa.
- **Les marqueurs sérologiques ou critères solubles** tels que la  $\beta$ 2-microglobuline ( $\beta$ 2M), la thymidine-kinase (TK) ou la CD23 soluble (sCD23) sont des marqueurs de pronostic corrélés à la progression de la maladie, s'ils sont supérieurs à la normale, ceci est généralement associé à un mauvais pronostic pour le patient. Ces marqueurs restent toutefois limités par l'usage de valeurs seuils qui varient d'un centre d'analyse à l'autre.

- **Le CD38** est un marqueur membranaire qui a pour but de montrer l'activation et la maturation lymphocytaire. Son pourcentage d'expression sur les lymphocytes est mesuré par cytométrie en flux. L'augmentation de son taux d'expression est la marque d'anomalie morphologique, et d'un mauvais pronostic.

La littérature est abondante en matière de facteurs pronostiques, mais ils ne sont pas tous systématiquement utilisés. L'étude des anomalies chromosomiques, l'étude du profil mutationnel des immunoglobulines (IGs) et la recherche de l'expression de ZAP70 en font partie.

La mutation des IGs (anticorps) permet de séparer en deux groupes les patients atteints de LLC selon deux degrés d'évolution distinctes. Les patients ayant la mutation possèdent une évolution favorable et ont peu de risque de développer une maladie agressive. Les patients ne possédant pas la mutation ont un profil à risque plus élevé pour développer une maladie évolutive et donc possèdent une survie raccourcie.

L'expression de la protéine ZAP70 [39] (Zeta-Associated-Protein-70) chez certains lymphocytes B tumoraux peut être corrélée au profil non muté des gènes des IGs, de ce fait, représente un mauvais pronostic. Malheureusement, cette corrélation n'est pas toujours vraie, ce qui implique qu'utiliser seul cette méthode n'apparaît pas toujours fiable.

L'analyse cytogénétique [39, 40] permet d'identifier des anomalies relatives au génome (anomalies chromosomiques). Il existe 4 types d'anomalies possibles : les délétions 11q22, 13q14 et 17p13 et la trisomie 12. La technique FISH (Fluorescence In Situ Hybridization) est l'examen de référence pour déterminer les facteurs pronostiques cytogénétiques. Il est à noter qu'au vu de la complexité et le coût important de ces tests, ces techniques restent peu abordables pour une application en routine clinique

### **3.3. Approches thérapeutiques**

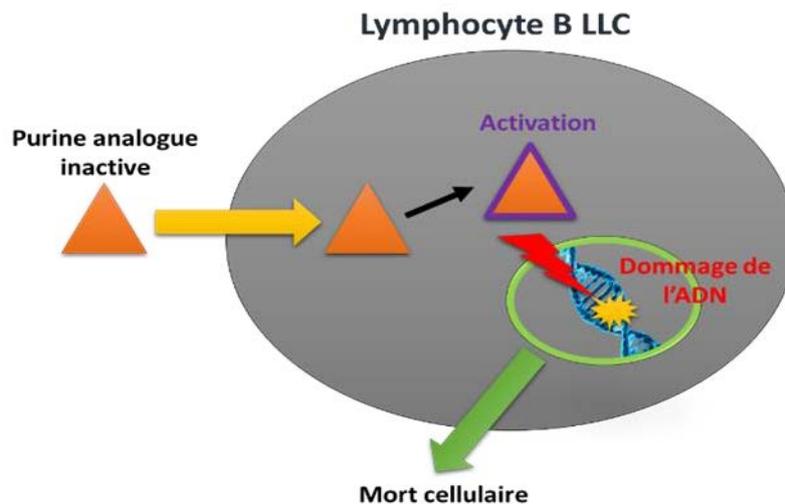
Généralement pour les formes précoces de la LLC (stade A), il n'est pas essentiel de mettre en place une solution thérapeutique, juste une surveillance attentive est nécessaire. Chez certains patients, une accumulation des cellules tumorales peut entraîner des infections récurrentes et/ou des cytopénies. Ainsi dans le cas d'une progression importante de la LLC et en fonction de l'état général du patient, un traitement adéquat sera mis en place. Il existe de nombreuses options thérapeutiques qui évoluent de façon régulière chaque année. Le principal objectif du traitement de la LLC est de ralentir l'accumulation de lymphocytes tumoraux,

revenir à un nombre de cellules sanguines normales et maintenir la qualité de vie des patients. Pour chaque patient, un schéma thérapeutique sera choisi parmi les nombreuses options existantes et adaptées en fonction du stade clinique, de la présence ou non de délétions génétiques et/ou des réponses antérieures des patients au traitement. Ces options de traitement sont présentées dans le tableau 3.

**Tableau 3 - Résumé des principales thérapies existantes pour la LLC, les molécules courantes utilisées et leur action**

Méthode thérapeutique	Mécanisme cible	Molécules courantes	Action	Références
<b>Agents chimiothérapeutiques</b>	Anthracyclines	Doxorubicine	Bloque la réplication et la transcription	[41, 42]
	Agents alkylants	Chlorambucil, Cyclophosphamide	Bloque la réplication et la transcription. Ils ont <i>in fine</i> un effet pro-apoptotique sur les cellules tumorales	[30, 43]
	Analogues purines	Fludarabine, Cladribine	Inhibent la synthèse d'ADN entraînant un arrêt de la croissance cellulaire	[44, 45]
	Immunomodulateurs Inhibiteurs BTK et BCL-2 Inducteurs d'apoptose	Ibrutinib, acalabrutinib, navitoclax, venetoclax, idelalisib, Flavopiridol	Agissent sur les voies de signalisation de survie des lymphocytes B leucémiques, inhibent la synthèse des molécules anti-apoptotiques et/ou augmentent la synthèse de molécules pro-apoptotiques	[30, 43, 46-50]
<b>Immunothérapie</b>	Anticorps	Alemtuzumab, rituximab	Inducteur d'apoptose, lyse par complément, cytotoxicité cellulaire anticorps-dépendante	[43, 51-55]
<b>Allogreffe</b>	Greffe de cellules souches hématopoïétiques	Cellules souches d'un donneur compatible	La greffe de cellules souches permet de remplacer les cellules souches qui forment le sang (hématopoïétiques) d'une personne. Les nouvelles cellules souches fabriquent des cellules sanguines saines	[56]

Les principaux médicaments utilisés dans le traitement d'une LLC sont pour la majorité des cas, des molécules de chimiothérapie telles que les anthracyclines, les agents alkylants ou les analogues des purines qui agissent directement sur la synthèse de l'ADN (Figure 7) et qui sont très efficaces pour éliminer les cellules cancéreuses de la LLC. Malgré leur efficacité avérée, ces médicaments sont toxiques également pour les lymphocytes sains, entraînant une forte sensibilité des patients aux pathogènes.

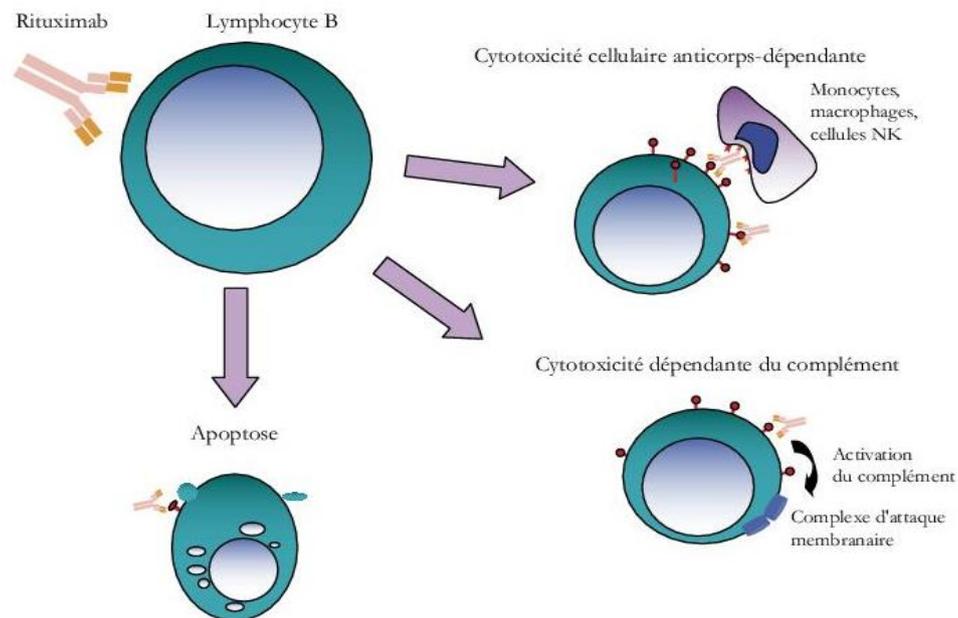


*Figure 7 - Schéma présentant l'effet des analogues aux purines sur un lymphocyte B LLC - Les purines ont pour effet de bloquer la réplication de l'ADN du lymphocyte, ceci va engendrer l'apoptose du lymphocyte*

Depuis quelques années, on a pu voir une réelle évolution dans les nanotechnologies et en particulier dans le domaine des thérapies ciblées [43]. La combinaison de la compréhension du mécanisme de prolifération des cellules cancéreuses et des connaissances sur les thérapies ciblées ont permis de développer de nouveaux médicaments agissant directement sur les principales voies de signalisation qui interviennent dans la survie cellulaire et la prolifération leucémique. Parmi ces nouvelles molécules, il y a les modulateurs, les inhibiteurs ou les inducteurs d'apoptoses. Ces nouvelles molécules semblent prometteuses, notamment l'inhibiteur ibrutinib (PCI-32765) de Bruton tyrosine kinase ou BTK [46] ou l'inhibiteur d'isoforme delta de la phosphoinositide 3-kinase (PI3K) [47].

Les schémas de traitement de la LLC ont évolué en remplaçant progressivement les monothérapies par des agents combinant chimiothérapie et immunothérapie [30]. Ainsi des anticorps monoclonaux humanisés ou chimériques, dirigés contre les antigènes de surfaces des cellules, ont vu le jour, permettant une déplétion des lymphocytes B chez le patient atteint de

la LLC. Ces molécules jouent sur 3 mécanismes principaux : induction d'apoptose, lyse par le complément et cytotoxicité cellulaire anticorps-dépendante. Un exemple peut être vu sur la Figure 8 qui présente les étapes d'action du Rituximab (anticorps monoclonal chimérique)



*Figure 8 - Schéma représentatif des étapes d'action du Rituximab sur un lymphocyte B - Le rituximab s'accroche spécifiquement à l'antigène CD20 des lymphocytes pré-B et B matures. Cet antigène s'exprime dans plus de 95 % des cellules B [55]*

Dans le cas où les traitements conventionnels n'ont aucun effet sur le patient, il est possible d'effectuer une transplantation de cellules souches hématopoïétiques. Mais cet acte n'est pas anodin pour le patient et demande la prise en compte de critères d'éligibilité [56] comme l'âge, l'état général du patient, absence de réponse ou rechute rapide et surtout la compatibilité du donneur.

Bien qu'il existe de nombreuses thérapies, la LLC est une maladie encore incurable, car il n'est pas possible de détruire toutes les cellules cancéreuses de l'organisme (cellules résistantes au traitement), de plus les traitements actuels visent à ralentir sa progression et à en contrôler les symptômes. Les traitements pour lutter contre la LLC peuvent provoquer des effets secondaires comme des anémies ou thrombopénies. La recherche de nouveaux traitements est encore d'actualité afin de trouver le meilleur schéma thérapeutique possible. Néanmoins, une nette augmentation du taux de survie a été observée [57] (+8 points à 5 ans entre 1989-1993 et 2005-2010). Cette augmentation est due principalement à une meilleure compréhension de la LLC, de meilleurs facteurs pronostics, une amélioration de la prise en charge des patients et du schéma thérapeutique tel que l'association de la chimiothérapie et la thérapie ciblée.

L'émergence de nouvelles molécules permettra peut-être d'améliorer encore plus ce taux de survie ou d'atteindre une éradication totale de la maladie.

#### 4. Les techniques de spectroscopie vibrationnelle

##### 4.1. Introduction

Lors d'une interaction entre un rayonnement électromagnétique et la matière (Figure 9), on observe divers phénomènes physiques tels que l'émission, l'absorption ou la diffusion. Ce rayonnement électromagnétique correspond à un processus vibratoire caractérisé par 3 grandeurs physiques que sont sa vitesse de propagation dans le vide  $c$ , sa fréquence  $\nu$  et sa longueur d'onde  $\lambda$ . Elles sont liées par la relation de **Planck-Einstein** exprimée par l'équation 1 :

$$E = h\nu = hc/\lambda \quad (\text{éq. 1})$$

Où :

- $E$  est l'énergie de l'onde électromagnétique
- $\nu$  est la fréquence de l'onde
- $h$  : la constante de Planck ( $6,625 \cdot 10^{-34}$  J.s)
- $c$  : célérité de la lumière dans le vide ( $3 \cdot 10^8$  m. s<sup>-1</sup>)
- $\lambda$  : longueur d'onde de la lumière

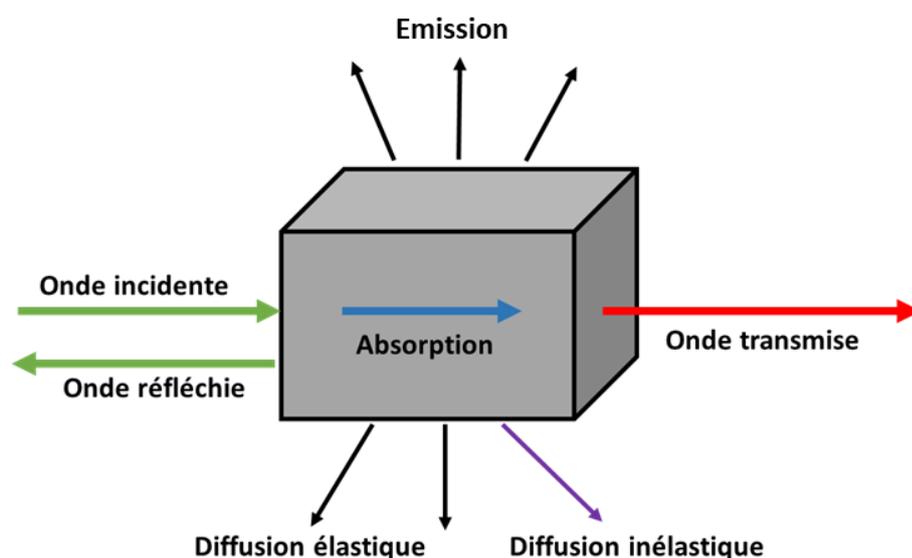


Figure 9 - Représentation schématique des différents phénomènes lors d'une interaction d'une lumière/matière

L'analyse de ces différents phénomènes d'interactions, grâce aux techniques de spectroscopie optique, permet de caractériser la composition de la matière en accédant aussi à

des informations d'ordre structural. Selon la gamme du spectre électromagnétique analysée, il existe diverses techniques de spectroscopie, telles que les spectroscopies d'absorption X (10 nm -100 pm) et d'absorption ultraviolet (UV : 200-400 nm) – visible (400-750 nm) ou d'émission de fluorescence dont le principe repose sur la mise en évidence de variation des distributions électroniques atomiques ou moléculaires. En outre, on distingue aussi les spectroscopies vibrationnelles qui permettent de sonder les liaisons chimiques composant un échantillon moléculaire (Figure 10).

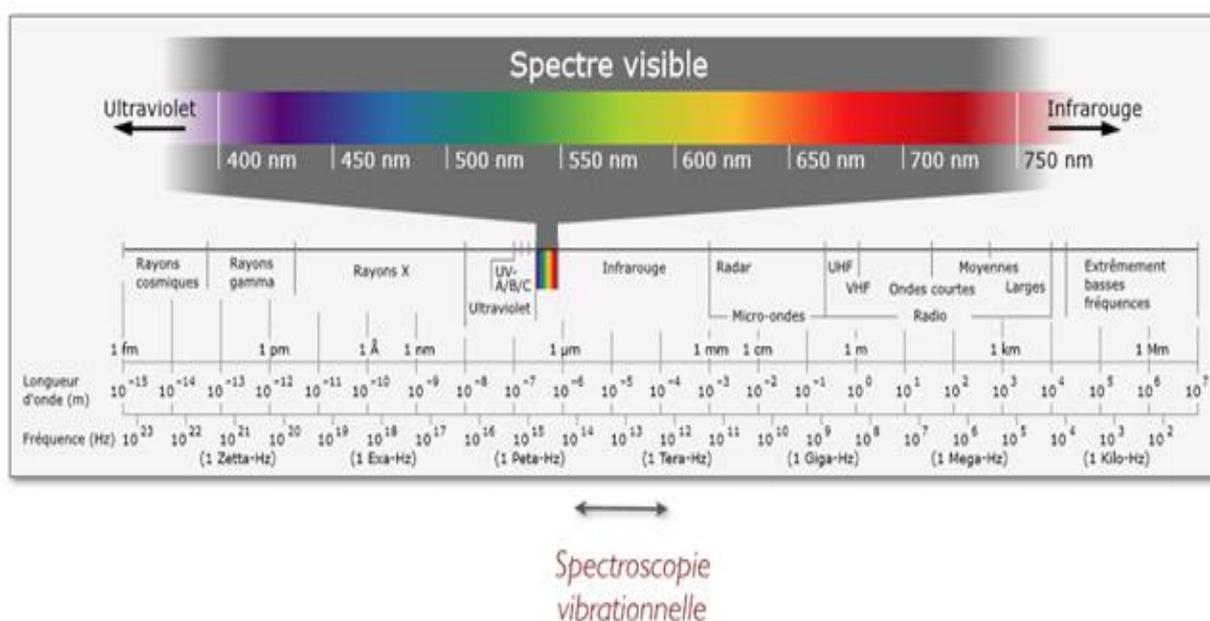


Figure 10 - Ensemble des domaines électromagnétiques

#### 4.1.1. Les techniques vibrationnelles

Les spectroscopies vibrationnelles sont des techniques permettant d'avoir un accès direct à l'empreinte moléculaire ou biochimique d'un échantillon sans en altérer son intégrité. Ces techniques sont très utilisés dans différents domaines, tel que la science des médicaments, la Physique des matériaux, en agroalimentaire et récemment dans le domaine de la médecine translationnelle et en biologie. L'émergence de ces technologies est principalement due aux avancées technologiques des années 70, avec le développement des interféromètres et l'apparition des premiers lasers.

Il existe deux types de spectroscopies vibrationnelles, la spectroscopie de diffusion Raman et la spectroscopie d'absorption infrarouge. Elles reposent sur l'interaction de la lumière avec la matière, ce qui provoque des transitions sur les niveaux vibrationnels [58, 59]. L'onde

incidente ne peut interagir qu'avec les modes de vibrations des molécules car son énergie est faible. Ceci permet d'avoir des informations structurales d'ordre moléculaire. En effet, les liaisons chimiques entre les atomes d'une molécule sont élastiques, les molécules peuvent donc se mouvoir et vibrer autour de leur centre de gravité (Figure 11). Ces vibrations peuvent être théorisées selon trois modèles : modèle harmonique, modèle quantique et anharmonique.

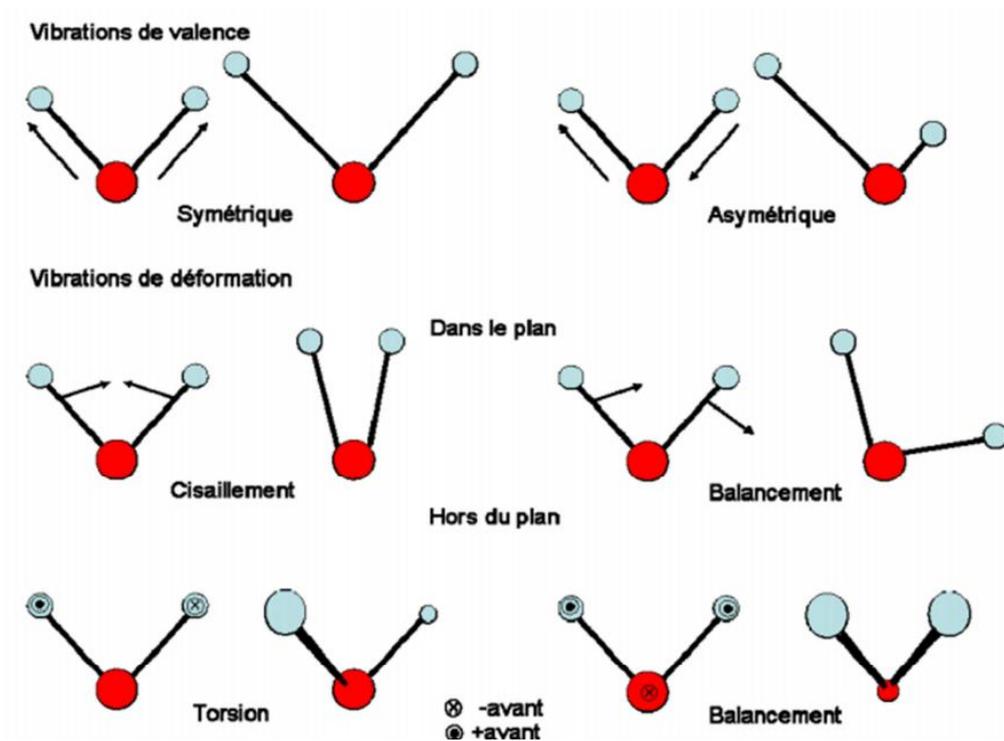


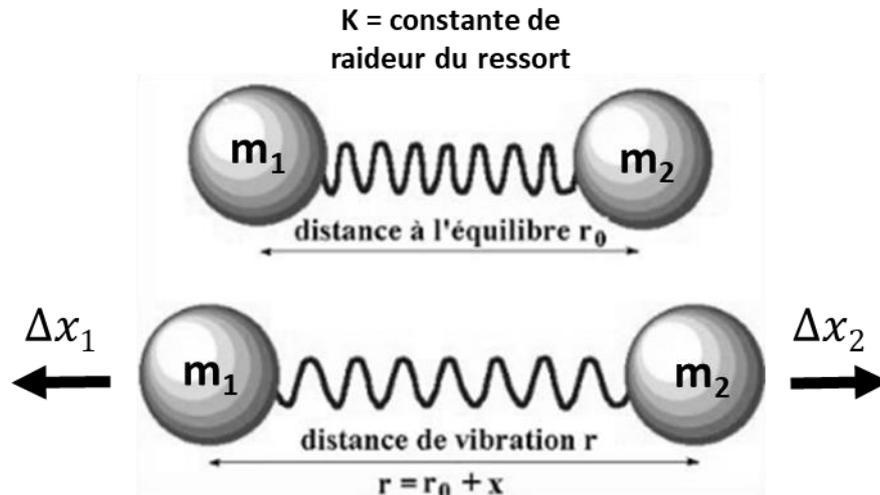
Figure 11 - Représentation schématique de vibration de valence

### - Modèle de l'oscillateur harmonique

Dans le cas idéal, en première approximation, les vibrations moléculaires peuvent être décrites selon le modèle de l'oscillateur harmonique (Figure 12) qui obéit à la loi de Hooke. Soit une molécule constituée de deux atomes de masses  $m_1$  et  $m_2$  reliés par un ressort (liaison chimique) de constante de raideur  $k$ , séparés par une distance  $r_0$  à l'équilibre et dont la masse réduite  $\mu$  est définie par :

$$\frac{1}{\mu} = \frac{1}{m_1} + \frac{1}{m_2} \quad (\text{éq. 2})$$

$$\mu = \frac{m_1 \cdot m_2}{m_1 + m_2} \quad (\text{éq. 3})$$



*Figure 12 - Oscillateur mécanique*

$\Delta x_1$  et  $\Delta x_2$  correspondent aux déplacements des masses  $m_1$  et  $m_2$  par rapport à leurs positions d'équilibre  $x_1$  et  $x_2$ .

Soit la variable d'élongation  $q$  du ressort (liaison) lorsque le système sort de son état d'équilibre qui vaut :  $q = r - r_0 = \Delta x_1 - \Delta x_2$

Par introduction de l'énergie cinétique du système, nous obtenons :

$$T = \frac{1}{2} \sum m_i \cdot v_i^2 = \frac{1}{2} \mu \frac{dq^2}{dt} \quad (\text{éq. 4})$$

$$\text{avec } \mu = \frac{m_1 \cdot m_2}{m_1 + m_2}$$

Car  $m_1 x_1 = m_2 x_2$

Maintenant il faut déterminer l'énergie potentielle  $V$  du système qui est défini par l'équation :

$$V = \int -F dq = \frac{1}{2} k q^2 \quad (\text{éq. 5})$$

avec la force de rappel associée au mouvement qui vaut :  $F = -kq$

$$\text{Soit } \frac{d}{dt} \left( \frac{\partial T}{\partial \dot{q}} \right) + \frac{\partial V}{\partial q} = 0$$

$$\text{Ce qui donne } \mu \ddot{q} + kq = 0$$

Le système possède un mouvement sinusoïdal :

$$\text{On a donc } Q = a \cos(\omega t + \phi) \text{ où } \omega^2 = \frac{k}{\mu}$$

$$\text{Le système vibre donc à une fréquence de } \nu = \frac{1}{2\pi} \sqrt{\frac{k}{\mu}} \text{ (Hertz).}$$

En spectroscopie, on utilise couramment le nombre d'onde ( $\text{cm}^{-1}$ ) qui vaut :

$$\bar{\nu} = \frac{\nu}{c} = \frac{1}{2\pi c} \sqrt{\frac{k}{\mu}}$$

Cette grandeur dépend ainsi de la masse réduite  $\mu$  et de la constante de force de liaison  $k$ . L'énergie potentielle agissant sur les liaisons de vibration vaut:

$$E_p = \frac{1}{2} kq^2 = \frac{1}{2} k(\mathbf{r} - \mathbf{r}_0)^2 \text{ (éq. 6)}$$

On peut en déduire que le système global possède la même énergie potentielle et la même énergie cinétique qu'une masse unique  $\mu$  (la masse réduite), c'est-à-dire que le système est représenté par une masse liée à un ressort fixé en un point fixe. L'énergie potentielle  $E_p$  forme une parabole en fonction de  $(\mathbf{r}-\mathbf{r}_0)$  (Figure 13).

### - Modèle quantique

Cependant la mécanique classique ne suffit pas pour décrire complètement ce modèle, car elle n'introduit pas la quantification des niveaux énergétiques, c'est pourquoi on fait appel à la mécanique quantique. En effet, les énergies de vibrations sont quantifiées et sont représentées par la résolution de l'équation générale de Schrödinger suivante :

$$\frac{\hbar^2}{2m} \Delta \Psi + V\Psi = E\Psi \quad \text{(éq. 7)}$$

En introduisant l'Hamiltonien dans la fonction de Schrödinger, on obtient l'équation suivante :

$$\frac{\hbar^2}{8\pi^2\mu} \frac{d^2\Psi(\mathbf{q})}{dq^2} + \frac{1}{2}kq^2 \cdot \Psi(\mathbf{q}) = E \cdot \Psi(\mathbf{q}) \quad (\text{éq. 8})$$

La résolution de l'équation 8 permet d'avoir les niveaux d'énergie vibrationnelle  $E_v$  et amène à la solution suivante :

$$E_v = \left(v + \frac{1}{2}\right) h\nu = \left(v + \frac{1}{2}\right) hc\bar{\nu} \quad (\text{éq. 9})$$

Avec  $v = 0, 1, 2 \dots$

$v$  correspond au nombre quantique de vibration c'est-à-dire le niveau d'énergie du système et  $\nu$  correspond à la fréquence de vibration.

Au vu de ces équations, la mécanique quantique permet donc d'introduire la quantification des niveaux énergétiques de vibrations. Ces niveaux quantifiés sont équidistants où la différence entre deux niveaux consécutifs est égale à  $h\nu$ . D'autre part si  $v = 0$ , alors  $E_v = 1/2 h\nu$ , c'est-à-dire que l'énergie de vibration n'est jamais nulle.

#### - Modèle réel

Dans le cas des molécules réelles, le modèle harmonique n'est plus valable. En effet, une molécule réelle n'obéit pas exactement à la loi de Hooke, le puit de potentiel  $E_p$  ne forme plus une parabole mais plutôt une courbe de Morse (Figure 13). En effet, les liaisons chimiques ne sont pas des ressorts parfaits, il en résulte que les niveaux de vibrations ne sont plus équidistants, en effet si  $v$  augmente la différence entre deux niveaux d'énergie consécutifs diminue. Si la molécule tend à se dissocier, la différence entre deux niveaux d'énergie tend vers 0.

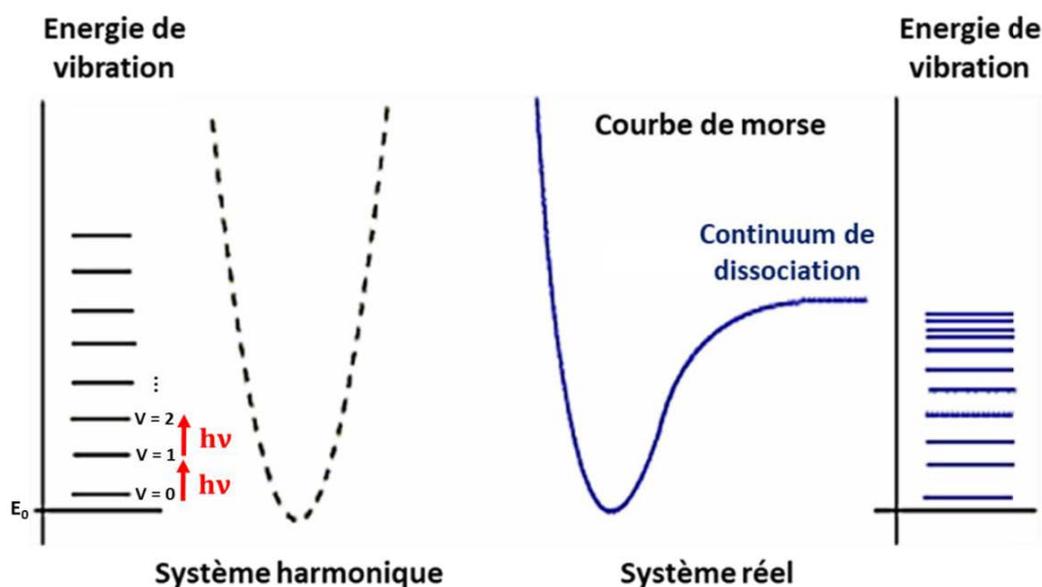


Figure 13 - Niveaux vibrationnels d'un système harmonique et d'un système réel

#### 4.1.2. Règles de sélection en infrarouge et Raman

Les techniques vibrationnelles sont actuellement en plein développement, de nombreuses études et publications scientifiques apparaissent régulièrement dans de nombreux domaines de la recherche et de l'industrie, tels que la biologie, la physique, biotechnologie, les nanosciences et la médecine. L'infrarouge (proche et moyen) et le Raman reposent sur des règles de sélection pour permettre les transitions vibrationnelles.

Dans le cadre de l'infrarouge, la vibration doit engendrer une variation du moment dipolaire permanent de la molécule. En ce qui concerne le Raman, il doit y avoir une variation du moment dipolaire induit de la molécule, c'est-à-dire un changement dans sa polarisabilité [58]. La polarisabilité représente la façon dont le nuage électronique est capable de se déformer sous l'effet d'un champ électrique  $\vec{E}$ .

Une vibration très active en infrarouge engendrera un signal faible en Raman et vice-versa ça. C'est-à-dire que si la molécule possède un centre de symétrie, il n'existera aucune vibration commune en infrarouge et en Raman. Les vibrations symétriques seront donc actives en Raman mais inactives en infrarouge. Dans le cas contraire, les vibrations antisymétriques seront inactives en Raman mais actives en infrarouge (règle de l'exclusion mutuelle), c'est ce qui fait la complémentarité des deux techniques. Certaines vibrations peuvent n'apparaître ni en spectroscopie Raman ni en spectroscopie infrarouge. Si la molécule possède au moins un

axe de symétrie d'ordre supérieur à deux, des modes dégénérés apparaissent. Les modes de vibrations dégénérés sont alors confondus en une seule raie. D'autres vibrations peuvent également être actives à la fois en Raman et en infrarouge, comme il est possible de voir sur la Figure 14 avec les liaisons C=C et C=O de l'ester méthylique d'acide oléique.

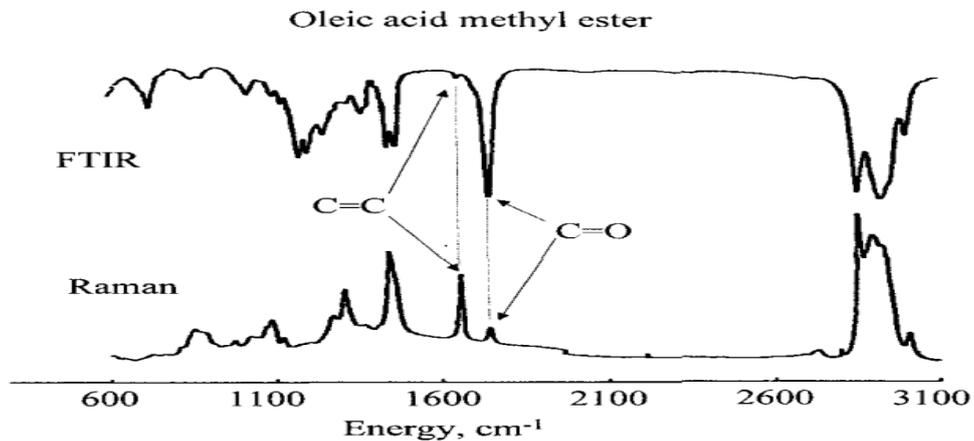


Figure 14 - Spectre infrarouge et Raman d'ester méthylique d'acide oléique. [60]

Chaque molécule possède un riche éventail de transitions vibratoires. Une vibration moléculaire est représentée par un pic sur le spectre comme sur la Figure 15. La spectroscopie vibrationnelle consiste donc à déterminer la fréquence de vibrations des molécules, ces fréquences dépendent de l'énergie des liaisons entre les atomes constituant la molécule. Il est donc possible d'extraire d'un spectre des informations structurelles ou sur la nature d'une molécule.

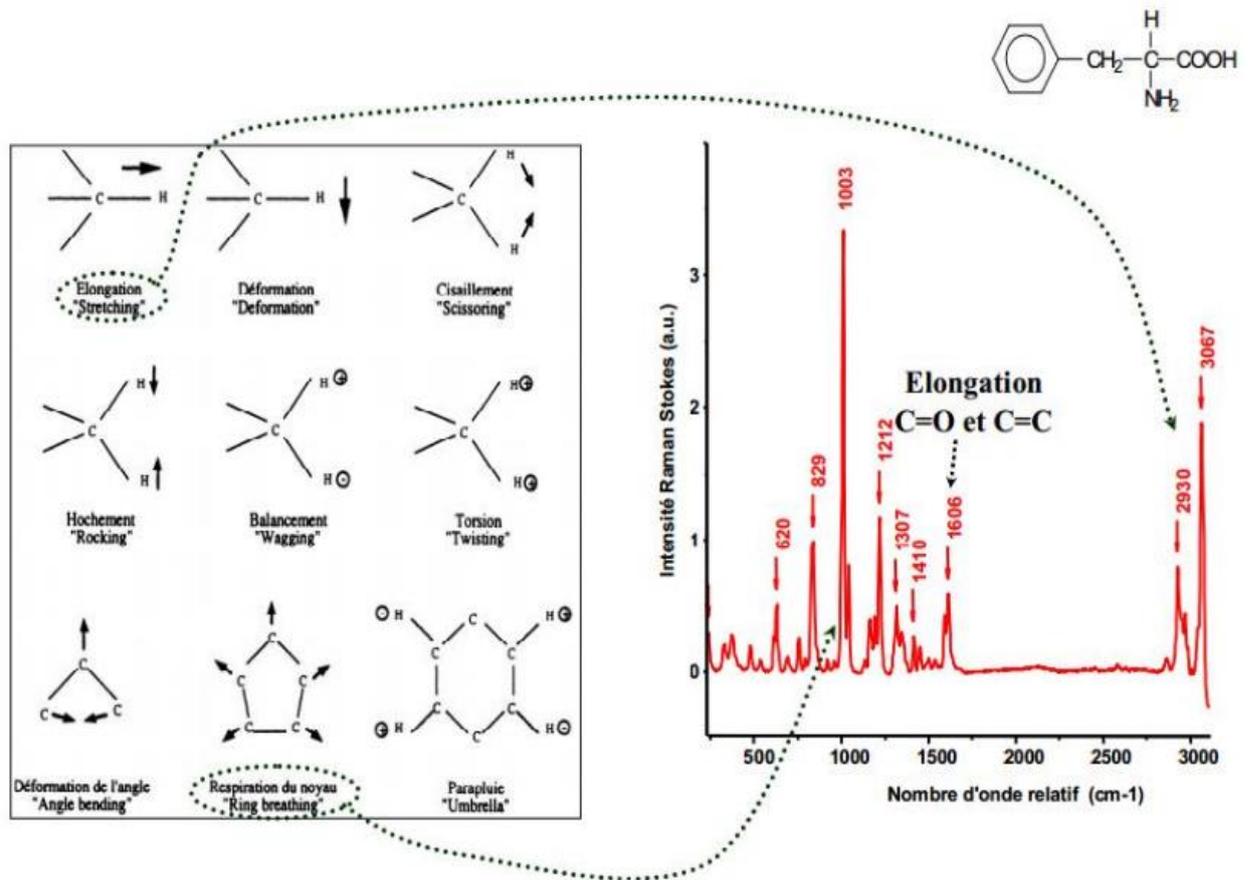


Figure 15 - Exemples de modes fondamentaux de vibration moléculaire et de leurs pics Raman associés

## 4.2. Spectroscopie d'absorption Infrarouge

### 4.2.1. Historique

Situé entre la région du visible et des ondes hertziennes, le domaine du rayonnement infrarouge (IR) comme vu précédemment s'étend de 0.8 $\mu\text{m}$  à 1000  $\mu\text{m}$  et a été découvert par Frédéric Wilhelm Hershel en 1800. C'est seulement 124 ans plus tard, que les scientifiques se sont aperçus qu'il existait une concordance entre l'énergie interne des molécules et l'énergie du rayonnement IR, permettant ainsi de relier l'absorption IR d'une molécule et sa structure chimique.

### 4.2.2. Considérations théoriques

Comme nous avons pu le voir précédemment, l'interaction entre la lumière et la matière crée différents phénomènes. Généralement, le phénomène d'absorption énergétique de la lumière incidente (laser), permet grâce à la forte énergie des photons, le passage des électrons de l'état fondamental à un état plus énergétique « excité » (exemple : phénomène de

fluorescence ou phosphorescence). Dans le cas de l'absorption infrarouge, l'énergie des photons incidents issus d'une lumière polychromatique n'est pas suffisamment grande pour provoquer ces transitions électroniques au sein du cortège mais induit des variations au sein des niveaux d'énergie vibrationnelle.

Le domaine IR peut être subdivisé en trois régions selon l'application : le lointain IR (4000 – 13000  $\text{cm}^{-1}$ ), le moyen IR (400 – 4000  $\text{cm}^{-1}$ ) qui permet d'étudier les vibrations moléculaires et le proche IR (10 – 400  $\text{cm}^{-1}$ ) qui permet l'étude de vibrations plus complexes comme les harmoniques. Ainsi lors de l'analyse, le rayonnement polychromatique incident est envoyé sur l'échantillon et le rayonnement transmis est mesuré. Si l'intensité du faisceau transmis est inférieure à l'intensité du faisceau incident, l'échantillon a absorbé alors une partie du rayonnement.

Dans le cas de la spectroscopie moyen IR, cette absorption se produit lorsque l'énergie (ou fréquence) des photons incidents, coïncide avec l'énergie nécessaire pour faire vibrer certaines liaisons chimiques [61]. On représente ainsi le spectre infrarouge par une absorbance  $A$  ou transmittance  $T$  en fonction du nombre d'onde.

$$\tilde{\nu}(\text{cm}^{-1}) = \frac{10^4}{\lambda (\text{m})}$$

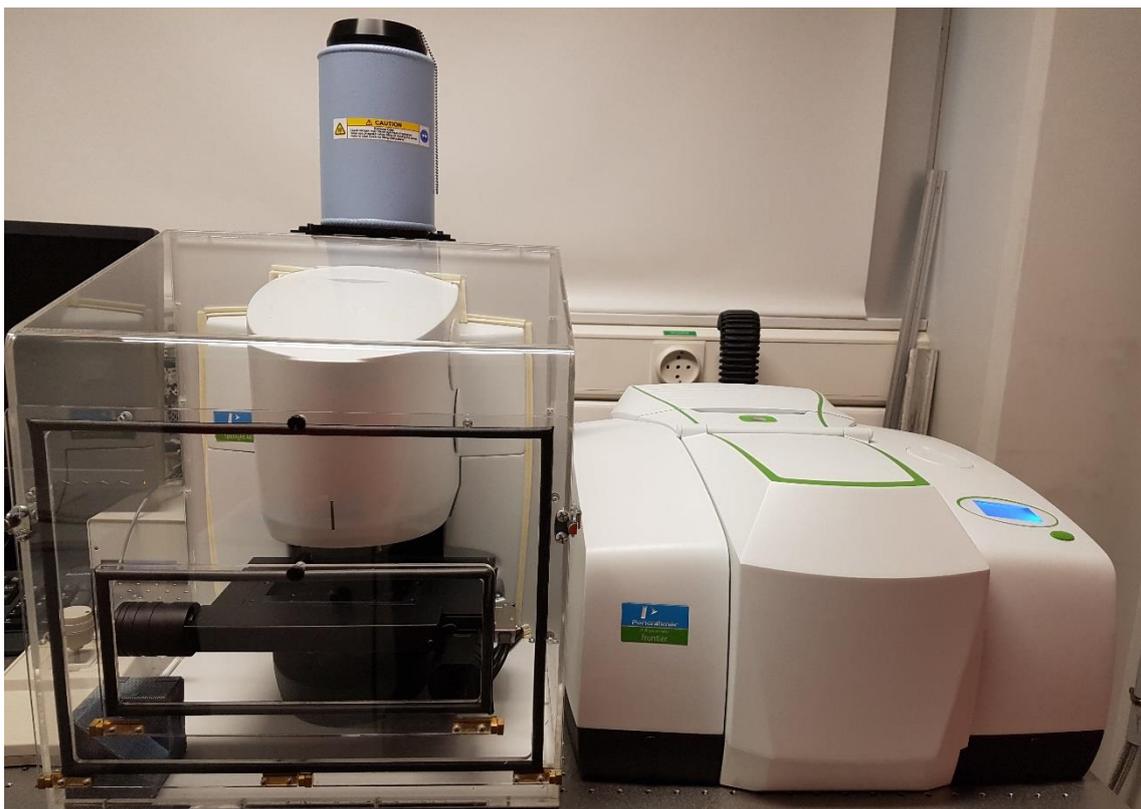
$$\text{Soit } A = \log_{10} \left( \frac{I_0}{I} \right) = \log_{10}(T)$$

$T$  : transmittance (%)

$I_0 = I_{\text{source}}(\mathbf{x})$  : intensité de la source polychromatique incidente

$I$  : intensité transmise après le passage de l'échantillon

Cette technologie nécessite l'utilisation de supports transparents pour l'infrarouge : bromure de potassium (**KBr**), séléniure de zinc (**ZnSe**) ou fluorure de calcium (**CaF<sub>2</sub>**). En effet, les supports conventionnels en verre sont opaques aux radiations infrarouges. Il n'est possible que de travailler dans les hauts nombres d'onde.



*Figure 16 - Appareillage Spotlight 2 FTIR, le système est constitué d'un microscope Spotlight 300 de Perkin Elmer associé à un spectromètre infrarouge à transformée de Fourier (Spectrum One, Perkin Elmer)*

### **4.3. Spectroscopie de diffusion Raman**

#### **4.3.1. Historique**

Chandrasekhara Venkata Râman né le 7 novembre 1888, à Tiruchirapalli, est un physicien indien de renom qui découvrit en 1928, le phénomène de diffusion inélastique : la diffusion Raman. En effet, il fut montré qu'une molécule irradiée par une onde électromagnétique, émet un rayonnement qui contient des photons ayant la même énergie que le faisceau incident (diffusion Rayleigh). Ce rayonnement contient également un faible pourcentage de photons (1 photon sur 1 million, soit 0.0001%) ayant une longueur d'onde légèrement différente, soit avec une énergie légèrement supérieure (radiation anti-Stokes) à l'onde incidente, soit légèrement inférieure (radiation Stokes). En 1930, C.V Râman reçoit le prix Nobel de physique pour sa découverte sur la diffusion de la lumière. C'est à partir de cette période que la spectroscopie Raman devient une technique très populaire comme méthode non-destructive pour l'analyse chimique. Mais peu de temps après la seconde mondiale, cette méthode d'analyse révolutionnaire a été supplanté par une méthode devenu routinière et plus

facile à utiliser par le fait de la révolution électronique et le développement des capteurs infrarouge : la spectroscopie infrarouge. Dans les années 60, le développement des lasers est en plein essor, malgré cela la spectroscopie Raman restera confinée dans les laboratoires et souffrant d'un gros handicap lié au phénomène de fluorescence qui pollue fortement le signal Raman et nécessitant des opérateurs très qualifiés et des chambres noires.

Il faudra attendre 1986, pour voir arriver le premier couplage d'un interféromètre commercial infrarouge à transformée de Fourier et une source d'excitation proche de l'infrarouge qui ont permis d'obtenir des spectres Raman de qualité tout en éliminant le phénomène de fluorescence. En effet, une source monochromatique proche de l'infrarouge ( $\lambda=1\mu\text{m}$ ) réduit fortement ce phénomène de fluorescence et avoir une meilleure sensibilité, tout en permettant qu'augmenter la puissance d'excitation sans avoir de photo-destruction car cette source est beaucoup moins énergétique que les lasers visibles.

C'est grâce à ces innovations que le Raman a pu se développer et permettre de constituer une base donnée indiquant les fréquences de vibration des liaisons atomiques des molécules. Ces bases de données sont actuellement comparables à celle obtenues en spectroscopie infrarouge. Ceci permet de réduire considérablement le temps d'identification d'une molécule. Le Raman est devenu accessible à un nombre beaucoup plus grand de scientifiques.

#### **4.3.2. Théorie de la diffusion Raman**

Dans le cas de la diffusion Raman spontanée [62], qui est comme vue précédemment, un processus résultant de l'interaction lumière/matière, des phénomènes de diffusion se produisent. Lors de cette interaction, le champ électrique  $\vec{E}$  de l'onde électromagnétique incidente réagit avec le nuage électronique de la molécule considérée et induit un moment dipolaire oscillant. Cette oscillation provoque la diffusion de radiations qui sont pour la majorité des diffusions élastiques ayant la même fréquence que l'onde incidente (diffusion Rayleigh). Mais une portion infime de ces radiations possède une fréquence différente avec gain ou perte d'énergie, ce sont des diffusions inélastiques. Pour comprendre ce phénomène, deux approches sont nécessaires : la mécanique classique et la mécanique quantique.

##### **a) Considération théorique : approche classique**

Lorsque l'on applique un champ électrique  $\vec{E}$  sur une molécule, son nuage électronique se déforme et il se crée un moment dipolaire induit de vecteur  $\vec{P}$  [62]. Ce moment dipolaire induit s'exprime par :

$$\vec{P} = \epsilon_0 \alpha \vec{E} \quad (\text{éq. 10})$$

Où  $\epsilon_0$  est la permittivité du vide,  $\alpha$  est la polarisabilité de la liaison et  $\vec{E}$  le vecteur du champ électrique incident.

Si on définit un vecteur champ  $\vec{E}$  qui oscille à la fréquence  $V_0$  et qui s'exprime selon l'expression suivante :

$$\vec{E} = E_0 \cos(2\pi V_0 t) = E_0 \cos(\omega_0 t) \quad (\text{éq. 11})$$

Où  $E_0$  est l'amplitude du champ électrique,  $V_0$  est la fréquence de l'onde incidente,  $\omega_0$  est la pulsation de l'onde incidente et t correspond au temps.

Et que pour chaque liaison vibrant à fréquence donnée, le déplacement du barycentre par rapport à sa position d'équilibre s'exprime selon l'équation suivante :

$$dQ = Q_0 \cos(2\pi V_{\text{vib}} t)$$

Où  $Q_0$  est l'amplitude maximale ou le déplacement maximal autour de la position d'équilibre,  $V_{\text{vib}}$  est la fréquence du mode normal de vibration.

Et que l'on suppose que les déplacements sont de faible amplitude, en approximant le tenseur de polarisabilité comme étant égale à :

$$\alpha = \alpha_0 + \frac{\partial \alpha}{\partial Q} Q \quad (\text{éq. 12})$$

Où  $\alpha_0$  est l'amplitude de la polarisabilité au repos.

On obtient :

$$\vec{P} = \epsilon_0 \alpha \vec{E} = \epsilon_0 \alpha_0 E_0 \cos(2\pi V_0 t) + \frac{\partial \alpha}{\partial Q} \epsilon_0 Q_0 \cos(2\pi V_{\text{vib}} t) \times E_0 \cos(2\pi V_0 t) \quad (\text{éq. 13})$$

Soit (éq. 14) :

$$\cos(a+b) = \cos(a) \cos(b) - \sin(a) \sin(b)$$

$$\cos(a-b) = \cos(a)\cos(b) + \sin(a) * \sin(b)$$

$$\cos(a+b) + \cos(a-b) = 2 * \cos(a) * \cos(b)$$

$$\cos(a) * \cos(b) = \frac{1}{2} * \cos(a+b) + \cos(a-b)$$

On a donc

$$\vec{P} = \epsilon_0 \alpha \vec{E} = \epsilon_0 \alpha_0 E_0 \cos(2\pi V_0 t) + \frac{\partial \alpha}{\partial Q} \frac{\epsilon_0 Q_0 E_0}{2} (\cos(2\pi t(V_0 - V_{vib})) + \cos(2\pi t(V_0 + V_{vib}))) \quad (\text{éq. 15})$$

Pour qu'il y ait un phénomène Raman il faut donc que  $\frac{\partial \alpha}{\partial Q} \neq 0$ . Ainsi selon cette équation, parmi les photons diffusés, la plupart ont la même fréquence que le rayonnement excitateur ( $V_0$ ). Ce phénomène de diffusion sans perte d'énergie et ayant la même fréquence  $V_{vib}$  que la radiation incidente ( $V_{vib} = V_0$ ), s'appelle la diffusion Rayleigh (diffusion élastique). Le phénomène Raman étant très rare, il arrive occasionnellement que des photons incidents soient diffusés avec changement de fréquence donc perte ou gain d'énergie, c'est la diffusion Raman (diffusion inélastique). Cet effet Raman inclut deux types de diffusion: la diffusion Raman Stokes et Raman anti-Stokes. Lors de la diffusion Raman Stokes, la fréquence de la lumière diffusée ( $V_0 - V_{vib}$ ) est plus petite que celle de la lumière incidente (perte d'énergie). Par contre, lors de la diffusion anti-Stokes la fréquence de la diffusion ( $V_0 + V_{vib}$ ) est plus élevée que celle de la lumière incidente (gain d'énergie).

### **b) Considération théorique : approche semi-quantique**

L'approche semi-quantique reste l'approche la plus adaptée afin de décrire la diffusion Raman de façon plus complète car elle permet d'imager l'interaction, quel que soit l'état physique du système étudié (solide, liquide ou gaz). L'excitation d'un milieu se traduit par un changement d'état électronique (Figure 17). Ainsi sous l'influence du faisceau laser, les atomes que compose la matière (exemple : cellule lymphocytaire) sont portés à un état quantique excité « virtuel ». On parle d'état « virtuel » car contrairement à la fluorescence, l'énergie extérieure apportée par les photons incidents ne permet pas la transition de l'état fondamental  $E_0$  au premier état excité  $E_1$ . Les molécules (ADN, protéines, lipides...) absorbent donc une partie du rayonnement incident, puis retourne à un état plus stable correspondant soit au niveau de base soit à un niveau d'énergie vibrationnelle précis. L'énergie de vibration est alors quantifiée en « paquet d'énergie ». Ces transitions entre états vibrationnels (mécanique quantique), correspondent à différents types de vibrations moléculaires (mécanique classique). Ainsi si

l'énergie du rayonnement diffusé est la même que l'énergie du rayonnement incident, nous sommes dans le cas d'une diffusion Rayleigh. En ce qui concerne la diffusion Raman, le système va se désexciter vers un état d'énergie vibrationnel différent de celui de départ. Si l'état final après désexcitation est un niveau plus excité que l'état initial, nous sommes dans le cas de la diffusion Stokes. Si au contraire, l'état de départ est dans un état excité et l'état final après désexcitation est dans un état moins excité, nous sommes dans le cas de figure d'une diffusion anti-Stokes. La différence entre l'état initial et l'état final correspond à l'énergie d'un niveau de vibration des molécules.

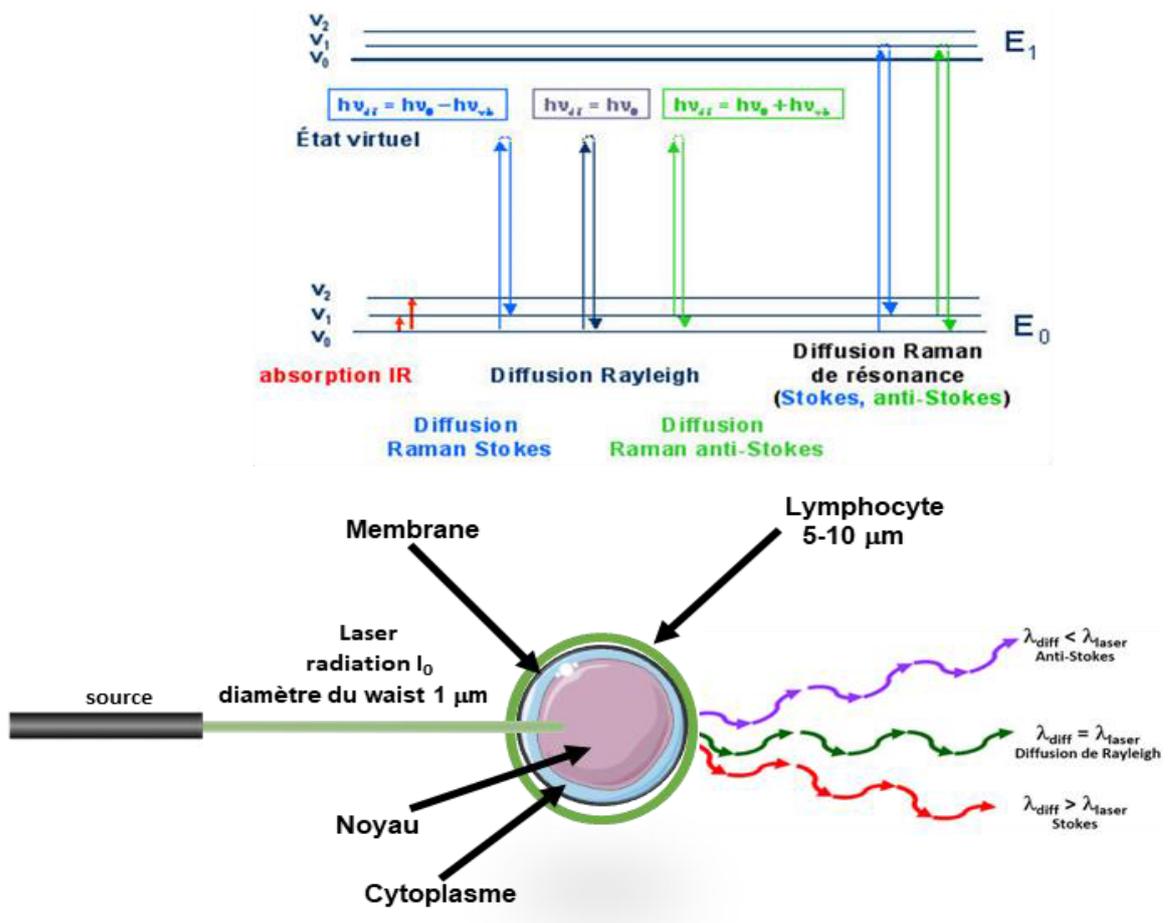


Figure 17 - Schéma montrant les principaux processus de diffusion Rayleigh, Raman Stokes et Raman anti-Stokes

Lors d'une analyse Raman, l'effet principalement utilisé est la diffusion Stokes car beaucoup plus intense à température ambiante que son homologue anti-Stokes comme on peut le voir sur la Figure 18. Lorsque la température évolue, le peuplement des niveaux d'énergie d'un ensemble de molécules obéit à la distribution de Maxwell-Boltzmann. Ainsi seules les

molécules se trouvant dans un état excité peuvent donner une transition anti-Stokes de type  $v \rightarrow v - 1$ . Ceci est régi par la relation :

$$\text{Soit } \frac{I_{as}}{I_s} = \left( \frac{V_0 + V}{V_0 - V} \right)^4 e^{\left( \frac{h\nu}{kT} \right)} \quad (\text{éq. 16})$$

Où  $k$  est la constante de Boltzmann ( $1.38 \times 10^{-23} \text{ J.K}^{-1}$ ),  $h$  la constante de Planck ( $6.62 \times 10^{-34} \text{ J.s}$ ),  $T$  est la température absolue,  $\nu_0$  est la fréquence de l'onde excitatrice et  $\nu$  est la fréquence du mode de vibration excité.

Il est possible grâce à ce rapport de déterminer la température d'un échantillon sans contact avec celui-ci. Ainsi dans une expérience donnée, si on augmente la température, on augmente la détection de la diffusion anti-Stokes à cause de la répartition des niveaux d'énergie.

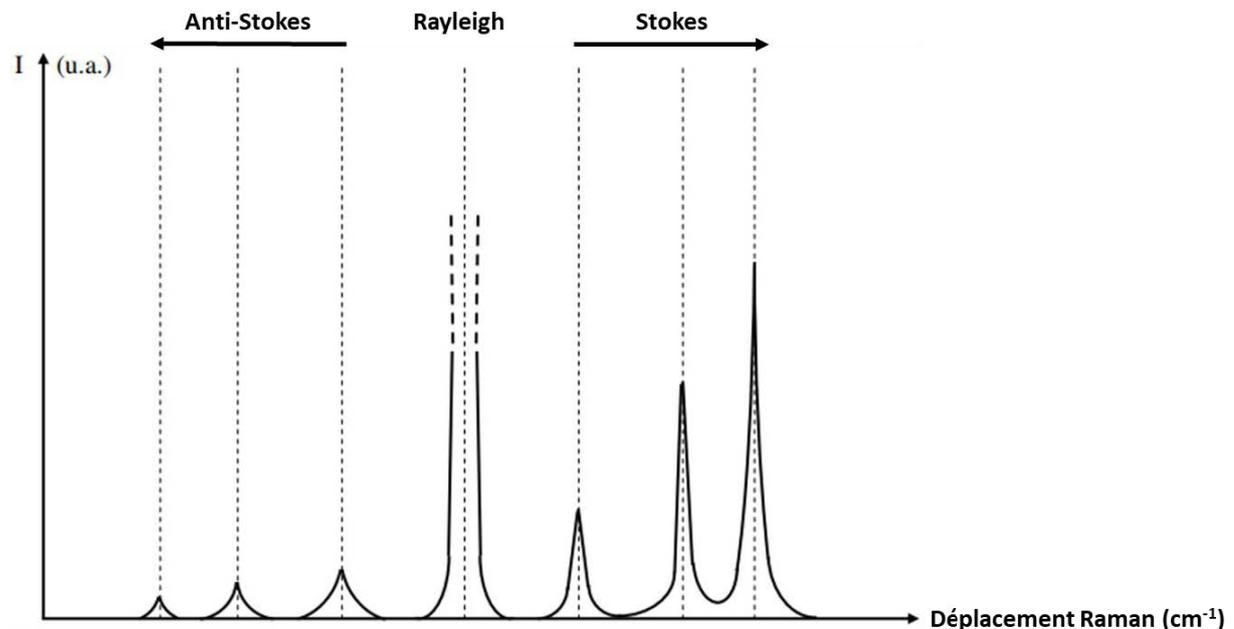


Figure 18 - Représentation d'un spectre de diffusion Raman à température ambiante

#### 4.4. Application des spectroscopies vibrationnelles en clinique

Au cours des dernières années, de nombreuses applications ont pu être rapportées dans le domaine biomédical, en oncologie, dermatologie, odontologie... [1-15, 17, 18, 63]. Ces nouvelles technologies ont été utilisées principalement dans l'optique de développer de nouveaux outils d'aide au diagnostic. La polyvalence des spectroscopies vibrationnelles permet d'étudier des processus biologiques fins tels que la détection de modifications génétiques et

épigénétiques induits par l'environnement ou par une pathologie [19, 20, 25]. Des études ont permis également d'évaluer les états de différenciations cellulaires [21-24] ou d'observer de façon dynamique l'évolution cellulaire, comme par exemple l'évolution au cours du temps de la molécule de cytochrome C dans le cadre de l'apoptose cellulaire [26].

Depuis quelques temps, les avantages des spectroscopies vibrationnelles émergent également dans le domaine de l'hématologie. Par exemple, en parasitologie hématologique, différentes publications font référence à la détection de la malaria en spectroscopie [64-68]. En immunologie, des équipes de chercheurs ont été capable de réaliser une séparation des sous-ensembles de cellules immunitaires : lymphocytes T, lymphocytes B, lymphocytes NK et cellules dendritiques [69, 70]. Ils ont pu montrer également qu'il était possible de faire la différence entre des cellules immunitaires activés et non-activés [69, 71, 72], soit par l'identification de bandes spectrales liées à l'évolution du taux d'immunoglobuline, soit par l'augmentation du taux de cytochrome C dans le cycle de respiration au niveau mitochondrial, montrant ainsi une augmentation de l'activité métabolique cellulaire. Cette dernière étude a montré que la spectroscopie Raman pouvait être utilisé afin de mettre en évidence des phénomènes temporels tel que la dynamique cellulaire. En effet, il a pu être observé au cours du temps (sur 24h) des changements biochimiques lents au niveau des lymphocytes après activation.

En ce qui concerne l'oncohématologie, des études ont également été menées afin d'identifier des marqueurs spectroscopiques caractérisant des cellules hématopoïétiques normales et néoplasiques sur des lignées cellulaires de type Jurkat et Raji [73] mais également sur des cellules du syndrome myélodysplasique et de leucémie myéloïde aigüe [74]. Une autre étude, a montré une identification de cellules tumorales, à des fins diagnostiques, dans le cadre de la leucémie lymphoblastique aigüe à cellule B [75]. Cette étude a été mené sur des lignées cellulaires RS4 ;11, REH et MN60. Il a pu être montré que le Raman était capable d'identifier le stade de maturation de ces cellules.

A l'heure actuelle, le déploiement des spectroscopies vibrationnelles en clinique est fortement limité par son incompatibilité avec les lames de verre qui sont utilisées comme consommable en milieu hospitalier. En effet, la préparation d'échantillon en hématologie pose encore problème car les techniques actuelles comme HemaPrep produisent des frottis sur lame de verre. Cependant, contrairement à la spectroscopie infrarouge qui n'est pas du tout compatible avec des supports en verre, il a été montré que l'utilisation d'un laser 532nm en

spectroscopie Raman, diminue la contribution du verre dans le signal [76], cependant celle-ci reste toujours problématique car interférant avec le signal moléculaire de l'échantillon. Dans la majorité des études relevées dans la littérature, les supports de CaF<sub>2</sub> ou quartz restent les alternatives les plus utilisées en spectroscopie vibrationnelle. Mais ces supports sont peu pratiques pour des applications à grande échelle en milieu hospitalier ou en routine car ce sont des substrats assez coûteux par rapport aux lames de verre. Dans ce contexte, la mise au point d'une nouvelle méthode reproductible et automatisable est nécessaire pour envisager une implémentation en clinique.

Dans le cas d'analyse Raman sur lame de verre, il est possible d'appliquer une correction numérique pour s'affranchir des interférences spectrales associées au verre [76]. En effet, la méthode « Extended Multiplicative Signal Correction » (EMSC), qui est une extension de la méthode « Multiplicative Signal Correction » (MSC), vise à minimiser l'influence de facteurs contaminants dans un signal [77]. Par rapport aux techniques conventionnelles de soustraction du signal parasite, l'EMSC semble plus efficace pour neutraliser la contribution du verre dans le signal Raman [78]. Ces autres techniques sont basées sur l'utilisation de polynômes pour modéliser le signal du verre; elles laissent une trace significative de la contribution du verre dans le signal. L'EMSC aurait également l'avantage de prétraiter les données en bloc (simultanément), en réalisant à la fois une normalisation, une élimination des interférences et de corriger l'effet de la fluorescence parasite. Ainsi par son efficacité et sa polyvalence, l'utilisation de cet algorithme semble produire des résultats fiables et reproductibles. L'EMSC apparaît donc être une méthode efficace pour une standardisation du prétraitement des spectres Raman en vue d'une exploitation diagnostique en clinique.

Ces publications contribuent à démontrer le potentiel des spectroscopies vibrationnelles dans des applications biomédicales. Les différentes méthodes utilisées ont permis de définir des marqueurs spectroscopiques, différenciant les tumeurs des tissus sains, les cellules pathologiques des cellules saines en traduisant des variations de structure ou de contenu en lipides, protéines ou acides nucléiques.

Dans ces études, l'un des principaux inconvénients est l'utilisation de modèles animal [69] ou (RS4 ;11, REH, MN60, Jurkat ou Raji) [72, 73, 75, 79, 80]. En effet, les lignées cellulaires ne reflètent pas spécifiquement la réalité, du fait qu'elles ne se trouvent pas dans leur environnement « naturel » et que par conséquent, leurs activités ne sont pas régulées par les autres cellules comme dans le cas d'un organisme complexe. Par contre, l'utilisation de lignées

cellulaires est un atout précieux pour le développement technologique, l'optimisation des protocoles de mesure et des traitements de données (optimisation des longueurs d'onde utilisées, puissance laser, temps d'acquisition, évaluation des prétraitements...). L'identification de biomarqueurs associés à la pathologie, transférables en clinique, est difficilement envisageable à partir des lignées cellulaires, du fait de l'existence potentielle d'une variabilité inter-lignées mais également d'une variabilité biochimique induite par le milieu de culture.

Une des caractéristiques de ce travail de thèse est de se rapprocher le plus possible d'un environnement clinique, pour une future implémentation en milieu hospitalier. Au vu de l'incompatibilité de la spectroscopie infrarouge avec les supports en verre, nous avons opté pour la spectroscopie Raman. Nous avons donc mis au point une standardisation des protocoles de préparation des frottis, des protocoles d'acquisition, de prétraitements, de création des bases d'apprentissage et d'analyse de données.

#### **4.5. La spectroscopie Raman pour l'aide au diagnostic**

La spectroscopie Raman est une technique biophotonique "label free" pour l'analyse d'échantillons biologiques complexes, tels que les biofluides [8, 16], les cellules [21-24] et les tissus [1-10, 12-15]. La spectroscopie Raman peut détecter des changements biochimiques causés par la maladie et induire des variations dans les spectres Raman. Cette technologie a donc le potentiel d'être utilisée pour le diagnostic dans la pratique clinique de routine. De ce fait, cette technologie a été largement employée dans de nombreuses études comme preuves de concept lors de campagnes de mesure ponctuelles pour l'analyse d'échantillons biomédicaux. En ce qui concerne l'oncohématologie, beaucoup d'études préliminaires ont été menées afin de trouver des différences moléculaires entre des cellules pathologiques et saines [69-75].

Le potentiel discriminatoire de la spectroscopie Raman est également renforcé par des outils mathématiques, des algorithmes d'apprentissage et de classification (supervisée ou non). Certains outils ont pour but de préparer les données Raman afin de les rendre comparables et analysables. D'autres outils servent à réduire la taille de l'ensemble des données, d'extraire seulement les informations pertinentes et de réduire le temps de calcul lors de création de modèles de prédiction. Le couplage de la spectroscopie Raman et de l'utilisation d'algorithmes permettant le traitement et la classification des données permet d'apporter un concept de standardisation et d'automatisation dans le processus de préparation et d'analyse des données Raman. Cependant, apporter cette technologie au chevet du patient nécessite de corriger

certaines manques de transférabilité des données Raman acquises dans différentes conditions expérimentales et instrumentales lors de campagnes de mesures multicentriques. En effet, ces changements expérimentaux et instrumentaux peuvent influencer sur les données Raman.

Dans le cadre de cette thèse, différents axes de recherche ont été menés. Il a été proposé, dans un premier temps, d'évaluer une solution consistant en l'application d'un prétraitement des données Raman spécifiquement développé afin d'éliminer la variabilité spectrale induite par les différents changements de conditions. Il a été montré que des modifications au niveau expérimental pouvait avoir une influence sur les spectres Raman [81, 82]. Nous avons comparé les données Raman de lymphocytes acquis au cours de deux campagnes de mesures indépendantes à partir de frottis sanguins frais non colorés, étalés sur lame de verre, et provenant de patients atteints de la LLC à un stade avancé ainsi que de personnes saines. Les différences entre ces campagnes étaient basées principalement sur l'optimisation de la configuration instrumentale, des acquisitions multicentriques, mais également du changement de la méthode de préparation des frottis (automatique et manuel). Dans le cadre d'une étude précédente, un prétraitement efficace avait été développé pour corriger les données Raman [83]. Or ce prétraitement s'est avéré obsolète pour des données multicentriques. Ainsi, une seconde correction numérique basée sur la correction multiplicative étendue du signal (EMSC) [77] a permis d'homogénéiser les données en neutralisant la variabilité du signal entre les deux campagnes de mesure.

Le second axe de recherche a été d'évaluer différentes stratégies de classifications sur les données Raman collectées dans différentes conditions cliniques, ceci afin de pouvoir créer et optimiser des modèles de prédiction pour réaliser le diagnostic de la LLC au niveau patient puis au niveau cellulaire. La LLC a été choisie comme sujet d'étude car il est possible d'avoir accès à une grande quantité de matériel tumoral. De plus de nombreuses technologies existent déjà pour caractériser les cellules leucémiques, comme l'imagerie morphologique, la cytométrie en flux ou la spectroscopie de masse. Toutefois, ces techniques peuvent être destructives, nécessitent des ressources importantes et coûteuses mais également une préparation spécifique de l'échantillon, ce qui empêche leur analyse par une autre modalité complémentaire.

## **5. Analyse des données**

Le processus d'analyse numérique se déroule suivant les étapes suivantes : 1) une phase de préparation des données (prétraitement), 2) une phase de réduction de données, 3) une phase d'apprentissage et d'évaluation des modèles (validation).

Afin d'avoir des modèles performants, il est primordial de procéder dans un premier temps, à une préparation ou à un nettoyage des données brutes (prétraitement) afin de les mettre en forme, de les rendre comparables et exploitables. Les données prétraitées peuvent subir ensuite une extraction de variables ou une sélection de variables afin de réduire l'information aux variables les plus pertinentes. Cette sélection de caractéristiques se fait par le biais d'algorithmes de corrélation statistique. Il est à savoir que les variables sélectionnées doivent être les plus décorréélées possible.

L'apprentissage automatique est une notion qui peut être assimilée au fait d'entraîner des algorithmes complexes afin de leur faire apprendre une action ou créer de la connaissance à partir de données existantes et ainsi accomplir des tâches sans être programmé. Cette science de l'information permet de créer des modèles à partir de données d'échantillonnage afin de prendre des décisions qui peuvent être appliquées sur de l'information inconnue. Il existe tout un panel d'algorithmes qui peuvent être utilisés selon la complexité et les caractéristiques des données à analyser : quantité de données, homogénéité ou dimensions. Dans cette partie, il sera donc question de voir les différentes méthodes les plus courantes de préparation de données et d'apprentissage supervisé et non supervisé (Figure 19).

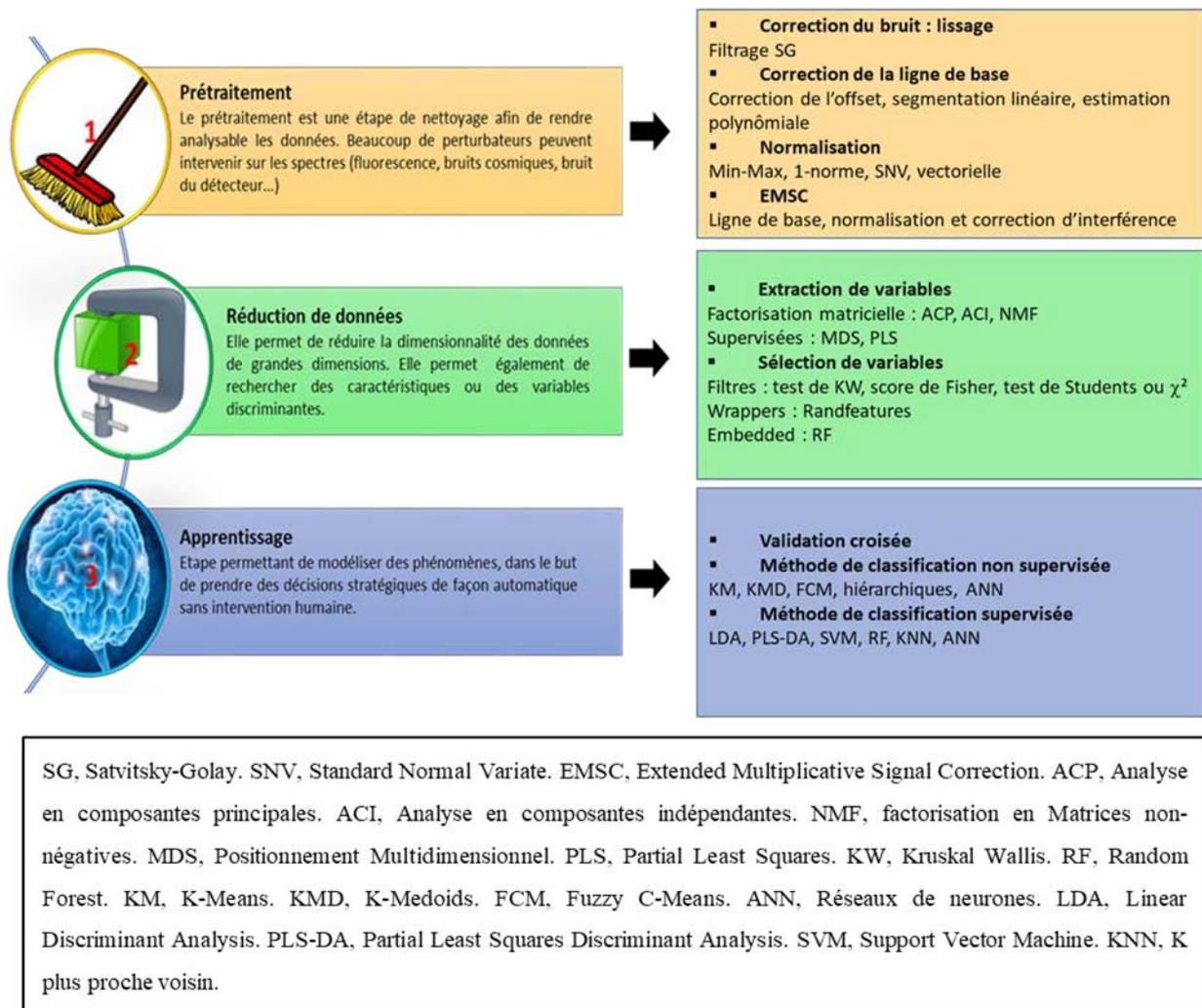


Figure 19 - Diagramme des principales étapes de traitement numérique pour des spectres Raman.

## 5.1. Prétraitement des données

Après l'acquisition Raman, la première étape consiste à prétraiter les données afin de les rendre analysables et fiables. En effet, divers phénomènes perturbateurs peuvent provoquer des interférences spectrales tels que la fluorescence de l'échantillon, le bruit de fond du détecteur, les bruits cosmiques, la défocalisation du laser ou la différence d'épaisseur de l'échantillon. Ces perturbateurs rendent quasi impossible l'analyse des données brutes [84]. Il est important de bien choisir son protocole de prétraitement afin d'éliminer correctement les interférences sans altérer l'information d'intérêt. Il existe différents types de prétraitements selon les perturbateurs à éliminer.

### **5.1.1. Lissage spectral**

Le lissage spectral est généralement le premier prétraitement qui est appliqué, il a pour but de réduire le bruit résiduel d'un spectre et ainsi augmenter son rapport signal/bruit. Selon l'application, il existe différents filtres de lissage, tel que les filtres passe-bas, les filtres binomiaux ou les filtres moyens [85]. Le plus couramment utilisé en spectroscopie Raman est une méthode introduite en 1964, le filtre de Savitzky-Golay [86]. Ce filtre est un bon compromis entre la réduction du bruit et la conservation de la morphologie spectrale (largeur de pic, intensité...). Cette méthode repose sur l'utilisation d'une fenêtre glissante de taille définie et une estimation polynomiale d'un certain degré, celui-ci appliqué localement le long de la fenêtre. Le point central de la fenêtre est remplacé par une approximation polynomiale. La taille de fenêtre et l'ordre du polynôme doivent être adaptés aux données et choisis avec soin. En effet, un choix mal adapté de ces paramètres peut supprimer de l'information essentielle pour discriminer les variables intéressantes.

### **5.1.2. Correction de la ligne de base**

La correction de la ligne de base est le second prétraitement appliqué sur les spectres afin de corriger les déviations d'intensité des spectres qui sont la conséquence de divers phénomènes endogènes et exogènes à l'échantillon ou à l'expérience. En spectroscopie Raman, la déviation spectrale est due principalement à la fluorescence provenant de l'échantillon, mais aussi des optiques ou lumières parasites non filtrées [87, 88]. Cette déviation est caractérisée par la variation lente d'une courbe dans le spectre ou un offset de celui-ci (déviation de la ligne de base). L'intensité de la ligne de base peut être très grande par rapport à l'intensité du signal rendant ainsi les spectres difficilement analysables. Il est donc nécessaire d'éliminer cette ligne de base numériquement grâce à des algorithmes spécifiques. La correction de l'offset est la méthode la plus simple, elle consiste à soustraire à un spectre une ligne de base droite et horizontale de sorte qu'au moins un de ses points soit égal à zéro. La correction par segmentation linéaire s'effectue en définissant un certain nombre de points sur le spectre, où chaque point est relié par une ligne droite puis cette ligne de base est ramenée à zéro. La méthode la plus utilisée en spectroscopie Raman est l'estimation par une fonction polynomiale de degré défini, c'est-à-dire que l'algorithme recherche un polynôme pour que celui-ci colle au mieux à chaque spectre. Lieber et al [89] ont proposé une méthode itérative qui estime la ligne de base par moindres-carrés en modifiant le signal à chaque itération jusqu'à ce que celui-ci

ressemble à la ligne de base. Il est à noter que dans la littérature, il existe bien d'autres méthodes [90-92].

### **5.1.3. Normalisation**

Le troisième prétraitement appliqué est la normalisation. Elle sert à homogénéiser tous les spectres afin de les ramener à une échelle commune. L'épaisseur hétérogène de l'échantillon, la variation d'intensité du laser ou les différences de focalisation de l'objectif sont des facteurs qui influent directement sur l'intensité du spectre [87]. Grâce à ce prétraitement, il est possible de s'affranchir de toutes ces variations et prioriser la forme des spectres sur son intensité relative. Il existe plusieurs types de normalisation qui peut être utilisée. La normalisation « Min-Max » est un procédé qui joue sur l'intensité maximale de chaque spectre. Il consiste par un procédé de transformation à ramener le minimum de chaque spectre à 0 et le maximum à 1. Les normalisations « 1-norm » et « vectorielle » reposent sur le principe de centrage/réduction des spectres, c'est-à-dire que dans un premier temps chaque spectre est centré sur la moyenne, puis chaque intensité est divisée soit par la somme des valeurs absolues de toutes les intensités pour « 1-norm », soit par la racine carrée de la somme du carré de toutes les intensités pour la normalisation « vectorielle ». La méthode « Standard Normal Variate » (SNV) consiste en une transformation par la déviation standard, c'est-à-dire que le spectre est centré sur la moyenne, puis celui-ci est réduit ou réajusté en divisant chacune de ces intensités par l'écart-type  $\sigma$  du spectre, ramenant ainsi sa moyenne à 0 et son écart-type à 1.

### **5.1.4. Extended Multiplicative Signal Correction (EMSC)**

Contrairement aux méthodes précédentes, l'Extended Multiplicative Signal Correction (EMSC) est une méthode corrective conjointe et « tout-en-un », dérivée de la méthode MSC développée par Martens et al [93, 94]. Cette méthode de prétraitement permet en même temps de corriger la ligne de base des spectres Raman grâce à l'application d'un polynôme de degré  $k$ , de neutraliser les effets de signaux contaminants dans les spectres comme le verre [78] ou la paraffine [95, 96] grâce à une matrice d'interférence, et de normaliser les spectres. Cette méthode est très efficace et très adaptative pour corriger l'hétérogénéité d'un jeu de donnée, car elle repose sur l'utilisation d'un spectre de référence qui permet de guider la correction des données (généralement la moyenne des données). Mathématiquement, elle considère les données comme une combinaison linéaire de signaux pondérés par des coefficients estimés par la méthode des moindres carrés. Le caractère polyvalent de l'EMSC en fait une méthode à fort

intérêt pour les spectroscopies vibrationnelles. Cette méthode est utilisée dans de nombreuses applications telles que des échantillons de colon humain [97], sur des cellules tumorales de poumons [98] ou dans le cas des biofluides [99].

## 5.2. Réduction de données

La spectroscopie Raman génère des spectres riches en information composés de plusieurs milliers de variables, complexifiant ainsi le traitement de données et l'extraction de l'information d'intérêt. Il est donc primordial après le prétraitement, de réduire la dimensionnalité des données afin de simplifier le problème et de proposer lors de la phase d'apprentissage des modèles optimaux et de réduire le temps de calcul. Pour cela, il existe deux types de méthodes permettant de réduire la dimensionnalité des données : l'extraction de variables et la sélection de variables.

### 5.2.1. Extraction de variables

La première catégorie a pour principe de réduire la dimensionnalité des données en les compressant par une transformation. **Les méthodes d'extraction de variables** combinent linéairement ou non linéairement les variables originales des données dans un espace différentiel réduit, appelé « feature space ». Ainsi à partir de  $p$  variables originales, un nouvel ensemble est créé contenant  $d < p$  variables. Ces méthodes permettent d'obtenir une vision globale plus réduite d'un ensemble d'observations multidimensionnelles. En spectroscopie Raman, plusieurs méthodes sont utilisées comme **les méthodes par factorisation matricielle (FM)** qui projettent les observations multivariées d'un espace de dimension  $d$  ( $d$  étant le nombre de variables observées) vers un espace plus petit en conservant un maximum d'information. Ainsi nous considérons tout objet comme étant la somme pondérée de plusieurs composantes sous-jacentes. Chaque objet est ensuite projeté sur ces composantes estimées afin d'obtenir de nouvelles coordonnées. En spectroscopie, différentes variantes sont utilisées telles que l'analyse en composantes principales (ACP) [100, 101], l'analyse en composantes indépendantes (ACI) [102], la factorisation en matrice non-négative (NMF) [103] et la transformation en ondelettes [104]. Dans la littérature, l'ACP est souvent combiné avec un algorithme de classification afin d'améliorer ces performances de prédictions. L'ACP a été utilisée dans différentes études afin d'extraire des caractéristiques diagnostiques. La combinaison de l'ACP et du Support Vector Machine (SVM) a été employée afin d'étudier les altérations osseuses dues à l'invasion tumorale par le cancer du sein [105]. Dans d'autres cas,

la combinaison de l'ACP et de l'analyse discriminante linéaire (LDA) a été utilisée: pour l'étude du cancer du col de l'utérus [106], pour la quantification de la fibrose rénale [107], afin d'analyser les effets radiobiologiques des rayonnements ionisants [108], ou pour détecter des hématies parasitées [109].

**Les méthodes supervisées (MS)** sont différentes des FM, car elles considèrent K étiquettes ou classes d'appartenances (labels) afin de construire une matrice de transformation. Elles cherchent une représentation de faible dimension des données afin de maximiser la variance interclasse et minimiser la variance intra-classe, cela permet une séparation optimale des K groupes dans les données. Les méthodes les plus courantes sont le Positionnement Multidimensionnel (MDS) [110] et moindres carrés partiels (PLS) [111].

### 5.2.2. Sélection de variables

Contrairement à l'extraction de variables, ces méthodes consistent en une recherche d'un sous-ensemble de variables explicatives parmi les variables d'origine, sans passer par une transformation. La réduction de donnée se fait par le biais de la sélection des variables les plus pertinentes pour répondre un phénomène spécifique. Une sélection manuelle peut être utilisée quand l'utilisateur connaît déjà approximativement ou avec précision les variables d'intérêt pour son sujet d'étude. Dans le cas de la spectroscopie, l'utilisateur réduira la taille de ces données en sélectionnant la gamme spectrale qui l'intéresse. En Raman, la gamme de  $400\text{ cm}^{-1}$  à  $1800\text{ cm}^{-1}$  (« Fingerprint ») est utilisée couramment. Pour la majorité des études, il est quasiment impossible de prévoir à l'avance les ensembles de longueurs d'ondes qui caractérisent au mieux l'empreinte moléculaire de l'échantillon étudié. C'est pour cela qu'il est primordial d'utiliser des techniques statistiques de sélection de variables. Trois familles d'approches de sélection de variables existent : **les méthodes filtres**, **les méthodes wrappers** et **les méthodes embedded**.

La première famille repose sur des tests univariés et permet de sélectionner des variables sans passer par un processus de classification. Elle se base seulement sur les caractéristiques intrinsèques de l'ensemble de données en leur attribuant un score. Seules les variables ayant un score élevé sont gardées pour la suite de l'étude. Il existe différentes méthodes comme le test de Kruskal Wallis [112], le score de Fisher [113], le test de Student ou le test du  $\chi^2$ .

La seconde famille propose des techniques multivariées, associées à un algorithme d'apprentissage afin de sélectionner un sous-ensemble de variables. L'avantage de ces

techniques est leur caractère multivarié. L'inconvénient est que ces méthodes sont très coûteuses en temps de calcul, longues et parfois complexes dans la mise en œuvre car dépendant de nombreux paramètres. Un autre inconvénient est la dépendance des variables sélectionnées au classifieur choisi. En effet, chaque algorithme de classification a ses propres critères et spécificités et si l'algorithme de classification est changé, rien ne dit que les variables sélectionnées seront les mêmes. De plus, ces méthodes peuvent être sujettes au sur-apprentissage. Pour éviter cela ces méthodes sont généralement couplées avec une validation croisée. Les méthodes *randfeatures* et les algorithmes génétiques font partie de ces méthodes [114-116]

La troisième famille est constituée par des méthodes « intégrées » permettant de sélectionner directement un sous-ensemble de variables discriminantes lors du processus d'apprentissage. Le sous-ensemble sera choisi de sorte qu'il optimise le critère d'apprentissage de l'algorithme de classification. Le *random forest* [117] fait partie de ces méthodes *embedded*. Ce sont des méthodes beaucoup plus rapides et moins complexes que les méthodes *wrappers*.

Toutes ces méthodes de réduction de données sont des outils indispensables dans le domaine des spectroscopies vibrationnelles afin d'améliorer les performances d'apprentissage et de prédiction des classifieurs.

### **5.3. Algorithmes de classification automatiques**

Les algorithmes de classification automatique sont des méthodes statistiques qui permettent de diviser un set de données en plusieurs groupes appelés « cluster ». Chaque groupe contient un sous-ensemble des données originelles ayant de grandes similitudes mais qui sont très différentes des données des autres groupes. Ces outils statistiques sont donc utilisés pour faciliter l'analyse de données de grandes dimensions. La taille du jeu de données et le critère d'homogénéité sont autant d'éléments à considérer afin de choisir l'algorithme de classification le mieux adapté. Il existe deux familles de méthodes de classification: non-supervisée et supervisée.

#### **5.3.1. La classification non supervisée**

La classification non supervisée (ou *clustering*) [118] est une méthode d'apprentissage qui est capable à partir de données non labellisées et de paramètres au préalable définis par l'utilisateur (nombre de clusters, métrique de calcul, paramètre de flou...), de trouver sans a priori des correspondances au sein de données d'entrée et de les regrouper en catégorie (cluster)

de manière à ce que les données d'une même catégorie soient les plus similaires possible et les plus éloignés de ceux des autres catégories. Il existe deux types d'algorithme de classification non supervisée : **hiérarchique et par partitionnement**.

- **Les classifications hiérarchiques** sont des méthodes d'agrégations itératives basées sur une représentation d'arbre hiérarchique ascendant (HCA) [119] ou descendant (HCD) [120], appelé « dendrogramme ». Ce dendrogramme peut être coupé à différents niveaux (Figure 20) pour obtenir le nombre de classes désirées. Il est construit en calculant les distances deux à deux de l'ensemble des données. Pour la méthode ascendante, au départ un seul objet est considéré par classe, puis au fur et à mesure, les classes qui plus proches sont fusionnées jusqu'à obtenir une classe unique contenant toutes les données. La méthode descendante consiste à partir d'une seule classe contenant toutes les données, de diviser successivement celle-ci au fur et à mesure des itérations. A la fin du processus, il y aura autant de classes que de données.

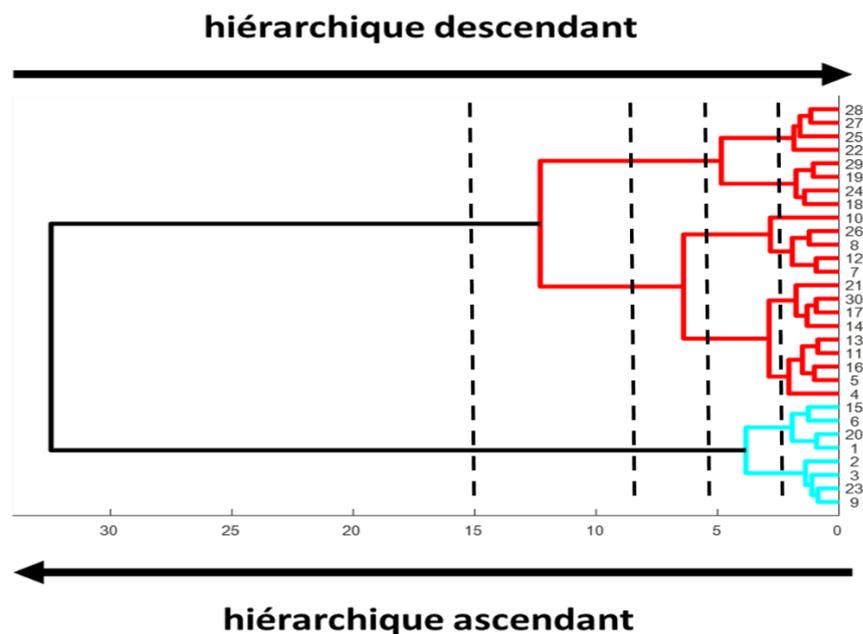


Figure 20 - Exemple de dendrogramme de clustering hiérarchique

- **Les classifications par partitionnement** sont des méthodes dits « centres mobiles » qui permettent de rassembler des données qui se ressemblent dans une même classe. Ces données sont rattachées à des « centroïdes » (données représentatives de chaque classe). Ces rattachements sont mis à jour au cours d'un processus itératif jusqu'à convergence de l'algorithme. L'optimisation de ces méthodes se fait par le biais d'une fonction objective.

Lorsque cette fonction objective converge, le processus s'arrête. Les algorithmes de tels que les K-Means [121] ou les Fuzzy C-Means [122] sont les plus couramment utilisés en spectroscopie vibrationnelle [123-125].

### **5.3.2. La classification supervisée**

Les méthodes de classifications supervisées ont pour but de construire un modèle de prédiction à partir de données accompagnées de leurs labels, afin que le classifieur puisse apprendre les règles de décisions qui serviront à distinguer les différentes classes de données. Ainsi le principe du classifieur est de modéliser à partir des données et de leurs labels une représentation qui permet de distinguer au mieux les différents groupes. Du point de vue décisionnel, on veut avoir la meilleure affectation possible pour classer de nouvelles données dont on ne connaît pas les labels. Par exemple, dans ce mémoire, l'algorithme supervisé est alimenté avec des données Raman de cellules cancéreuses labellisées comme LLC et d'autres comme cellules saines. Étant modélisé sur ces données, l'algorithme d'apprentissage supervisé devrait être capable d'identifier plus tard des signaux Raman de cellules LLC ou saines non labellisées et ainsi faire la différence entre un patient sain et LLC. Il existe différents types d'algorithmes: i) linéaires : Partial Least Squares - Discriminant Analysis (PLS-DA) [126], ii) non linéaires : Support Vector Machine (SVM) [127, 128] et Random Forest (RF) [127] dont les bonnes performances en spectroscopie vibrationnelle ont largement été prouvées dans la littérature. PLS-DA [126] est une méthode linéaire utilisée pour modéliser la structure de covariance entre les données et les labels par des variables latentes. Cette méthode est particulièrement efficace pour les données multi colinéaires comportant plus de caractéristiques que d'observations, ce qui est généralement le cas dans les études de spectroscopie Raman. SVM [129] est une méthode non linéaire qui utilise une fonction noyau (kernel) et un hyperplan afin de séparer les classes en maximisant une marge permettant de séparer au mieux les différents groupes. Afin de pouvoir séparer des données non linéairement séparables, SVM projette implicitement ces données non linéaires, dans un espace à haute dimension dans lequel ces données seront linéairement séparables. RF [117] est une méthode non linéaire basée sur la construction d'une multitude d'arbres de décision. Chaque arbre, avec une profondeur prédéfinie, est construit en utilisant un échantillon aléatoire des données et un nombre prédéfini de variables sélectionnées au hasard. La décision finale de RF est basée sur le vote à la majorité de tous les arbres de décision.

Il existe beaucoup d'autres algorithmes ayant tous leurs propres spécificités telles que LDA, les K plus proches voisins (KNN), les réseaux de neurones artificiels (ANN), les algorithmes génétiques.

# **CHAPITRE 2 : MATERIELS ET METHODES**

Ce chapitre aborde les aspects pratiques et expérimentaux qui sont spécifiques à l'élaboration de frottis en conditions cliniques mais également de leurs analyses. Ainsi ici sont détaillés les caractéristiques, les protocoles et les algorithmes de traitement et de classification des spectres Raman utilisés lors de cette étude.

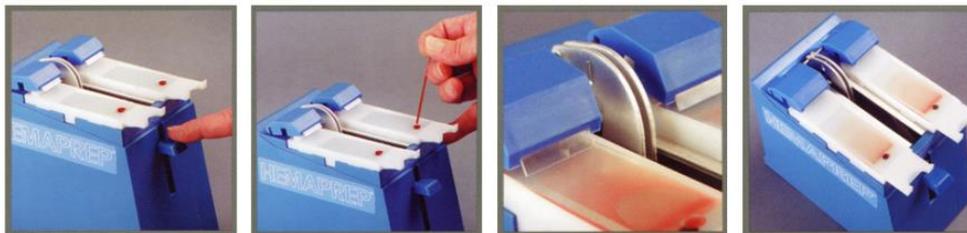
## 1. Préparation des échantillons sanguins

### 1.1. Groupes d'étude

Dans le cadre de cette étude, un groupe de 61 patients sains et un groupe de 79 patients atteints de LLC, avec un score de Matutes supérieur à 3 et un stade A dans la classification Binet, ont été formés à partir de deux campagnes de mesure différentes. La première a été réalisée en 2010-2011 dans le cadre du projet ANR TecSan IHMO et a concerné 25 patients sains et 45 patients B-LLC recrutés au Centre Hospitalier de Reims Champagne-Ardenne (RCA-HC), comme décrit précédemment dans l'article [83]. La deuxième campagne, réalisée au cours des années 2015-2016 dans le cadre du projet européen CIP ICT PSP M3S, a inclus 36 patients en bonne santé et 34 patients B-LLC recrutés au centre hospitalier Mont Godinne-Namur (MGN-HC). Les deux campagnes ont été réalisées avec le consentement des patients et approuvées par les comités d'éthique locaux.

### 1.2. Frottis sanguin

Pour chaque patient, un frottis a été préparé directement à l'hôpital de recrutement, en étalant une goutte de sang déposée sur une lame de verre conventionnelle afin de travailler dans des conditions cliniques classiques. Chez RCA-HC, un étalement manuel a été utilisé, tandis qu'un étalement automatique (Figure 21) (hémaprep) a été utilisé chez MGN-HC, les deux sans traitement chimique préalable.



*Figure 21 - Étaleur de frottis automatisés Hemaprep*

### **1.3. Cytométrie en flux (MFC)**

Le diagnostic de la LLC nécessite l'utilisation de la cytométrie en flux multiparamétrique (MFC) afin de caractériser les lymphocytes B [34]. En effet, il est difficile de distinguer avec la microscopie conventionnelle une cellule saine d'une cellule tumorale et il est très difficile de distinguer les sous-types lymphocytaires (les lymphocytes B, T et NK). Le pronostic repose sur le stade de la maladie déterminé par la classification de Binet (données cliniques et numération) et plusieurs biomarqueurs moléculaires validés tels que les marqueurs membranaires. La détermination de ces marqueurs nécessite l'utilisation de différentes méthodes de haute technologie telles que les analyses MFC.

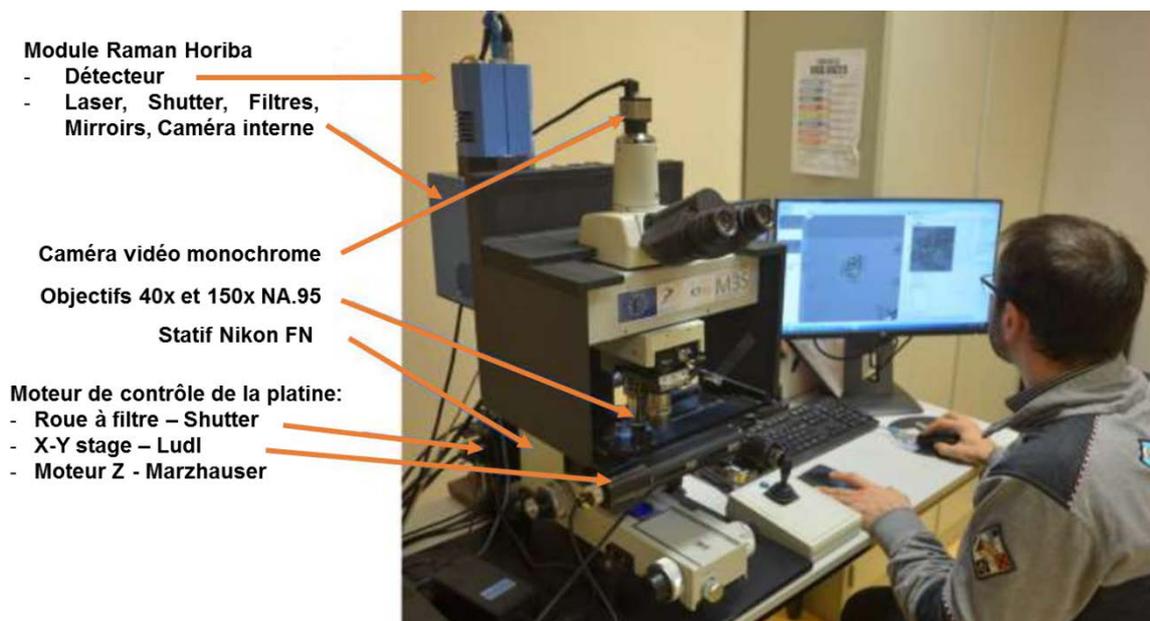
Dans le cadre du projet, les proportions de sous-population de lymphocytes (T, B et NK) ont été déterminées à l'aide de l'analyse MFC. Des échantillons de sang ont été prélevés dans des tubes K2 EDTA. La MFC a été réalisée sur un appareil FACS Canto-II (Becton Dickinson). Pour l'analyse, un cocktail de marqueurs a été utilisé : CD19, CD20, CD5, kappa, lambda, CD10, CD3, CD4, CD8 et CD56. L'acquisition et l'analyse des données ont été réalisées à l'aide du logiciel DiVa (Becton Dickinson). Pour chaque échantillon, nous avons déterminé la proportion de lymphocytes B, T et NK. Les mêmes échantillons ont été analysés sur la plateforme M3S.

### **1.4. Vérité terrain**

Dans le cadre de cette thèse, deux vérités terrain ont été utilisées. Dans une première étude, afin de travailler à l'échelle du patient, nous avons utilisé le statut du patient (LLC ou sains). Puis dans un second temps, afin de travailler au niveau cellulaire, nous avons utilisé un immunomarquage a posteriori des acquisitions Raman. Afin de réaliser cette étape d'immunomarquage, nous avons utilisé des anticorps porteurs d'une enzyme qui produit une coloration rouge en présence de révélateurs, les anticorps sont dirigés contre le complexe CD20, spécifique des lymphocytes B, et le complexe CD2, spécifique des lymphocytes T et NK. Les anticorps ont une spécificité de l'ordre de 100% et une sensibilité de l'ordre de 90% ce qui implique un risque de faux négatif dans les résultats obtenus. Afin de réduire ce risque la population minoritaire des lymphocytes est systématiquement marquée: B (CD20) chez les patients sains, T et NK (CD2) chez les patients LLC. La procédure d'immunomarquage des frottis est réalisée après une fixation par un mélange méthanol/acétone (50%/50%). Le protocole employé est présenté en Annexe.

## 2. Instrumentation et workflow des acquisitions

Le système M3S (Figure 22) génère différents types de données qui comprennent les spectres Raman, l'imagerie d'intensité et les images QPI (Phase Quantitative). Deux campagnes de mesures ont été faites d'abord au sein de l'URCA puis ensuite dans les centres cliniques. Au cours du processus d'analyse des frottis, le système M3S utilise différentes méthodes d'acquisition. Les données morphologiques et Raman sont acquises sur des lames non marquées, ces lames subissent ensuite l'immunomarquage qui permet d'obtenir une vérité terrain pour chacune des cellules observées. Afin de labelliser chaque cellule, les images en couleur correspondantes ont été collectées et labellisées par les cliniciens. Les acquisitions ont nécessité un calendrier précis afin d'assurer une immunocoloration efficace des frottis après les analyses Raman.



*Figure 22 - Plateforme d'acquisition de spectres Raman M3S - Le système est composé d'un microscope conventionnel, d'un module Raman, d'un laser 532nm, d'un éclairage par LED et d'une platine motorisée XYZ.*

### 2.1. Dispositif Raman

La Figure 23 présente le schéma du chemin optique typique d'un appareillage Raman permettant l'acquisition de spectres. En règle générale, une lumière monochromatique est générée par un laser, ainsi la lumière incidente est focalisée par l'objectif du microscope sur l'échantillon. Les rayonnements de diffusion Raman issue de l'interaction lumière/matière

suivent le même trajet que la lumière incidente au niveau de l'objectif, c'est une configuration de rétrodiffusion. La lumière diffusée est ensuite envoyée vers un filtre de réjection, filtre notch ou edge, afin d'éliminer la diffusion Rayleigh et la rétrodiffusion du laser. Le signal filtré est dirigé vers le système de dispersion (le spectromètre) et enregistré à l'aide d'une caméra CCD refroidie. Signalons que la spectrométrie Raman a l'avantage de permettre l'étude d'échantillon très diversifié (poudres, liquides, cellules, tissus . . .).

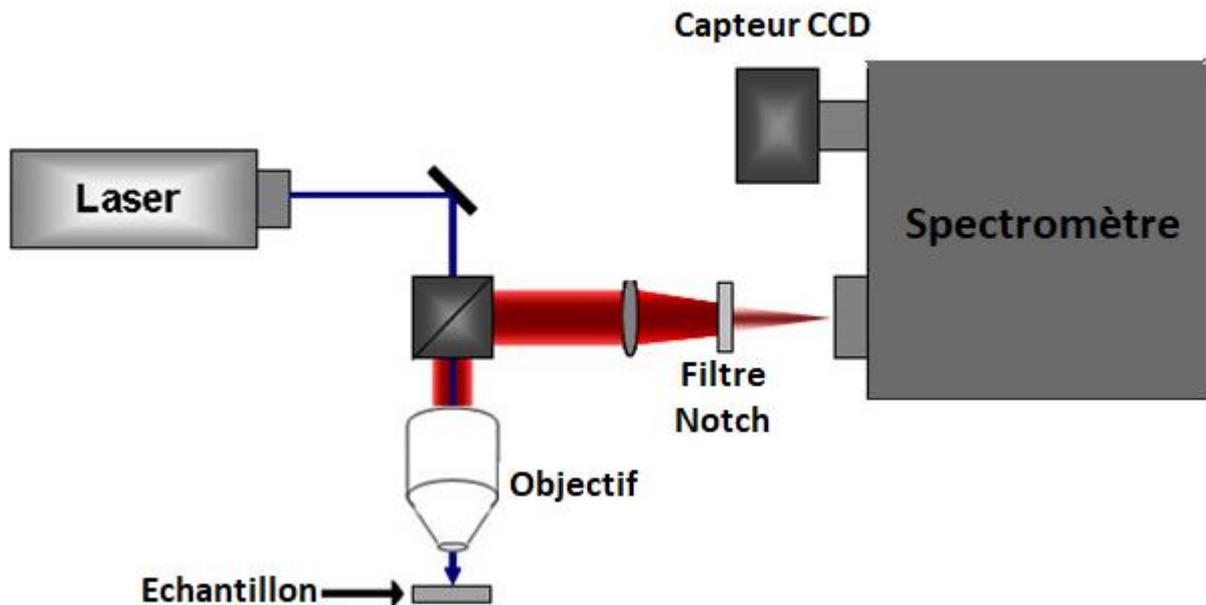


Figure 23 - Schéma optique d'un spectromètre Raman

#### - La source laser

Le laser fournit une radiation monochromatique de fréquence et d'énergie stable. Le choix de la longueur d'onde du laser est fait selon la fragilité et la nature de l'échantillon à analyser (peau, cellules, sérum, matériaux...) et le but de l'analyse selon s'il faut favoriser les hautes ou les basses fréquences. Notre système Raman est équipé d'un laser fournissant une radiation dans le visible d'une longueur d'onde verte de 532 nm (Toptica Photonics, Munich, Germany). Dans le cadre de ce projet, nous avons choisi cette longueur d'onde, car elle permet d'avoir une plus faible contribution du verre dans le signal Raman par rapport aux autres longueurs d'onde.

#### - Le microscope

Le système Raman est couplé à un statif Nikon FN1 équipé de deux lentilles à sec (Nikon). Le microscope a pour fonction de focaliser le spot laser incident sur la surface de

l'échantillon, via un objectif à fort grossissement (100x/NA 0,9 dans le projet IHMO et 150x/NA 0,95 dans le projet M3S), puis dirige les photons diffusés vers le spectromètre. Les globules blancs sont automatiquement localisés sur le frottis par analyse d'image grâce au logiciel embarqué développé par la société Tribvn. La localisation se fait avec un objectif 40x à sec (NA 0.6, Nikon SA, Champigny Sur Marne, France). Le microscope est équipé d'une platine XYZ motorisée (Ludl Electronic Products Ltd, New-York, USA).

#### - **Le filtre de réjection**

Afin de séparer le signal d'intérêt (diffusion Raman Stokes) du signal Rayleigh beaucoup plus intense, un filtre de réjection holographique de type « Notch » (filtre passe-haut) est placé sur le trajet optique du faisceau diffusé. Ainsi la diffusion Rayleigh est fortement atténuée mais la totalité de la diffusion Raman est conservée.

#### - **Le monochromateur**

Les photons Raman diffusés entrent par une fente dans le monochromateur qui va permettre de localiser spatialement la lumière. Un des éléments principaux d'un monochromateur est le réseau holographique (réseau de diffraction) qui fonctionne comme un prisme. Ce réseau est dit « blazé », c'est-à-dire qu'il est gravé de plusieurs traits parallèles les uns les autres. Le nombre de traits par millimètre conditionne la résolution spectrale. La résolution spectrale ou pouvoir de dispersion du réseau, correspond au pouvoir de séparation de deux longueurs d'onde différente. Ce pouvoir de séparation est lié directement à la densité de trait sur le réseau holographique et est régi par la loi de Bragg [130]. Ainsi dans notre étude, les photons sont dirigés vers un réseau holographique de 1200 traits/mm permettant leur dispersion selon leur énergie (Figure 24). Le choix du nombre de traits du réseau holographique se fait en fonction de la longueur d'onde incidente, de la résolution spectrale désirée et de la gamme spectrale à couvrir. Le faisceau est dispersé ensuite sur le détecteur CCD. Toutes les données spectrales ont été recueillies sur la plage spectrale de 700 à 3170  $\text{cm}^{-1}$  avec une résolution spectrale de 4  $\text{cm}^{-1}$ . Il est important de savoir que plus la résolution augmente plus la fenêtre spectrale est réduite et plus l'intensité du signal diminue.

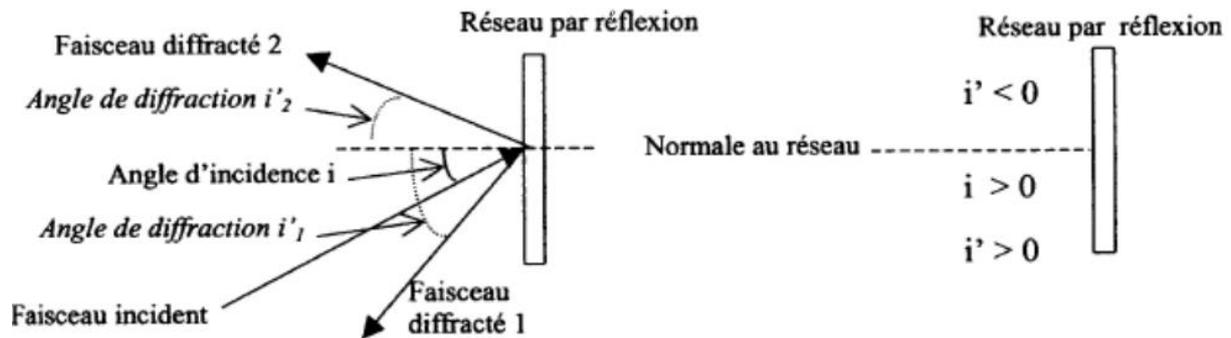


Figure 24 - Principe de fonctionnement d'un réseau holographique

- **Détecteur**

Dans notre cas, il s'agit d'un détecteur CCD (Charge Coupled Device) de 1024 x 256 pixels (Andor Technology, South Windsor, CT, USA), permettant de recueillir les informations spectrales concernant l'échantillon analysé. Le détecteur est refroidi à  $-70^{\circ}\text{C}$  par effet Pelletier.

- **Résolution, profondeur et pénétration**

**La résolution latérale (X et Y)**, est la taille minimale du spot laser sur l'échantillon. Elle est d'autant meilleure que la longueur d'onde du laser est petite et que l'ouverture numérique (NA) est grande. En effet, ceci découle de la spécificité ondulatoire d'une onde électromagnétique qui produit une fois focalisée sous l'effet de l'interaction avec un objet, une tâche de diffraction de diamètre RL (Figure 25).

Elle est définie par la relation :

$$RL = \frac{1,22\lambda}{2NA} = \frac{0,61\lambda}{NA} \text{ (éq. 17)}$$

Soit  $\lambda$  : la longueur d'onde de radiation incidente et NA : l'ouverture numérique des objectifs où  $NA = n \cdot \sin\alpha$  où n est l'indice du milieu [131]. Dans le cadre du projet, la résolution latérale est d'environ  $1 \mu\text{m}$ .

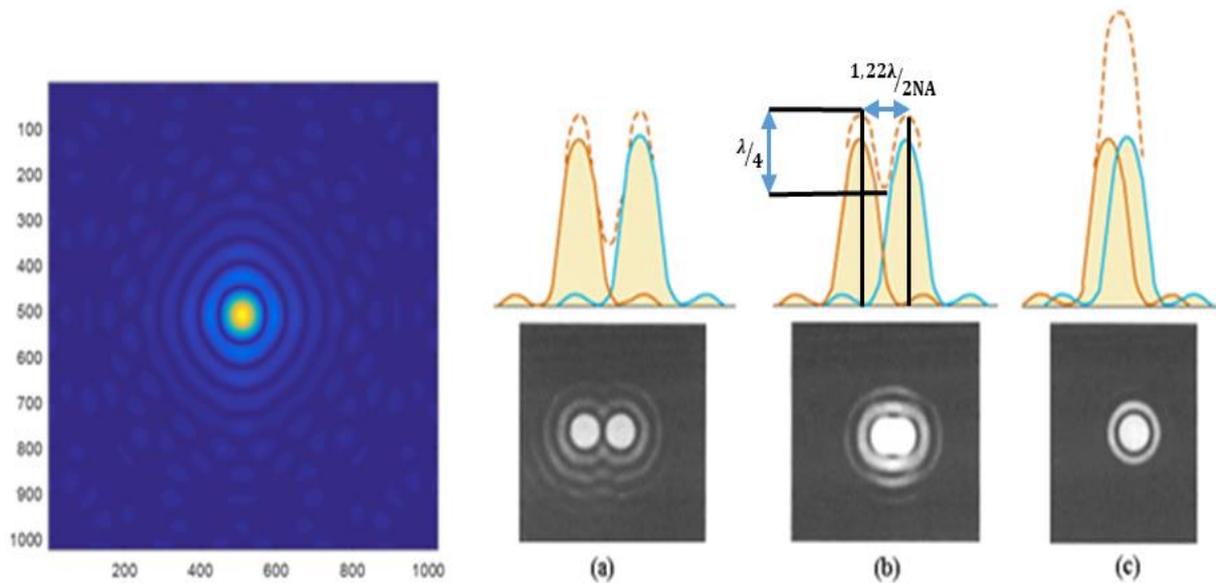


Figure 25 - Tâche de diffraction de la lumière [132]

**La résolution axiale (Z)**, correspond au volume optique du laser dans l'échantillon. Elle est directement proportionnelle à l'indice de réfraction de la lumière dans un échantillon donné. Sur le système utilisé dans cette étude, nous avons une résolution axiale d'environ  $2 \mu\text{m}$  et s'exprime selon l'équation 18.

$$RZ = \frac{1,4\lambda n}{NA^2} \text{ (éq. 18)}$$

## 2.2. Acquisition des spectres Raman

Pour la campagne IHMO, un spectre Raman a été acquis avec un objectif 100x sur le centre du noyau de 2596 cellules saines et 4257 cellules LLC avec un temps d'acquisition fixé à deux accumulations de 10 secondes, correspondant à 6853 spectres Raman acquis, soit un total d'environ 100 cellules par patients.

Dans le projet M3S, un nombre de trois spectres a été recueilli sur chaque cellule avec un objectif 150x, avec une distance d'environ  $1,5 \mu\text{m}$  entre les points d'acquisition. Le laser illumine une surface d'environ  $1 \mu\text{m}$  de diamètre qui permet d'acquérir trois spectres Raman sur le noyau des globules blancs avec un temps d'intégration de 10 secondes et 1 accumulation. Tout ceci correspond à 1804 cellules saines et 3128 cellules LLC, pour un total de 14796 spectres Raman acquis. Chaque cellule est ainsi caractérisée par ses spectres Raman et son label

physiopathologique (sain ou LLC). Un total d'environ 100 cellules par patients. La Figure 26 montre l'interface d'acquisition du système M3S.

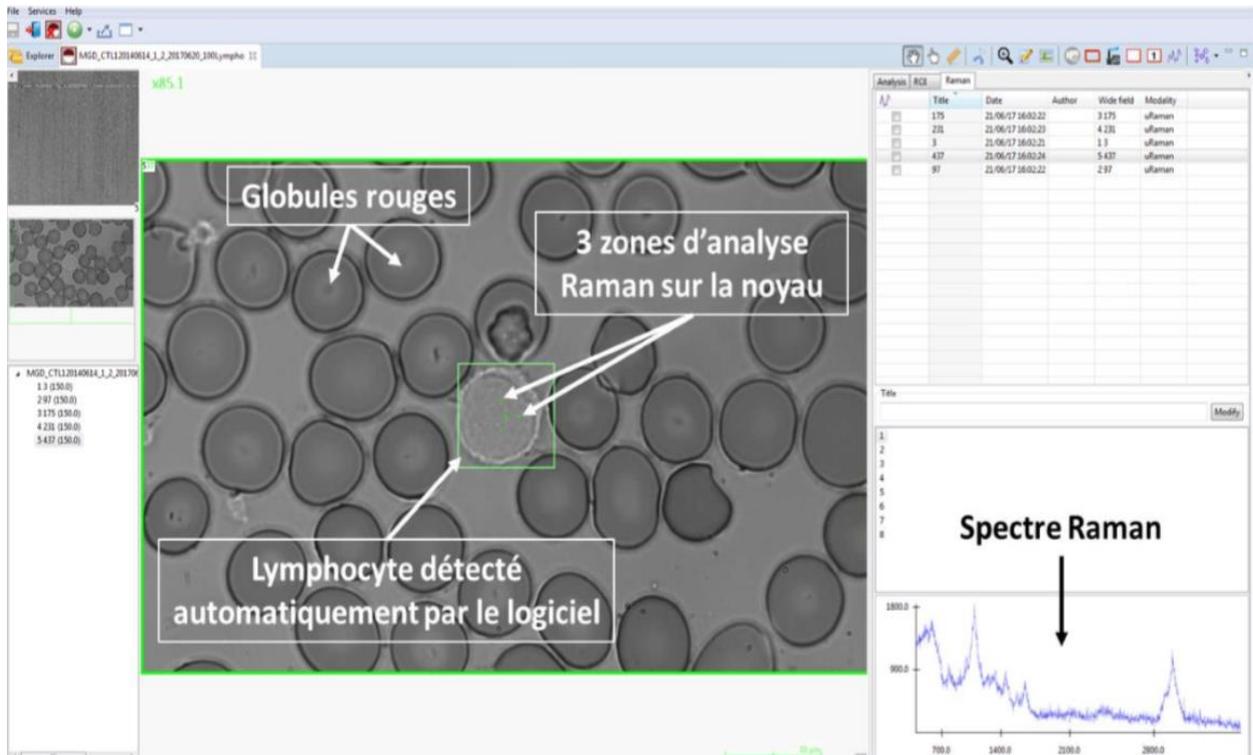


Figure 26 - Interface d'acquisition du système M3S

### 3. Tests qualités

Avant le prétraitement, il est important d'évaluer la qualité des spectres Raman bruts enregistrés. Les spectres Raman de mauvaise qualité doivent être rejetés. Seuls les spectres Raman de haute qualité doivent être conservés pour la construction de classificateurs robustes [133]. Les spectres bruts n'ayant subi aucun traitement au préalable, sont injectés dans un programme Matlab permettant de contrôler la qualité de ceux-ci. Quatre tests de qualité ont été conçus afin de quantifier la contribution des différents perturbateurs [83]. Un seuil a été défini pour chaque test de qualité. Ces seuils ont été fixés un à un manuellement afin d'avoir un bon compromis entre la qualité des spectres gardés et le nombre de spectres éjectés.

#### 3.1. Rapport Signal/Bruit

Le rapport Signal/Bruit (SNR) calcule le ratio de l'intensité intégrée entre une gamme spectrale connue pour être très informative et intense ( $2800$  à  $3150\text{ cm}^{-1}$ ) qui correspond aux protéines/lipides et l'écart-type d'autre gamme connue pour être une zone pauvre en signal

(1800 à 2200  $\text{cm}^{-1}$ ). Une valeur minimale de ce ratio a été fixée à 2, afin de ne conserver que les données les plus intenses. Les spectres avec un SNR inférieur à ce seuil sont supprimés de l'analyse (Figure 27). Pour les spectres Raman acquis sur les lymphocytes, le signal appartenant à la plage spectrale informative est spécifique aux protéines/lipides, tandis que la zone pauvre en signal est spécifique du bruit

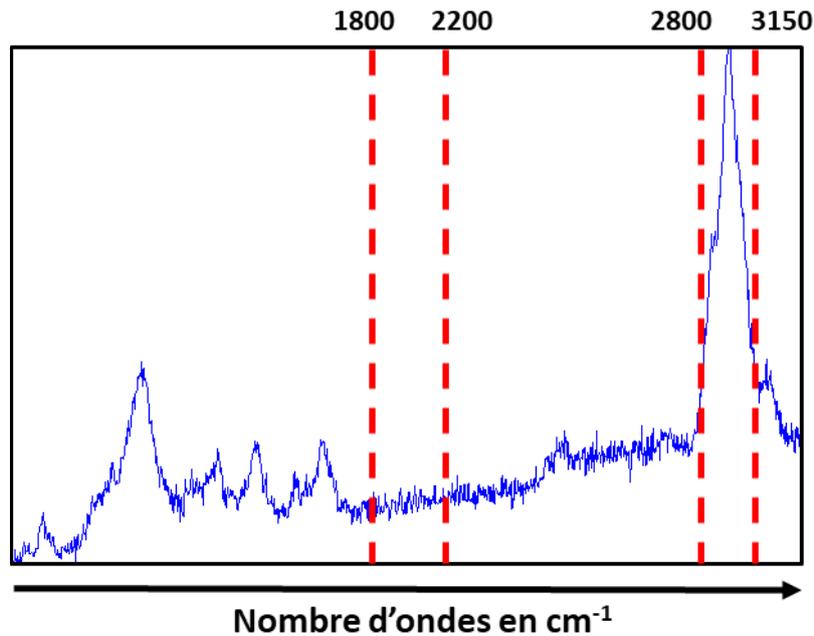


Figure 27 - Spectre Raman ayant un mauvais rapport signal sur bruit

### 3.2. Présence de signal de l'hémoglobine

Au niveau du frottis, les lymphocytes peuvent être proches ou superposés aux globules rouges. Les spectres Raman des lymphocytes enregistrés seront ainsi contaminés par une signature Raman des globules rouges (Figure 28). Ces spectres doivent donc être retirés de l'analyse. Afin d'évaluer ce niveau de contamination, le rapport entre l'intensité intégrée du signal dans la gamme 1656-1720  $\text{cm}^{-1}$  (Amide I) et l'intensité intégrée du signal dans la gamme 1540-1656  $\text{cm}^{-1}$  (hémoglobine) a été effectué. Les spectres avec un rapport inférieur à un seuil prédéfini sont supprimés de l'analyse. Un seuil de 1 a été déterminé comme étant optimal.

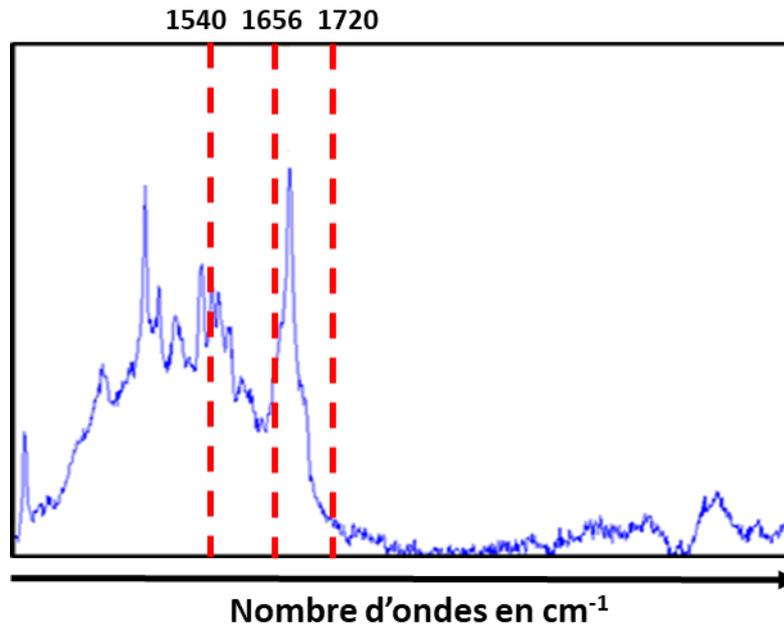


Figure 28 - Spectre Raman de la superposition d'un lymphocyte et d'un globule rouge

### 3.3. Détection des rayonnements cosmiques

En raison de l'utilisation de détecteurs CCD dans la configuration du système M3S, les spectres Raman acquis peuvent être parasités par des pics générés par les rayonnements cosmiques ou « Spikes » (Figure 29). Les spectres Raman contaminés par ces événements doivent être retirés.

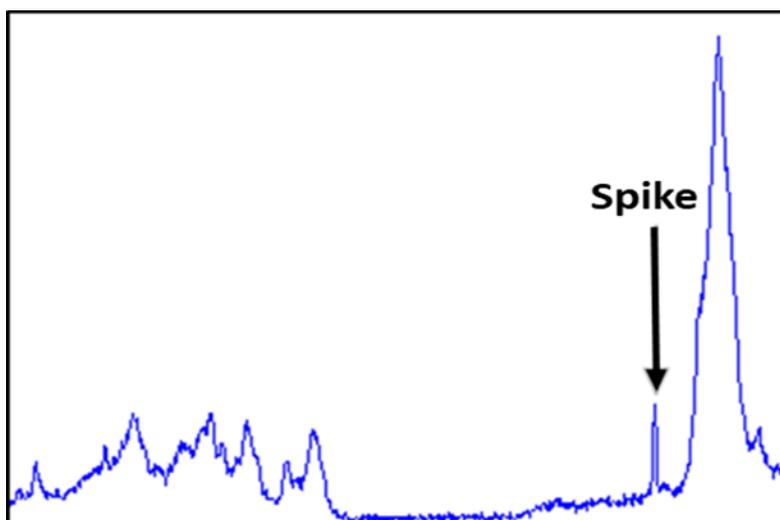


Figure 29 - Spectre de lymphocyte contenant un Spike (flèche noire)

### **3.4. Saturation des spectres**

En raison des dommages causés par l'irradiation sur certaines cellules ou de la fluorescence cellulaire, les spectres Raman peuvent être saturés en intensité. Ces spectres non informatifs doivent être supprimés de l'analyse. La saturation est caractérisée par plusieurs nombres d'ondes consécutifs ayant la même intensité maximale.

L'application de ces quatre tests « qualité » a abouti à l'élimination de 416 spectres (8%) du projet M3S et 822 spectres (12%) du projet IHMO.

## **4. Prétraitement des spectres Raman**

Afin de ne conserver que des informations cellulaires pertinentes, il est essentiel de corriger les spectres Raman des interférences spectrales telles que le bruit, le signal de verre, une dérive de la ligne de base principalement due à la fluorescence cellulaire.

Pour supprimer le bruit dans les spectres Raman, un lissage Savitzky-Golay [86] a été appliqué en utilisant une largeur de fenêtre de 9 points et une fonction polynomiale du troisième ordre. Ensuite, deux protocoles de prétraitement différents ont été appliqués dans cette étude.

### **4.1. Méthode de prétraitement « stepwise »**

La première utilisée dans le cadre du projet IHMO [83] est une méthode stepwise (étape par étape) appliquée individuellement à chaque spectre. La contribution du signal de verre a été estimée dans chaque spectre en ajustant un spectre Raman moyen du verre par la méthode des moindres carrés. Cette contribution a été soustraite du spectre. Ensuite, la ligne de base a été corrigée à l'aide d'une fonction polynomiale de second ordre. Enfin, le spectre a été normalisé selon la méthode « Standard Normal Variate » (SNV) qui consiste à retrancher à un spectre sa moyenne puis à diviser le résultat par son écart-type ramenant ainsi le spectre à une moyenne nulle et écart-type égal à 1 [134, 135].

### **4.2. Méthode de prétraitement « conjointe »**

La deuxième méthode consiste en l'application d'une méthode conjointe, basée sur la correction Extended Multiplicative Signal Correction (EMSC). Cette méthode utilise le spectre moyen de l'ensemble de données comme spectre de référence. La ligne de base et le signal du verre ont été neutralisés à l'aide d'une fonction polynomiale du quatrième ordre et d'un spectre Raman moyen du verre, respectivement, et les spectres ont été normalisés avec comme cible le spectre de référence.

Ce prétraitement est basé sur le modèle linéaire suivant :

$$\mathbf{S}_i = \mathbf{a}_i \widehat{\mathbf{S}} + \mathbf{b}_i \mathbf{P} + \mathbf{c}_i \mathbf{I} + \mathbf{e}_i \quad (\text{éq. 19})$$

Chaque spectre  $\mathbf{S}_i$  est modélisé comme la combinaison linéaire des termes suivants:

- $\widehat{\mathbf{S}} \in \mathbb{R}^{1 \times D}$  est une estimation de l'ensemble des données spectrales, c'est-à-dire le spectre moyen de l'ensemble des données :  $\widehat{\mathbf{S}} = \frac{1}{N} \sum_{i=1}^N \mathbf{S}_i$
- $\mathbf{I} \in \mathbb{R}^{M \times D}$  représente les spectres des interférences connues (le verre dans notre cas de figure). La matrice d'interférence  $\mathbf{I}$  du verre est modélisée par une analyse en composantes principales (ACP) grâce à des spectres de verre pur (lame de microscope) acquis au préalable en Raman. Cette matrice d'interférence est composée de  $M$  composantes principales modélisant la contribution du verre. Dans le cas de notre étude, nous avons utilisé 3 composantes principales représentant 99,5% de la variance du signal Raman du verre.
- $\mathbf{b}_i \mathbf{P}$  est un polynôme modélisant la ligne de base. Pour le projet nous avons utilisé un polynôme d'ordre 4.
- $\mathbf{e}_i$  est l'erreur du modèle estimé.
- $\mathbf{a}_i$  est un coefficient de régression de  $\widehat{\mathbf{S}}$  sur  $\mathbf{S}_i$ .
- $\mathbf{b}_i$  est le vecteur des coefficients de régression des composantes de  $\mathbf{I}$  sur  $\mathbf{S}_i$ .
- $\mathbf{a}_i$ ,  $\mathbf{b}_i$  et  $\mathbf{c}_i$  sont calculés par la méthode des moindres carrés afin de minimiser l'erreur de modélisation.

Ce prétraitement permet de neutraliser les variations dues aux interférences (verre) et à la ligne de base, en calculant :  $\mathbf{S}_i^c = \mathbf{S}_i - \mathbf{b}_i \mathbf{P} - \mathbf{c}_i \mathbf{I} = \mathbf{a}_i \widehat{\mathbf{S}} + \mathbf{e}_i$  (éq. 20)

Les spectres sont ensuite normalisés avec pour cible la moyenne  $\widehat{\mathbf{S}}$  de l'ensemble des données, en utilisant la formule :

$$\mathbf{S}_i^N = \frac{\mathbf{S}_i^c}{\mathbf{a}_i} = \widehat{\mathbf{S}} + \frac{\mathbf{e}_i}{\mathbf{a}_i} \quad (\text{éq. 21})$$

La figure 30 montre l'effet de cette méthode sur des spectres Raman de lymphocytes.

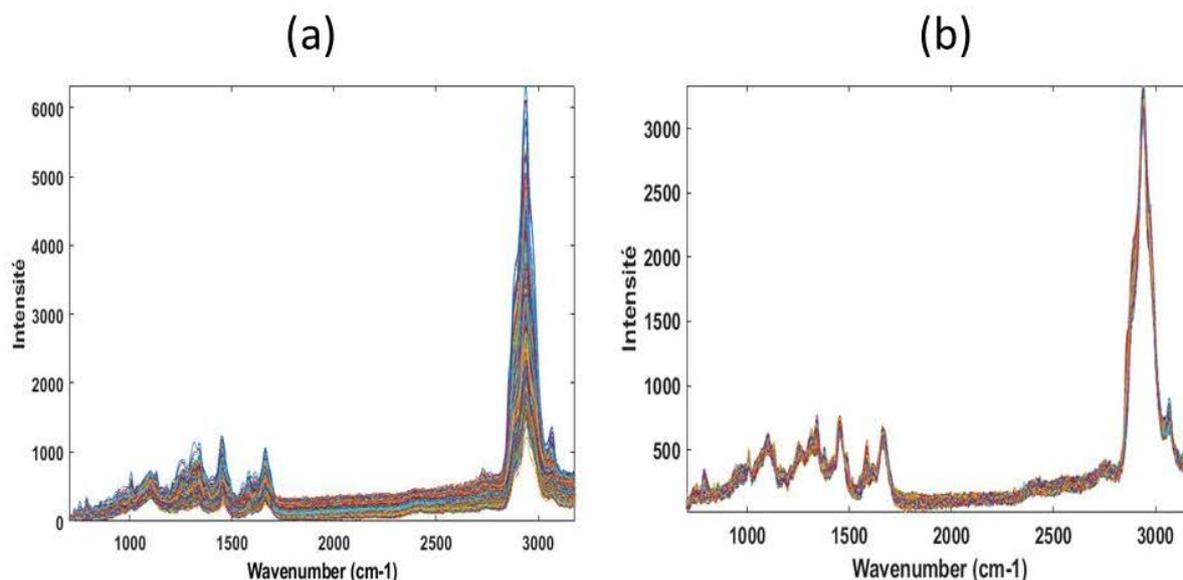


Figure 30 - Effet du prétraitement par EMSC. (a) Spectres Raman de lymphocytes avant l'EMSC, (b) spectres prétraités par EMSC

## 5. Analyses statistiques des spectres

### 5.1. Réduction de la dimension des données

L'objectif de la réduction de la dimension des données est de supprimer les variables redondantes dans l'ensemble de données [136]. Dans le cadre de ce travail, une réduction supervisée de la dimension des données à l'aide des labels cellulaires (sain ou LLC) a été développée. La première étape consiste à calculer le score de Fisher pour chaque longueur d'onde. La deuxième étape consiste à calculer la matrice de corrélation  $\mathbf{R}_\lambda$  pour tous les couples de longueurs d'ondes. À partir de l'ensemble de données. Dans la troisième étape, les éléments de  $\mathbf{R}_\lambda$  étant plus petits qu'un seuil fixé à 0,7 sont supprimés. Chaque ligne de  $\mathbf{R}_\lambda$  est donc composée des nombres d'onde fortement corrélés. Dans la quatrième étape, chaque ligne est réduite à son nombre d'onde ayant le score de Fisher le plus élevé calculé dans la première étape. La dernière étape consiste à calculer les nombres d'onde uniques qui composaient  $\mathbf{R}_\lambda$ , c'est-à-dire à supprimer les nombres d'onde redondantes. Cette méthode permet de trouver les nombres d'ondes les plus caractéristiques des spectres (avec le moins de redondance possible) qui sont également les plus discriminants entre les deux groupes.

### 5.2. Algorithmes de classifications supervisées

Le but de cette étude était de développer un modèle capable de diagnostiquer automatiquement les patients atteints de leucémie lymphoïde chronique grâce à la signature Raman de cellules lymphocytaires. Les algorithmes de classification supervisée sont des

techniques d'apprentissage permettant de réaliser cette tâche. Dans le cadre de ce projet, trois algorithmes différents de classification supervisée ont été choisis pour leurs applications en spectroscopie vibratoire: Partial Least Squares - Discriminant Analysis (PLS-DA) [137], Support Vector Machine (SVM) [127, 128] et Random Forest (RF) [127].

### 5.2.1. PLS-DA : Partial Least Square Discriminant Analysis

PLS-DA est un algorithme de classification supervisée linéaire, dérivé de la méthode de régression linéaire à variables latentes PLS-R (Partial Least Square Regression) introduite par Wold et al [138] (Figure 31). PLS-R est à l'origine conçu comme un outil de régression statistique qui permet de construire un modèle linéaire entre des variables d'entrée explicatives X et des variables expliquées Y. PLS-R réduit les données dans un premier temps en créant des composantes ou variables latentes qui représentent la combinaison linéaire de X auquel on attribue des poids (scores) proportionnels à la covariance entre X et Y.

PLS-DA met en relation une matrice de mesures expérimentales X et un vecteur Y correspondant aux indicateurs des différents groupes de mesure codifiés de façon binaire (0 ou 1). Le but est de trouver une solution maximisant la covariance entre X et Y. Dans le cadre de notre étude, la matrice de variables explicatives X contient les spectres Raman des différents lymphocytes et la matrice de variables à expliquer Y, les labels de ceux-ci (LLC ou sain). Cette méthode implique donc le calcul de modèles de régression entre X et Y.

D'un point de vue mathématique, les matrices X et Y sont décomposées linéairement en matrices de scores (poids) et de loadings (dimensions ou variables latentes) selon les équations fondamentales suivantes :

$$X = TP^T + E_x \quad (\text{éq. 22})$$

$$Y = UQ^T + E_y \quad (\text{éq. 23})$$

**avec T et U les plus corrélés possibles.**

Les matrices P et S contiennent les loadings de X et de Y.  $E_x$  et  $E_y$  sont les éléments résiduels. Il est à noter que PLS est une méthode basée sur un nombre de composantes, ainsi pour obtenir un modèle fiable en PLS-DA, il est opportun de faire varier le nombre de variables latentes  $n_{lv}$  et ainsi éviter le phénomène de sur-apprentissage (over-fitting en anglais). Ce paramètre d'optimisation se trouve grâce à une validation croisée qui sera présentée dans la suite de ce mémoire. Une fois le modèle de prédiction calculé, il est possible de prédire les

valeurs de Y pour des données d'origines inconnues. La prédiction s'effectue en projetant les données inconnues sur le modèle de régression calculée, ce qui permet d'obtenir une matrice contenant des scores pour chaque élément de la matrice X. Ces scores serviront pour attribuer une classe à chaque élément de la matrice X. Par exemple, pour un modèle binaire à 2 classes, l'attribution Y d'un élément de X correspondra donc à la valeur la plus élevée des deux valeurs prédites pour chaque classe. Par exemple, si pour un spectre de la matrice X, le modèle prédit les valeurs [0.95 0.05], l'algorithme retiendra 0.95 et ce spectre se verra attribuer la classe 1.

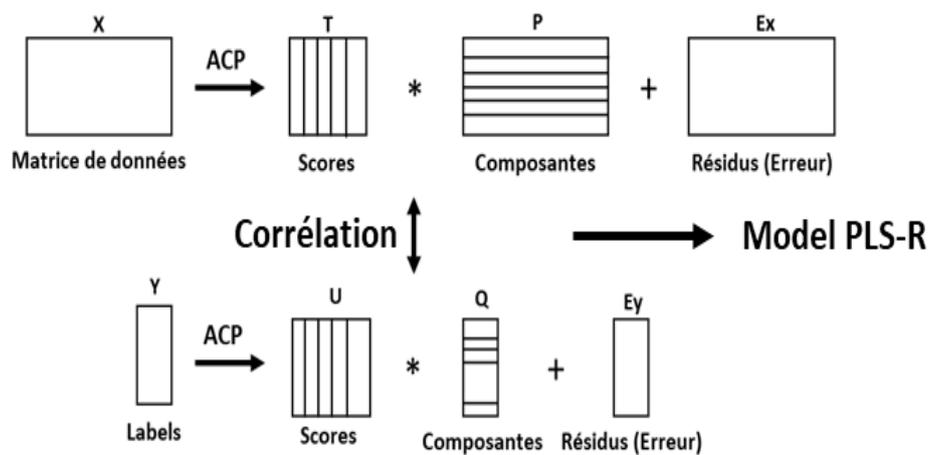


Figure 31 - Principe de l'algorithme PLS-R. X et Y sont les données de départ et les réponses (classes) du modèle, respectivement. T et U sont les scores des variables latentes issues de la modélisation. P et Q correspondent aux loadings [139].

### 5.2.2. SVM : Support Vector Machine ou séparateur à vaste marge

Le SVM est une méthode statistique de classification supervisée qui a été introduite en 1995 par Vapnik [140]. Comme toutes les méthodes supervisées, la méthode SVM repose sur l'apprentissage d'une fonction représentative d'un espace de données permettant de prédire au mieux d'autres données d'entrées qui n'ont pas encore été vues par l'algorithme.

Dans le cadre de notre projet, les données d'entrées sont des spectres Raman et les sorties sont leurs classes. Comme toutes les méthodes d'apprentissage supervisé, le SVM repose sur une phase d'apprentissage grâce à un jeu de donnée d'entraînement labellisé.

Le principe de cette méthode repose sur la recherche d'un hyperplan (Figure 32) qui permet de séparer les données de deux populations et qui en plus permet de maximiser la distance entre ces deux populations (exemple : LLC vs Sains). La distance maximale entre

l'hyperplan et les groupes, s'appelle la marge, d'où la dénomination de « Séparateur à Vaste Marge ». Il existe deux types de méthodes SVM : linéaires et non linéaires.

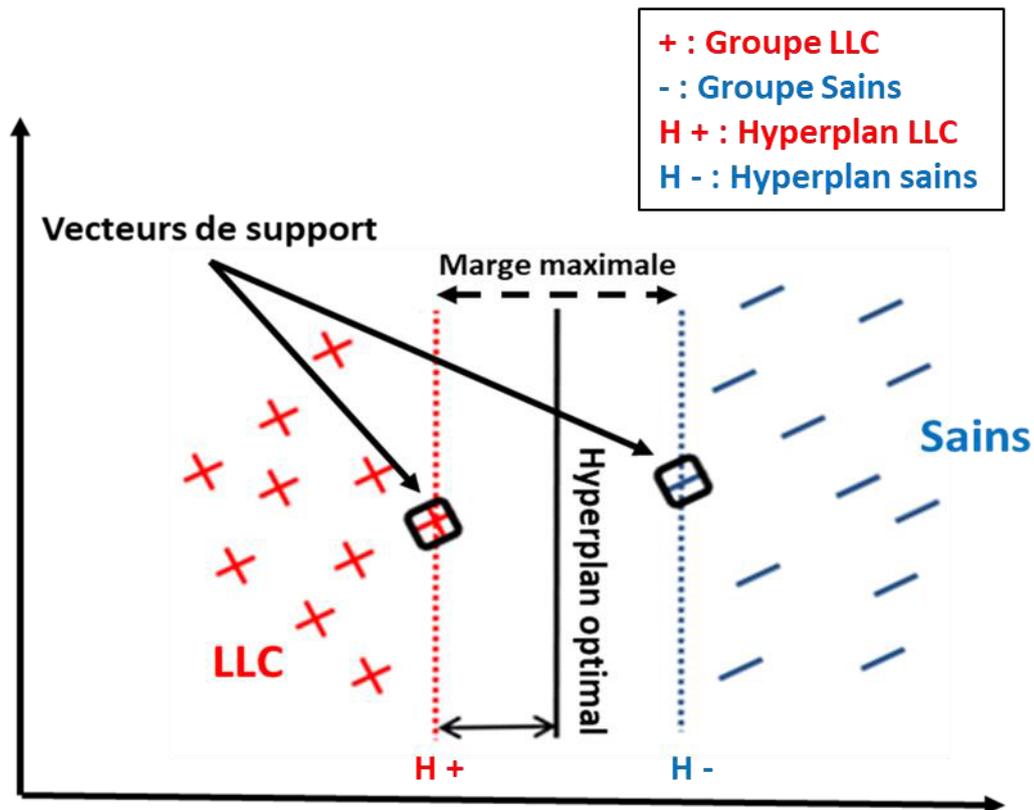
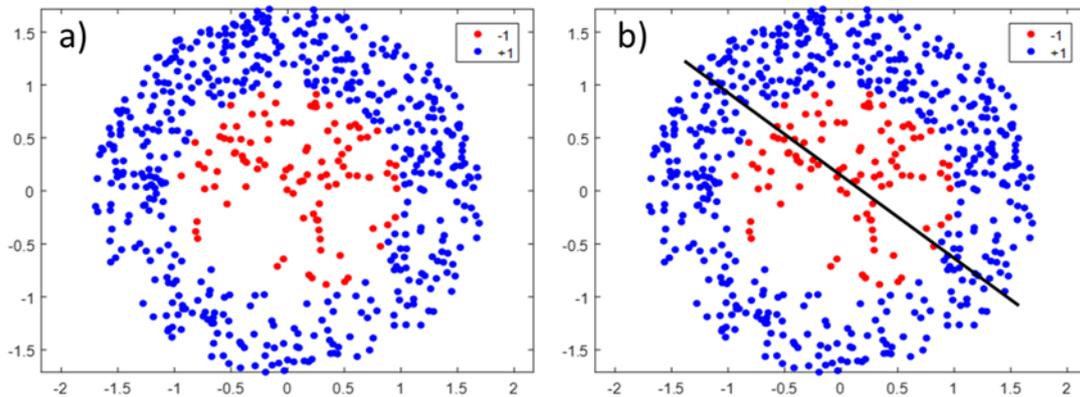


Figure 32 - Schéma présentant le calcul d'un hyperplan pour des données linéaire et séparable par l'algorithme SVM linéaire

Les méthodes linéaires sont les premières à avoir été développées. Le SVM linéaire est fondamentalement la plus simple des deux méthodes, car elle a pour caractéristique de séparer deux groupes linéairement séparables en traçant un hyperplan linéaire optimal entre les deux populations qui permet d'avoir la marge la plus grande possible. Cette séparation linéaire s'effectue sans transformation préalable de l'espace de données. Cependant pour la majorité des applications, ces méthodes linéaires ne sont pas suffisantes pour décrire tous les modèles existants et pour rendre compte des non-linéarités sous-jacentes dans les données. La Figure 33a et la Figure 33b présentent des données non linéairement séparables et l'application d'un hyperplan linéaire. Nous pouvons voir que dans ce cas la moitié des données est mal classée.



*Figure 33 – Représentation en deux dimensions de données imbriquées de façon concentrique et non linéairement séparable par un SVM linéaire.*

Les méthodes non linéaires reposent sur l’usage de fonctions noyaux (ou kernel) qui projettent les données dans un nouvel espace de plus grande dimension, appelé espace de « redescription ». Cette transformation d’espace ou changement de dimensions va permettre d’utiliser les méthodes linéaires dans l’espace de « redescription ». En effet, selon le théorème de Cover [141], plus l’espace de « redescription » sera grand (pouvant aller jusqu’à l’infini), plus la probabilité de pouvoir trouver un hyperplan linéaire séparateur entre les groupes sera élevée. La Figure 34a illustre un exemple d’une projection de données concentriques non linéairement séparables, vers un espace de dimension supérieure où il devient possible de trouver un hyperplan linéaire dans cet espace, permettant de séparer complètement ces deux populations.

Il existe plusieurs fonctions ou noyaux couramment utilisés pour transformer l’espace : polynomiale, sigmoïde, radial basis function (rbf) ou laplacien. La Figure 34b montre l’hyperplan choisi ainsi que les vecteurs de support pour un noyau rbf sur les données non linéaires.

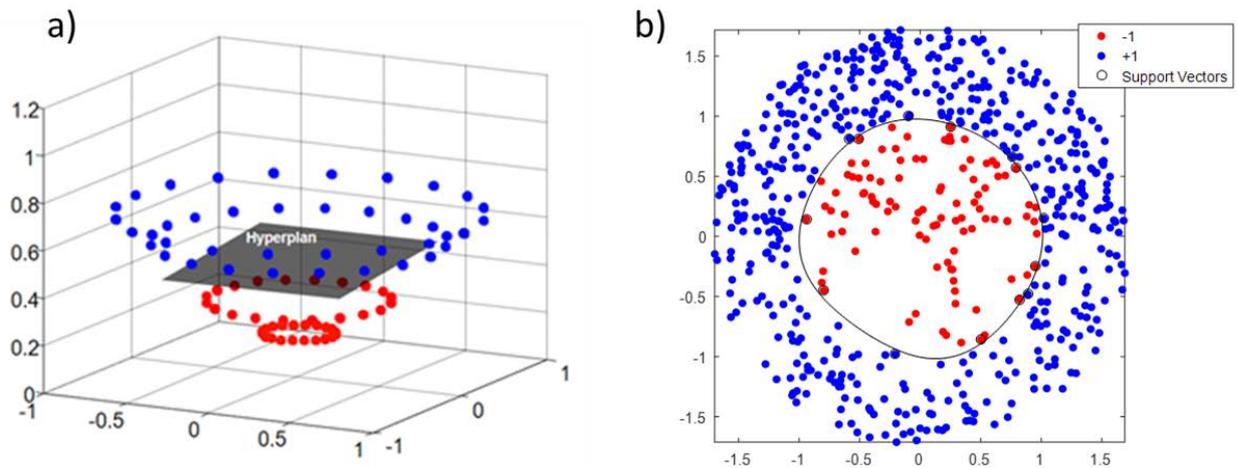


Figure 34 – a) Exemple d'une projection de données non linéairement séparables vers un espace des caractéristiques de dimension supérieure. b) Représentation en deux dimensions du résultat d'une SVM à noyau Rbf. Ici il est possible de voir l'hyperplan optimal choisi par l'algorithme

### 5.2.3. RF : Random Forest

Random Forest est un algorithme supervisé créé dans les années 2000 par Breiman [117]. Il se base sur la technique de « l'arbre de décision » et sur une approche ensembliste, c'est-à-dire que le choix final est pris par un vote majoritaire sur l'ensemble des arbres de la forêt (Figure 35). L'arbre de décision est par défaut un modèle non paramétrique.

Schématiquement, un arbre de décision se construit selon deux étapes. La première est de laisser pousser l'arbre embranchement par embranchement jusqu'à ce que l'information soit maximisée (colle autant que possible aux données). L'arbre grandit en utilisant une méthode de « partitionnement récursif » des données, c'est-à-dire que l'arbre pousse en posant une succession de questions binaires créant ainsi un arbre constitué de feuilles (les classes), de branches (réponses) et de nœuds (questions). Chaque nœud correspondant à une question binaire utilisant une variable du jeu de données et chaque branche partant d'un nœud correspond à une réponse possible à la question posée en ce nœud. La seconde étape est de réaliser un élagage de l'arbre, cela consiste à supprimer certaines branches de l'arbre ne servant à rien, afin de réduire le sur-apprentissage et de simplifier le modèle, le principe de scission des arbres se fait par CART [142].

L'arbre de décision suit des règles de classification basées sur des décisions prises grâce à une suite de tests étant organisés de façon arborescente. Les arbres de décision permettent de

trouver très rapidement les variables les plus discriminantes dans un jeu de donnée, en fonction du nombre de fois qu'ils apparaissent le long des nœuds. Malheureusement, la prédiction individuelle d'un arbre de décision est souvent mauvaise à cause de la variance de la prédiction mais aussi car il est sujet au sur-apprentissage. C'est pour cela que RF utilise une méthode ensembliste afin de résoudre ces problèmes.

Les forêts aléatoires forment donc un ensemble  $n_{dt}$  d'arbres de décisions de taille  $d$  (profondeur) entraînés individuellement (forêt aléatoire décisionnelle), qui diffèrent légèrement les uns par rapport aux autres. Chaque arbre de décision de la forêt possède à son entrée un sous-ensemble de données originelles légèrement différentes car les sous-ensembles sont tirés aléatoirement avec remise, c'est le principe du bagging (Bootstrap Aggregating) [143].

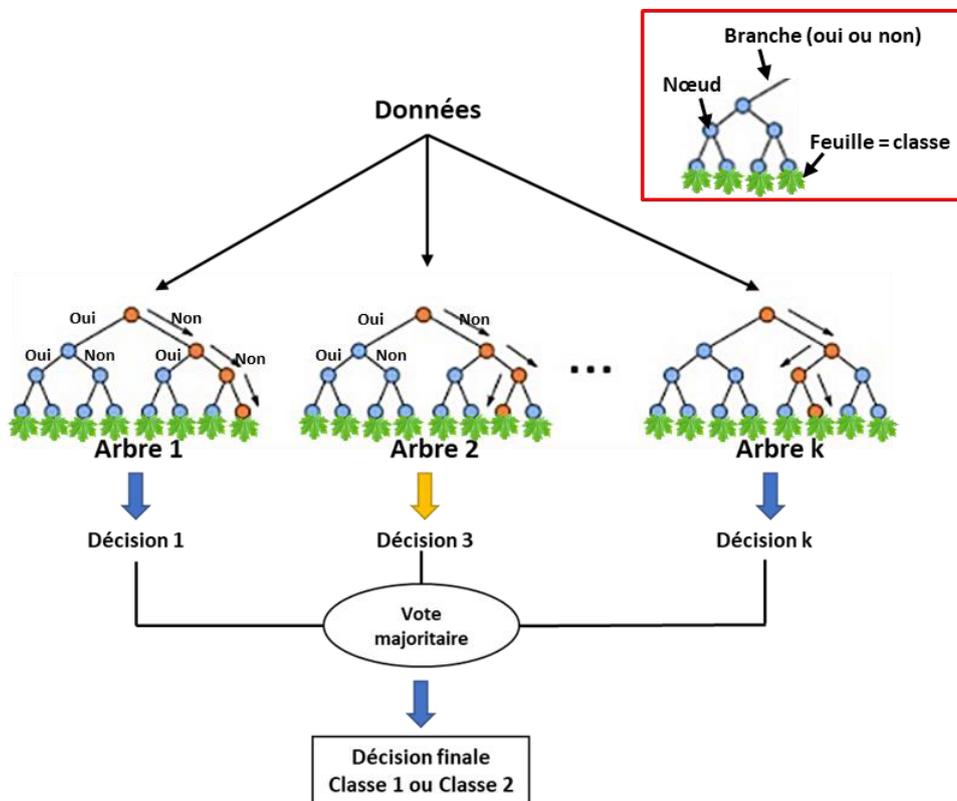


Figure 35 - Schéma de fonctionnement d'une forêt aléatoire décisionnelle

### 5.3. Validation croisée

La validation croisée est un processus qui permet d'évaluer les performances d'un modèle et d'optimiser ses paramètres. Pour mettre en place une validation croisée, la première étape consiste de séparer les données en deux parties: un jeu d'entraînement et un jeu de validation. En général le jeu d'entraînement est constitué des 2/3 des données, tandis que le jeu de validation est composé des 1/3 restant. Le jeu d'entraînement va servir à l'apprentissage et

la validation servira à tester le modèle créé afin d'évaluer le taux de mauvaises classification. Ainsi il sera possible d'évaluer les performances de ce modèle sur des données qui n'ont pas servi pour l'entraînement. Un troisième jeu de données peut également être formé afin de faire une évaluation en aveugle et vérifier les bonnes capacités de généralisation du modèle construit. Différentes méthodes de validation croisées peuvent être utiliser.

### **5.3.1. K-Fold cross-validation (KFCV)**

Dans le **KFCV**, les données  $E$  sont divisées en  $K$  parties (folds en anglais) de tailles identiques (ou quasi-identiques si le nombre de patients n'est pas un multiple de  $K$ ), ce qui induit  $K$  itérations. Tour à tour, chacune des  $K$  parties est utilisée comme jeu de validation. Le reste (autrement dit, l'union des  $K-1$  autres parties) est utilisé pour l'entraînement. Ainsi à chaque itération de la validation croisée un groupe de  $N$  patients est sorti. L'opération se répète ainsi  $K$  fois pour que chaque sous-échantillon ait été utilisé une fois comme ensemble de validation. Dans cette étude, la valeur  $K=10$  a été choisie. À la fin du processus, chaque patient a servi 1 fois dans un jeu de test et  $(K-1)$  fois dans un jeu d'entraînement.

### **5.3.2. Leave-one-patient-out cross-validation (LOPOCV)**

Le **LOPOCV**, est un cas particulier du **KFCV**. Soit  $M$  le nombre de patient total. A chaque itération de la validation croisée, un patient est sorti, ce qui induit  $M$  itérations. A chaque itération, l'ensemble d'entraînement est composé de  $M-1$  patients et le patient restant sert pour la validation. L'opération se répète ainsi  $M$  fois pour que chaque patient soit utilisé une fois comme ensemble de validation. L'inconvénient de cette méthode est que son utilisation augmente fortement le temps de calcul. De plus, les jeux d'entraînements utilisés à chaque itération de la cross-validation, seront très similaires entre eux.

### **5.3.3. Monte Carlo cross-validation (MCCV)**

A chaque itération du **MCCV**, l'ensemble de validation est composé de  $J$  patients sélectionnés au hasard. Les autres  $M-J$  patients composent le set d'entraînement. Dans ce travail, nous avons sélectionné un nombre  $J = 10$  patients pour l'ensemble de validation et un nombre de  $I = 100$  itérations de **MCCV**.

### **5.3.4. Double validation croisée répétée**

Dans le cadre de cette thèse, une stratégie hybride de méthode de validation croisée (**CV**) a été utilisée en raison de sa forte fiabilité dans l'estimation de la performance de

prédiction des modèles [144, 145] : la double validation croisée répétée (rdCV) [146]. La rdCV est la combinaison d'une boucle interne composée d'une CV classique (LOPOCV, KFCV et MCCV), d'une réduction dimensionnelle supervisée des données et d'un algorithme de classification supervisé (PLS-DA, SVM ou RF) pour former un classifieur et optimiser ses paramètres, et d'une boucle externe visant à construire plusieurs modèles à partir de différents jeux d'entraînement et des paramètres optimaux choisis dans la boucle interne. La Figure 36 présente le schéma de fonctionnement de la rdCV. Le processus est le suivant :

(1) A chaque itération de la boucle externe, les patients sont randomisés et répartis en un ensemble d'entraînement composé de 70% de patients et un ensemble de validation externe composé de 30% de patients. Comme un ensemble de données doit toujours être considéré au plus haut niveau hiérarchique afin d'évaluer correctement les modèles de classification [136], les données Raman n'ont pas été prises en compte au niveau spectral, ni au niveau cellulaire, mais au niveau du patient.

(2) A chaque itération de la boucle interne, un ensemble de paramètres de classification est sélectionné dans le cadre d'une recherche par grille (« Gridsearch »), ceci afin d'optimiser ces paramètres pour chaque algorithme de classification, c'est-à-dire le nombre de variables latentes  $n_{lv} \in \{1,2, \dots, 40\}$  pour PLS-DA,  $\gamma \in \{10^{-8}, 10^{-7}, \dots, 10^3\}$  et  $\nu \in \{0.1, 0.2, \dots, 0.9\}$  pour SVM, et le nombre d'arbres de décision  $n_{dt} \in \{20, 30, \dots, 200\}$  et la profondeur  $d \in \{10, 20, \dots, 100\}$  pour RF.

(3) En utilisant ces paramètres, un CV classique est exécuté.

(4) A chaque étape de la boucle CV, l'ensemble d'entraînement interne alimente la réduction dimensionnelle des données supervisées afin d'identifier les caractéristiques les plus discriminantes et les plus décorréelées. Comme suggéré dans l'article [136], la méthode de réduction des données est incluse dans la boucle interne de validation croisée afin d'empêcher le sur-apprentissage du modèle.

(5) Sur ces données réduites, un modèle de classification supervisé est formé en utilisant les paramètres sélectionnés à l'étape (2).

(6) L'ensemble de validation interne est réduit aux caractéristiques estimées à l'étape (4).

(7) Ces données sont injectées dans le modèle construit à l'étape (5) afin d'estimer sa performance.

(8) A la fin de la boucle CV, les sensibilités et les spécificités sont utilisées pour calculer la précision de l'algorithme de classification supervisé en utilisant les paramètres sélectionnés à l'étape (2).

(9) A la fin de la boucle interne, c'est-à-dire à la fin de la « Gridsearch », les paramètres de classification optimaux sont estimés comme étant ceux qui maximisent la précision calculée à l'étape (8).

(10) L'ensemble de la formation est utilisé pour identifier l'ensemble des caractéristiques les plus discriminantes et décorréliser par la réduction supervisée de la dimension des données.

(11) L'ensemble d'entraînement réduit est utilisé pour calculer un modèle de classification unique en utilisant les paramètres optimaux estimés à l'étape (9).

(12) Les données de validation externe sont réduites également avec les caractéristiques estimées à l'étape (10).

(13) Ils sont injectés dans le modèle construit à l'étape (11) afin d'estimer sa performance de généralisation en termes de sensibilité et de spécificité. Les étapes (1) à (13) sont répétées T = 100 fois, ce qui permet de construire 100 modèles optimisés.

*Tableau 4 - Nombre de patient et de cellules en entraînement, validation et test par tour*

	Set d'entraînement		Set de validation		Set de test	
	Patients	Cellules	Patients	Cellules	Patients	Cellules
<b>IMHO</b>	30	3221	20	1322	20	1456
<b>M3S</b>	37	2353	13	805	20	1358
<b>IMHO + M3S</b>	70	5088	30	2613	40	2814

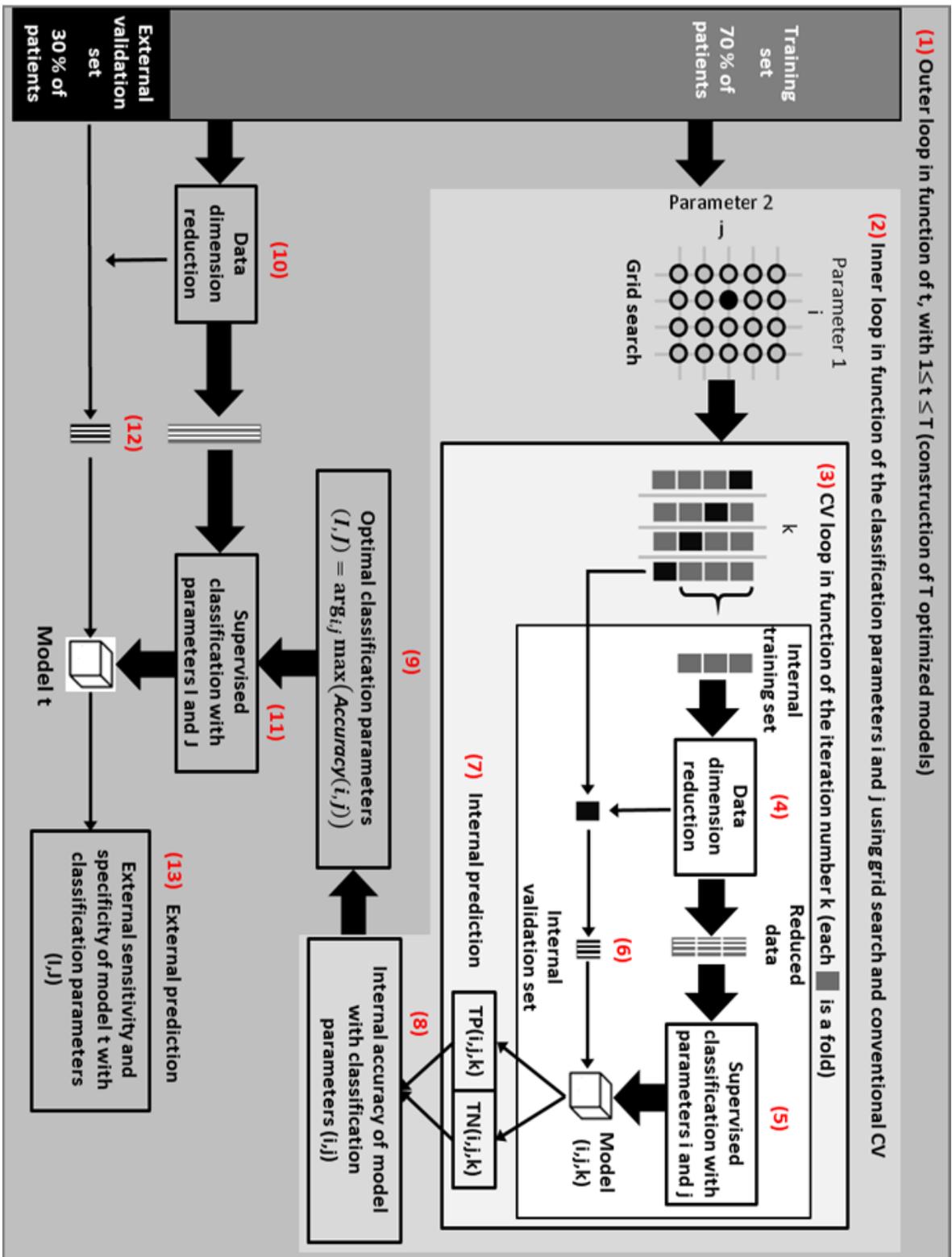


Figure 36 - Diagramme du principe de fonctionnement de la double validation croisée répétée (rdCV)

#### 5.4. Test en aveugle des classificateurs optimisés : vote majoritaire

La dernière étape consiste à tester les 100 modèles sur un ensemble indépendant de patients qui n'ont pas été utilisés dans le rdCV. Chaque modèle optimisé prédit l'état physiopathologique de chaque patient testé, c'est-à-dire LLC ou non LLC. Pour un patient test, chacun des 100 modèles optimisés donnera sa prédiction et la décision finale sera obtenue en utilisant une méthode ensembliste, c'est-à-dire un vote majoritaire (Figure 37). Tout comme pour l'algorithme de Random Forest, cette méthode permet d'agréger les réponses des classifieurs, réduire l'effet du sur-apprentissage et booster les résultats de prédiction. Bien sûr cette méthode n'a d'intérêt que si les classifieurs sont suffisamment différents les uns des autres. C'est pour cela que nous avons mis en place un tirage aléatoire des patients au début de chaque boucle externe de rdCV.

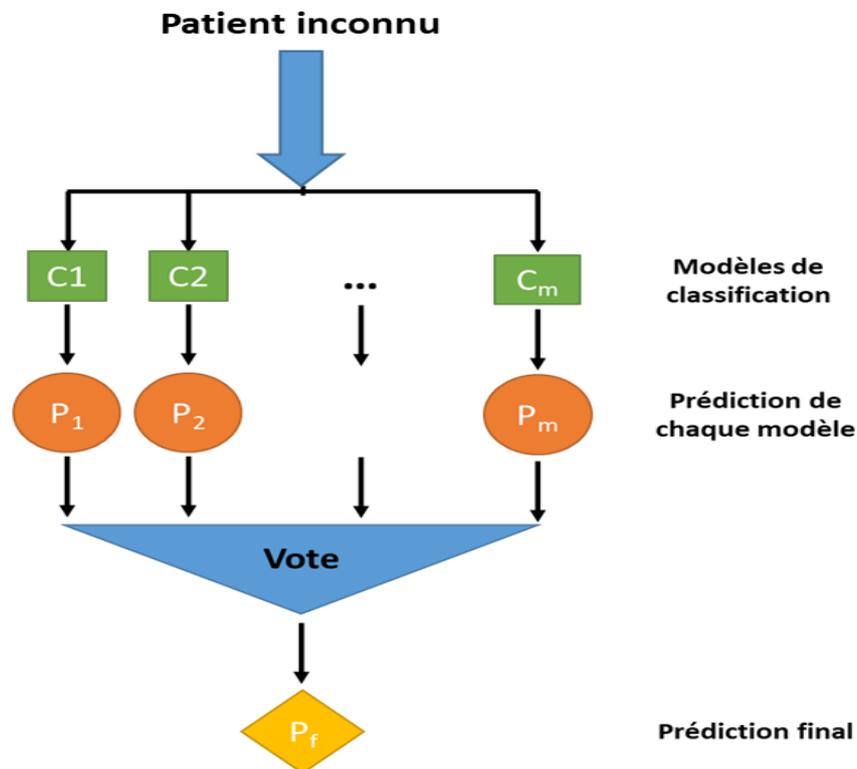


Figure 37 - Principe du vote majoritaire pour le diagnostic d'un nouveau patient

# **CHAPITRE 3 : CONSTRUCTION D'UNE BASE DE DONNEES MULTIMODALES**

Comme cela a été décrit dans la section « Matériels et méthodes », un groupe de patients en bonne santé et un groupe de patients atteints de LLC n'ayant subi aucun traitement, ont été formés à partir des deux campagnes de mesure différentes. La première campagne a été réalisée dans le cadre du projet IHMO financé par l'ANR (Agence Nationale de la Recherche). Tous les patients ont été recrutés au centre hospitalier de Reims Champagne-Ardenne (RCA-HC). La deuxième campagne de mesure a été réalisée dans le cadre du projet M3S financé par la Communauté européenne. Les patients ont été recrutés au centre hospitalier du Mont Godinne-Namur (MGN-HC). Tous les frottis ont été préparés directement à l'hôpital de recrutement, en étalant une goutte de sang déposée sur une lame de verre afin de travailler dans des conditions cliniques classiques. Chez RCA-HC, un étalement manuel est utilisé, tandis qu'un étalement automatique avec un dispositif de frottis sanguin automatisé (HemaPrep,) est préféré chez MGN-HC, les deux réalisés sans traitement chimique préalable. Pour les deux campagnes de mesure, les données Raman ont été acquises avec le même dispositif multimodal développé par la société TRIBVN (Châtillon, France). Le système combine un microscope conventionnel (ECLIPSE FN1, Nikon SA, Champigny-sur-Marne, France) et un spectromètre Raman (HORIBA FRANCE SAS, France).

Afin de comparer les données spectrales avec d'autres données cliniques et biologiques, une banque de données a été construite dans laquelle chaque signature spectrale est associée aux caractéristiques du patient analysé. Ainsi la première phase de cette thèse a été de réaliser de nombreuses acquisitions afin d'alimenter cette banque de données au fur et à mesure de l'avancée du projet. L'ensemble des données acquises sur les lymphocytes comprennent les spectres Raman, mais aussi des images en intensité et en Phase Quantitative (QPI). Pour développer les modèles prédictifs, une coloration par immunohistochimie a également été effectuée et des images couleurs correspondantes à chaque cellule analysée en Raman ont été recueillies. Le travail de cette thèse a été axé sur les analyses Raman. Ainsi au niveau de notre laboratoire, un calendrier très précis a été mis en place afin d'assurer une immunocoloration efficace des frottis après les analyses morphologiques, de phase et Raman. In fine, toutes les informations de cette base de données seront croisées statistiquement afin de construire un modèle pronostic.

## 1. Les différentes étapes de production de la base de données

Dans un premier temps, des données cliniques relatives aux caractéristiques du patient, à la numération cellulaire par cytométrie en flux et à la morphométrie ont été collectées pour fournir des vérités terrain, afin de développer ultérieurement des modèles de prédictions. Dans un second temps, les acquisitions ont été effectuées selon un protocole définies contenant différentes étapes manuelles et semi-automatique (figure 38) afin de maintenir des flux d'acquisition les plus fluides possibles (QPI, images d'intensité, Raman et images couleurs). Ces acquisitions ont permis de construire des classifieurs morphologiques et spectroscopiques capables de distinguer les patients sains et pathologiques mais également de reconnaître les lymphocytes des autres types de cellules. Ces acquisitions ont permis également de regrouper des informations morphologiques et moléculaires pour chaque cellule analysée.

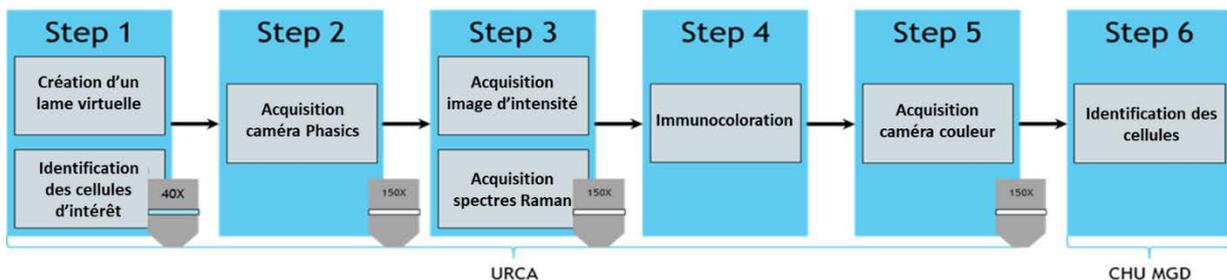


Figure 38 – Représentation des différentes étapes du flux d'acquisition d'un frottis sanguin sur le système M3S

### 1.1. Formulaire clinique électronique : e-CRF

Au moment de l'envoi de chaque frottis sanguin, les données cliniques des patients ont été recueillies par saisie électronique dans un formulaire de synthèse (e-CRF) développé par l'entreprise Quinten. Ce système a été conçu conformément aux normes actuelles de gestion des données.

L'e-CRF permet de regrouper toutes les données personnelles et cliniques de chaque patient (âge, sexe, numération, facteurs pronostiques). Toutes les données informatiques ont été saisies par le clinicien à l'aide d'une interface graphique. L'accès au système était contrôlé et limité au moyen d'un numéro d'identification individuel et d'un mot de passe. Chaque patient enregistré était identifié par un numéro d'identification unique et anonyme. Des contrôles systématiques des données d'entrées par une seconde personne, ont permis de limiter les erreurs de saisie des données. Chaque mise à jour des données ont fait l'objet d'un suivi.

## **1.2. Scanneur de lame et identification des cellules d'intérêt**

La première étape du processus d'analyse est l'acquisition par imagerie classique d'un frottis virtuel. Pour chaque région d'intérêt (ROI) du frottis, le système construit une lame virtuelle de la zone à analyser grâce à un objectif 40X et un frottis défocalisé de  $-4\mu\text{m}$  du même ROI. Sur la vue défocalisée, les cellules nucléées apparaissent comme de petits cercles blancs en raison de la diffraction du noyau. Un programme d'analyse d'images morphométrique permet de classer les lymphocytes par taille ce qui permet de les localiser sur le frottis. En effet, ces cellules sont les plus petites cellules nucléées du frottis.

## **1.3. Imagerie à phase quantitative : QPI**

Une fois les cellules d'intérêt repérées sur la lame virtuelle, une première phase d'acquisition a été réalisée par imagerie QPI. Il s'agit d'une double caméra qui acquiert 2 types d'images, en intensité monochrome et en phase quantitative. Les images d'intensité et de phase quantitative fournissent des informations optiques différentes sur les cellules non colorées.

L'imagerie en phase quantitative est liée à la densité optique de la matière. La QPI repose sur le déphasage qui se produit lorsque la lumière passe à travers des objets biologiques de densité diverse. Par conséquent, le contraste d'image acquis est fonction de l'indice de réfraction du matériau (cytoplasme, noyau, membrane ...). L'utilisation de la technologie QPI fournit des images très informatives sans préparation au préalable de l'échantillon et est prévue pour accélérer le flux de travail en permettant l'identification automatique des cellules pertinentes, c'est-à-dire les lymphocytes, au moyen d'un programme de reconnaissance automatique des cellules basé sur la technologie du Deep Learning.

Cette technologie a permis également d'améliorer la caractérisation des lymphocytes en apportant des informations optiques supplémentaire à la spectroscopie Raman. En utilisant l'imagerie d'intensité à haute résolution et de phase quantitative, une technique de visualisation pour améliorer la discrimination entre lymphocytes et autres leucocytes à partir de frottis non colorés a été développée par le centre de Morphologie Mathématique de l'école des Mines de Paris.

## **1.4. Spectroscopie Raman et imagerie monochrome à haute résolution**

Troisième étape : chaque cellule analysée en QPI est ensuite analysée en Raman puis simultanément capturer en imagerie d'intensité haut résolution.

## 1.5. Nomenclature

Afin de faciliter les échanges de données entre les partenaires et identifier les cellules analysées avec chaque modalité, une nomenclature a été mise place. Pour chaque cellule de chaque patient, un code (figure 39) a été attribué comprenant le statut du patient c'est-à-dire LLC, sain ou atteint d'une virose. La seconde partie du code correspond au numéro du patient, ensuite vient le type de lame soit frottis sanguin ou soit cellule triée, puis le numéro de ROI, et enfin le numéro de la cellule.

Exemple : **P2\_ULC23\_M1\_L1\_ROI1\_C1** désigne la première cellule acquise au 150X, dans la première région d'intérêt, du premier frottis sanguin, du 23ème patient recruté au CHUMGD qui est atteint par la LLC. Les données QPI, image d'intensité et Raman ont la même nomenclature afin de faciliter les échanges de données.

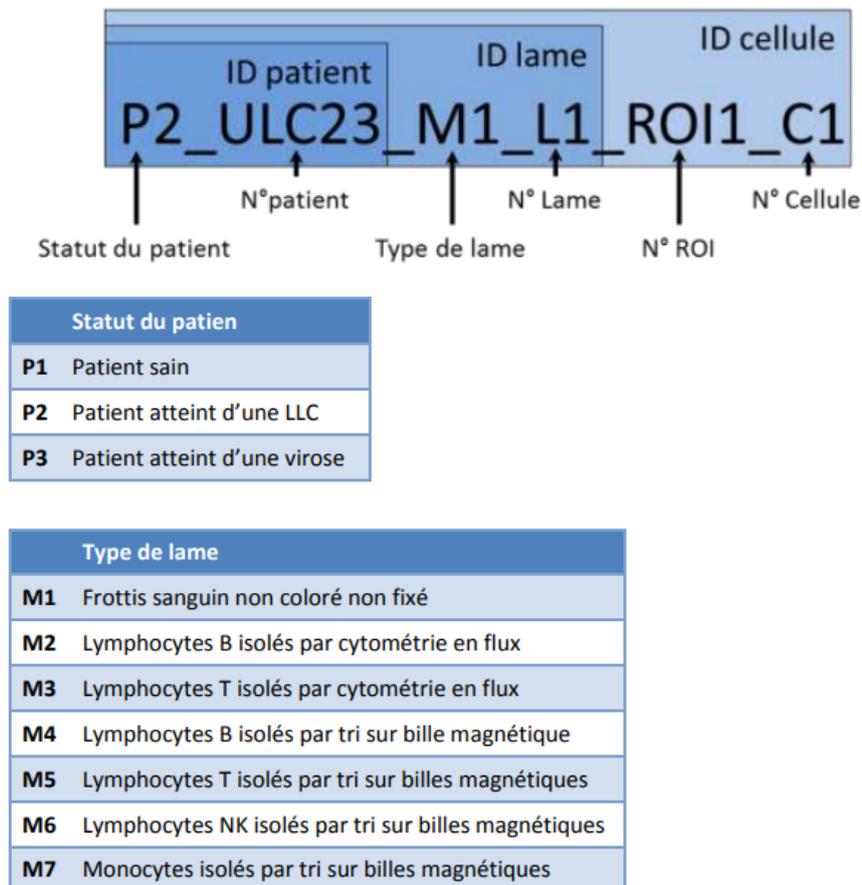


Figure 39 – Exemple de nomenclature type utilisée pour identifier chaque fichier de données

## **1.6. Identification de la classe des cellules analysées**

Après la phase d'analyse, la quatrième étape est l'immunomarquage. Cet immunocoloration a permis aux cliniciens d'identifier pour chaque lymphocyte analysé, la nature de celui-ci. Les cellules d'intérêt sont examinées par les cliniciens pour déterminer leur classe lymphocytaire. Pour réaliser cette identification, une photo a été acquise manuellement grâce à une caméra couleur pour chaque cellule analysée par les différentes modalités. Cette étape est très importante car le marquage des sous-types de lymphocytes sera considéré comme vérité terrain pour la construction des classificateurs morphologiques et Raman.

## **2. Problèmes rencontrés**

Lors du projet, différents problèmes ont été identifiés remettant en cause l'efficacité de l'immunomarquage comme Gold standard fiable et reproductible. En effet, dans certains lots d'images, aucune cellule ne porte de marquage ou le marquage est trop faible pour être visible. L'origine de ce problème peut provenir de différents cas de figure :

- ✓ Un échec du marquage lié au marqueur ou aux conditions d'expérience
- ✓ Un échec du marquage lié au vieillissement de la lame
- ✓ L'absence de cellules marquées parmi les cellules imagées.

Pour résoudre ce problème, une vérification systématique à l'œil nu a été réalisé sur chaque lame par un clinicien afin de vérifier la qualité du marquage cellulaire. Un autre problème qui a pu être rencontré, est le nombre d'acquisition possible de lymphocytes B sains. En effet, la méthode actuelle ne permet pas de recueillir assez de spectres Raman sur cette catégorie de lymphocytes très minoritaire. Les acquisitions sur cellules triées pourraient pallier ce manque de données.

# **CHAPITRE 4 : CONSTRUCTION DE MODELES RAMAN POUR LE DIAGNOSTIC AU NIVEAU PATIENT**

**Article #1 : Privilégier l'étape de prétraitement pour assurer la transférabilité clinique de données Raman multicentriques acquises sur les lymphocytes dans différentes conditions expérimentales et instrumentales.**

---

**Focus on pre-processing step to ensure the clinical transferability of Raman data acquired on lymphocytes in different experimental and instrumental conditions**

**M. Féré, O. Piot, L.H Liu, A. Beljebbar, V. Untereiner, D. Gheldof, M. Chollat, J. Klossa, B. Chatelain, C. Gobinet**

**Article soumis dans le journal Elsevier : Vibrational Spectroscopy.**

---

## **Préambule à l'article 1 :**

### **Contexte**

La spectroscopie Raman est une technique biophotonique de haute spécificité qui permet d'avoir accès à l'information biochimique d'un échantillon biologique. Elle génère des données complexes sous forme de spectres où chaque bande ou ensemble de bandes caractérisent des vibrations moléculaires spécifiques. Cependant cette technologie est également sensible aux changements expérimentaux et instrumentaux susceptibles de générer de la variabilité au sein des données Raman. Ainsi il se pose des problèmes de transférabilité lorsque l'on travaille sur des données multicentriques acquises dans différentes conditions expérimentales et instrumentales. Pour pouvoir apporter cette technologie au chevet du patient, il est nécessaire d'assurer la transférabilité de la méthodologie d'une étude à l'autre. Le choix du prétraitement des données est donc une première étape importante afin de créer des modèles de prédictions performants.

### **Objectif**

Dans la cadre de cette étude, nous avons examiné le rôle du prétraitement numérique pour compenser cette variabilité du signal introduite par différentes conditions expérimentales et instrumentales. Pour cela, nous avons comparé des données Raman de lymphocytes acquises au cours de deux campagnes de mesures indépendantes et multicentriques. Ces données ont été acquises à partir de frottis sanguins frais non-colorés et étalés sur lame de verre de personnes saines et de personnes atteintes de leucémie lymphoïde chronique (LLC) à un stade avancé. Entre ces campagnes de mesures, des modifications ont été apportées au niveau de la configuration instrumentale des dispositifs Raman, de la méthode de préparation des frottis, de la méthode d'acquisition.

### **Matériels et méthodes**

Les données Raman ont été analysées avec un dispositif multimodal développé par la société TRIBVN (Châtillon, France) combinant à la fois l'imagerie optique et la spectroscopie Raman. Les données Raman ont été acquises avec une source d'excitation laser de 532 nm, d'un réseau holographique de 1200 g/mm et d'un détecteur CCD (Charge Coupled Device). Pour la première campagne de mesures, un seul spectre Raman a été acquis grâce à un objectif 100x, sur le centre du noyau des lymphocytes avec un temps d'acquisition fixé à deux accumulations

de 10 secondes. Pour la seconde campagne de mesures, trois spectres Raman ont été acquis avec un objectif 150x, en trois positions différentes sur le noyau, avec un intervalle de 1  $\mu\text{m}$ . Le temps d'acquisition a été fixé à une accumulation de 10 secondes. Les frottis ont été réalisés en condition clinique sans traitement chimique préalable, avec un étalement manuel pour la première campagne de mesure et un étalement automatique pour la seconde campagne de mesure. Deux protocoles de prétraitement différents ont été appliqués dans cette étude. Le premier prétraitement, qui a été développé dans une étude précédente, est une méthode séquentielle appliquée individuellement à chaque spectre. Le second prétraitement est basé sur la correction multiplicative étendue du signal (EMSC). Afin de comparer les performances des deux méthodes, nous avons utilisé l'analyse en composantes principales (ACP) pour étudier la source de variabilité entre les deux campagnes de mesure et la méthode d'apprentissage supervisé PLS-DA afin d'évaluer le degré de séparabilité entre les patients sains et les patients atteints de LLC après prétraitement.

## **Résultats**

Les résultats ont montré que la première méthode de prétraitement (séquentiel) s'est avérée inefficace sur ces données multicentriques. En effet, une séparation des données a pu être observée avec le prétraitement séquentiel, ce qui démontre une hétérogénéité au sein des données. A contrario, l'EMSC a été capable d'homogénéiser ces ensembles de données afin de construire un modèle prédictif robuste pour le diagnostic de la leucémie lymphoïde chronique dans le cas de stades avancés.

## **Conclusion**

Ainsi, cette étude contribue à démontrer qu'il est possible de réaliser une transférabilité clinique de l'approche Raman malgré des modifications protocolaires ou instrumentales grâce à l'utilisation d'un prétraitement numérique adéquat. La seconde étape consistera à déterminer une stratégie permettant de construire des modèles de prédiction robustes sur les données provenant de patients.

# Focus on pre-processing step to ensure the clinical transferability of Raman data acquired on lymphocytes in different experimental and instrumental conditions

M. Féré<sup>1</sup>, O. Piot<sup>1,2</sup>, L.H Liu<sup>1</sup>, A. Beljebbar<sup>1</sup>, V. Untereiner<sup>2</sup>, D. Gheldof<sup>3</sup>, M. Chollat<sup>4</sup>, J. Klossa<sup>4</sup>, B. Chatelain<sup>3</sup>, C. Gobinet<sup>1</sup>

<sup>1</sup>University of Reims Champagne-Ardenne, BioSpecT EA 7506, Faculty of Pharmacy, Reims, France

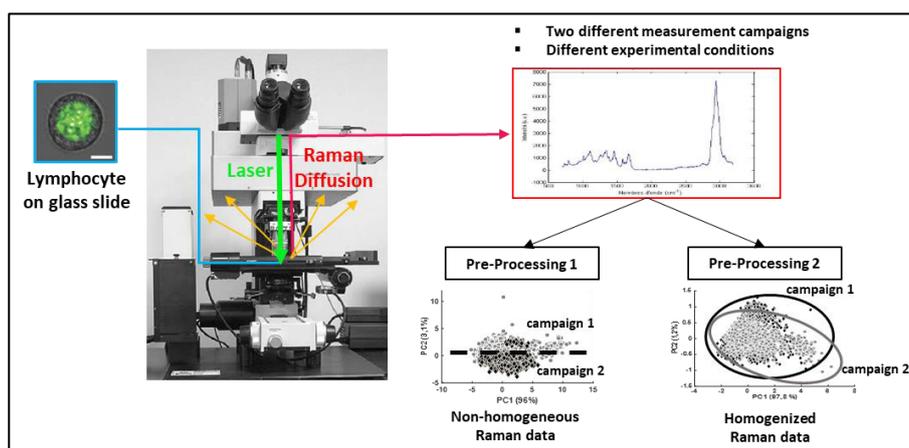
<sup>2</sup>Cellular and Tissular Imaging Platform PICT, Faculty of Pharmacy, University of Reims Champagne-Ardenne, Reims, France

<sup>3</sup>Catholic University of Louvain, CHU UCL Namur, Namur Thrombosis and Hemostasis Center, Hematology Laboratory, Yvoir, Belgium

<sup>4</sup>TRIBVN, Châtillon, France

## ABSTRACT

The efficiency of Raman spectroscopy for the analysis of biomedical samples has been largely demonstrated at the proof-of-concept stage during one-off measurement campaigns. However, bringing these results to the patient bedside requires to fill the gap for transferability of Raman data acquired in different experimental and instrumental conditions during multicentric measurement campaigns. In this study, we propose to evaluate a solution consisting in the application of Raman data pre-processing specifically developed to remove the spectral variability induced by such different conditions. For this purpose, we compared Raman data of lymphocytes acquired during two independent measurement campaigns from fresh unstained glass blood smears originating from healthy individuals and patients with a B-cell chronic lymphocytic leukemia (CLL) at an advanced stage. The differences between these campaigns were i) the instrumental configuration of the Raman devices, ii) the hospital partner, iii) the smear preparation method. A pre-processing developed previously and efficient for a specific measurement campaign is shown obsolete for these multicentric data. A second pre-processing based on Extended Multiplicative Signal Correction is able to homogenize the data by neutralizing the signal variability between the two measurement campaigns. These conclusions are drawn from the analysis of the data by Principal Component Analysis to study the source of variability between the two campaigns and by Partial Least Squares – Discriminant Analysis to assess the separability between healthy and CLL patients.



# 1. INTRODUCTION

Many technologies exist to diagnose leukemic cells, such as morphometric imaging[1], flow cytometry[1] or mass spectroscopy[2]. However, these techniques require important human resources, consumables and external labelling or specific preparation of sample preventing its analysis by another complementary modality.

Raman spectroscopy appears as an alternative method, which is a label free and nondestructive biophotonic technology probing the biochemical composition of a sample. Subtle pathophysiology -linked variations in a biological sample have a direct impact on its molecular signature, leading to detectable modifications in its Raman spectrum.

In recent years, many applications of Raman spectroscopy have been reported in the biomedical field, particularly at the cellular level, e.g. for the identification of cell state transitions[3-5], for the detection of epigenetic changes in living cells[6], for the identification of immune cell categories[7, 8], or for the detection of immune cell activation[7, 9, 10]. All these studies show that Raman microspectroscopy could be used as a new diagnostic tool in cytopathology.

However, this translational research remains far from the patient's bedside for several reasons[11], including the sensitivity of this technology experimental conditions, environment or sample preparation mode, that can generate detectable pernicious changes in the Raman data. Adaptive computational pre-processing methods are thus needed to develop transferable models able to compensate for the high sensitivity of Raman spectroscopy to these interferences changes.

In this context and based on two measurement campaigns (2010,2016) at two different sample periods with different experimental conditions, our objective was to evaluate the impact of these changes on Raman spectra acquired on lymphocytes from blood smears of healthy and B-cell Chronic Lymphocytic Leukemia (B-CLL) patients. Two different pre-processing procedures using a sequential method and a parallel method based on Extended Multiplicative Signal Correction (EMSC)[12] were developed in order to correct these spectral modifications due to the different experimental conditions. Their ability to homogenize the two different datasets were evaluated by Principal Component Analysis (PCA) and Partial Least Squares – Discriminant Analysis (PLS-DA).

## **2. MATERIALS AND METHODS**

### **2.1. Patients and sample preparation**

In this study, one group of healthy patients and one group of untreated B-CLL patients, with a Matutes score over 3 and a stage A in the Binet classification, were formed from the two different measurement campaigns. This study was approved by local ethics committees. The two campaigns were conducted under different experimental conditions (Table 1). The first campaign was achieved in the years 2010-2011 during the IHMO project funded by the ANR (Agence Nationale de la Recherche). All patients were recruited at the Reims Champagne-Ardenne hospital center (RCA-HC), as previously described[13]. The second campaign was performed in the years 2015-2016 during the M3S project funded by the European Community. All patients were recruited at Mont Godinne-Namur hospital center (MGN-HC).

Taken together, both campaigns resulted in the inclusion of 41 healthy patients (15 from IHMO and 26 from M3S) and 59 untreated B-CLL patients (35 from IHMO and 24 from M3S).

For each patient, one smear was prepared directly in the recruitment hospital according to its specific standard clinical conditions. At RCA-HC, a manual spreading is realized by depositing a blood drop on a conventional glass slide and spreading it with another glass slide. At MGN-HC, an automatic spreading is preferred. Smears are produced by depositing a blood drop on a conventional glass slide and spreading it with an automated device (HemaPrep). In the two recruitment hospitals, all samples were prepared without prior chemical treatment.

### **2.2. Acquisition of Raman spectra**

For the two measurement campaigns, Raman data were acquired with a multimodal device developed by the TRIBVN company (Châtillon, France). The system combines a conventional microscope (ECLIPSE FN1, Nikon SA, Champigny-sur-Marne, France) and a Raman spectrometer (HORIBA FRANCE SAS, France). The microscope was equipped with a motorized XYZ stage (Ludl Electronic Products Ltd, New-York, USA) and two dry lenses (Nikon): i) a 40x lens (NA 0.6) for automatic smear exploration and localization of lymphocytes, ii) a high magnification lens used for Raman acquisitions on lymphocytes. Between the two campaigns, different objectives with specific characteristics were used as indicated in Table 1.

The Raman spectrometer was composed of a 532 nm laser excitation source (Toptica Photonics, Munich, Germany) delivering a power of 13 mW on the sample, a holographic grating of 1200 g/mm and a -70°C cooled-Pelletier CCD (Charge Coupled Device) detector (Andor Technology, South Windsor, CT, USA) of 1024 x 256 pixels and a 100 µm confocal hole. This set-up leads a spectral range from 700 to 3170 cm<sup>-1</sup> and a spectral resolution of 4 cm<sup>-1</sup>.

For the IHMO campaign, the data acquisitions were realized in our laboratory by a biophysicist. A single Raman spectrum was acquired on the center of nucleus of 1862 healthy and 2681 CLL cells with an acquisition time fixed to two accumulations of 10 seconds, corresponding to 4543 acquired Raman spectra.

For the M3S project, half of the data acquisitions was realized in our laboratory by a biophysicist and the other half by a clinical technician in a clinical environment (MGN-HC, Belgium). Three Raman spectra were acquired at three different positions on the nucleus, with an interval of 1 µm, for one accumulation of 10 seconds, corresponding to 1449 healthy and 1709 CLL cells. For each cell, the three spectra were averaged, leading to a data set of 3158 spectra for this campaign.

Each cell is thus characterized by its Raman spectrum and its physiopathological label, healthy or CLL as determined on the basis of clinical criteria, used in further supervised classification. For each measurement campaign, 77 cells were analyzed in average per patient.

*Table 1 - Experimental and instrumental differences between the IMHO and M3S measurement campaigns.*

	<b>IMHO</b>	<b>M3S</b>
<b>Microscope lens</b>	100X/NA: 0,90	150X/NA: 0,95
<b>Blood smear spreading method</b>	Manual	Automatic
<b>Number of accumulations</b>	2	1
<b>Number of spectra per cell</b>	1	3
<b>Number of patients analyzed</b>	25 healthy and 45 B-CLL patients	36 healthy and 34 B-CLL patients
<b>Experimenter</b>	by a biophysicist	½ by a biophysicist and ½ by a clinical technician
<b>Acquisition place</b>	In a research laboratory (Reims, France)	½ in a research laboratory (Reims, France) and ½ in clinical environment (MGN-HC, Belgium)

### **2.3. Spectral data pre-processing**

To keep only relevant cell information for further classifications, it is essential to correct Raman spectra from unwanted effects such as noise, glass background, baseline mainly due to cell fluorescence, and variation in acquisition volume.

To reduce noise in the Raman spectra, a Savitzky-Golay smoothing[14] has been applied using a window length of 9 points and a third order polynomial function.

In this study, two different pre-processing protocols were applied separately, as two different options, on the whole dataset. The first one (Pre-processing 1), used during the IHMO project[13], is a stepwise method applied individually to each spectrum. The contribution of the glass signal was estimated in each spectrum by fitting a mean reference Raman spectrum of glass by ordinary least squares. This contribution was subtracted from the spectrum. Then, the baseline was corrected using a second order polynomial function. Finally the spectrum was normalized using the standard normal variate (SNV) method[15].

The second pre-processing protocol (Pre-processing 2) consists in the application of a joined method, i.e. pre-processing the Raman spectra altogether, based on Extended Multiplicative Signal Correction (EMSC)[12]. This method uses the mean spectrum of the data set as the reference spectrum to guide all corrections. The baseline and the glass signal were neutralized using a fourth order polynomial function and a mean Raman spectrum of glass[16], respectively. The dataset is normalized around the reference spectrum.

### **2.4. Principal Component Analysis**

Principal Component Analysis (PCA) is an unsupervised data analysis tool used to explore the structure of a dataset. More specifically, it projects data into a new orthonormal coordinates system computed to maximize the data variance on each of these new axes. PCA finds the sources of variability in the form of latent variables called principal components or loadings. The projection coefficients of each data on the principal components are called the scores and express the contribution of each principal component into each data. The first principal components express the principal sources of variation into the dataset and the visualization of their scores gives an easy way to analyze the dataset structure.

PCA has been widely used in vibrational spectroscopy due to its capacity to ease the visualization of multidimensional complex datasets.

In our study, PCA was used: i) in order to analyze the spectral differences induced by the experimental setup modifications between the two projects; and ii) to evaluate the capacity of the developed pre-processing protocols to homogenize the data set.

## 2.5. Feature selection and Partial Least Squares - Discriminant Analysis

In order to quantify the efficiency of the proposed pre-processing procedures, a supervised analysis was realized by computing sensitivity (B-CLL patients) and specificity (healthy patients) using a supervised processing composed of a feature selection and a supervised classification by Partial Least Squares - Discriminant Analysis (PLS-DA).

In this work, a canonical correlation analysis was applied to the data in order to achieve a supervised feature selection[17, 18]. In order further to decrease the number of selected variables, a supervised data dimension reduction using the cell labels (healthy or CLL) was developed and applied to the discriminant wavenumbers selected by the canonical correlation analysis. The first step consisted in the computation of the Fisher-score at each wavenumber. The second step computed the wavenumber correlation coefficient matrix  $\mathbf{R}_\lambda$  from the dataset. In the third step, the elements of  $\mathbf{R}_\lambda$  being smaller than a threshold fixed to 0.7 were removed. Each line of  $\mathbf{R}_\lambda$  was thus composed of highly correlated wavenumbers. In the fourth step, each line was reduced to its wavenumber having the highest Fisher-score computed in the first step, i.e. the most discriminant wavenumber. The last step computed the unique wavenumbers composing  $\mathbf{R}_\lambda$ , i.e. removed repeated wavenumbers. At the end of this very fast procedure, all the spectral information is summarized in its most representative and discriminant wavenumbers.

PLS-DA[19] is a linear supervised classification method used to model the covariance structure between data and labels by latent variables. This method has been proved particularly effective for multicollinear data composed of more features than observations, typically observed in Raman spectroscopy studies[20, 21]. In our project, PLS-DA was chosen for its simple mathematical model depending on a unique hyper parameter being the number of latent variables.

A PLS-DA classifier is optimized by a training step necessary to construct a predictive model, especially to select the optimal number of latent variables. To this end, data were randomly divided into two subsets, i.e. a training set composed of 70% of patients is used to optimize the classifier, while the 30% of remaining patients compose the external validation set used to

independently evaluate the predictive performance of this model in terms of sensitivity and specificity.

The training step is usually realized by a cross-validation strategy, such as leave-one out cross-validation. However, as a dataset must always be considered at the highest hierarchical level in order to properly evaluate classification models[22], data were considered at the patient level in our study. As a consequence, the popular leave-one-patient-out cross-validation strategy[20, 21] was used in our work. In each cross-validation cycle, the spectra corresponding to one patient of the training set are removed from the data set and used as the internal validation set. This process is repeated several times until each patient from the training set has been used once for the internal validation. Then, overall sensitivity and specificity in training are calculated. This procedure is repeated for a number of latent variables varying from 2 to 40. The optimal number of latent variables is the one giving the model with the highest sensitivity and specificity. Then, the entire training set is used to construct a new model with the selected number of latent variables.

The spectra acquired on the patients composing the external validation set are then injected into this final model. The performance of this classifier is evaluated in term of sensitivity and specificity estimated on this external validation set.

### **3. RESULTS AND DISCUSSION**

#### **3.1. Different experimental designs between IHMO and M3S projects**

At the end of the IHMO project, suggestions have been emitted in order to improve the quality of the collected Raman signal and to standardize the sample preparation protocol for clinical application. Experimental and instrumental modifications have thus been made during the M3S project. The switch from manual to automatic spreading has resulted in homogeneous smears and thicker cells. Replacing the IMHO 100X/0.90NA lens with a 150X/0.95NA lens in the M3S project increased confocality and signal-to-noise ratio and significantly reduced the contribution of glass to lymphocyte Raman spectra. Indeed, the large magnification and numerical aperture of the 150X lens enhance Raman scattering (excitation and collection) induced by a lower volume of analysis than the 100X lens.

As stated in previous studies[23, 24], such modifications impact the acquired Raman spectra. In the following, we will evaluate their influence on the Raman spectra acquired on lymphocyte nuclei.

### **3.2. Non-transferability of IHMO models on M3S data**

First, the data acquired during the IHMO project were pre-processed using pre-processing protocol 1 as described in the "Materials and methods" section. From these data, a feature selection was applied and a supervised PLS-DA classification model was developed. A sensitivity of 80% and a specificity of 84% were achieved in internal validation. In external validation, performance was 95% and 75% for sensitivity and specificity, respectively.

This pre-treatment protocol 1 and this trained PLS-DA model were blindly applied to the Raman data acquired on the 50 patients (26 healthy and 24 CLL) of the M3S measurement campaign. A sensitivity of 95% and a specificity of 27% were achieved, indicating the non-transferability of this supervised model to the M3S data. In order to investigate the origins of this limitative results, the raw data from the two campaigns were processed and compared in the next section.

In a previous study[13], data from the first campaign were also processed by Pre-processing 1 and Support Vector Machines (SVM), leading to an efficient classification model. Applied to the M3S dataset, this classifier failed in distinguishing healthy volunteers from B-CLL patients with similar results (data not shown). However, for the sake of clarity, we preferred to focus on a unique and simple supervised classification technique such as PLS-DA throughout this paper focused on the pre-processing step of raw data.

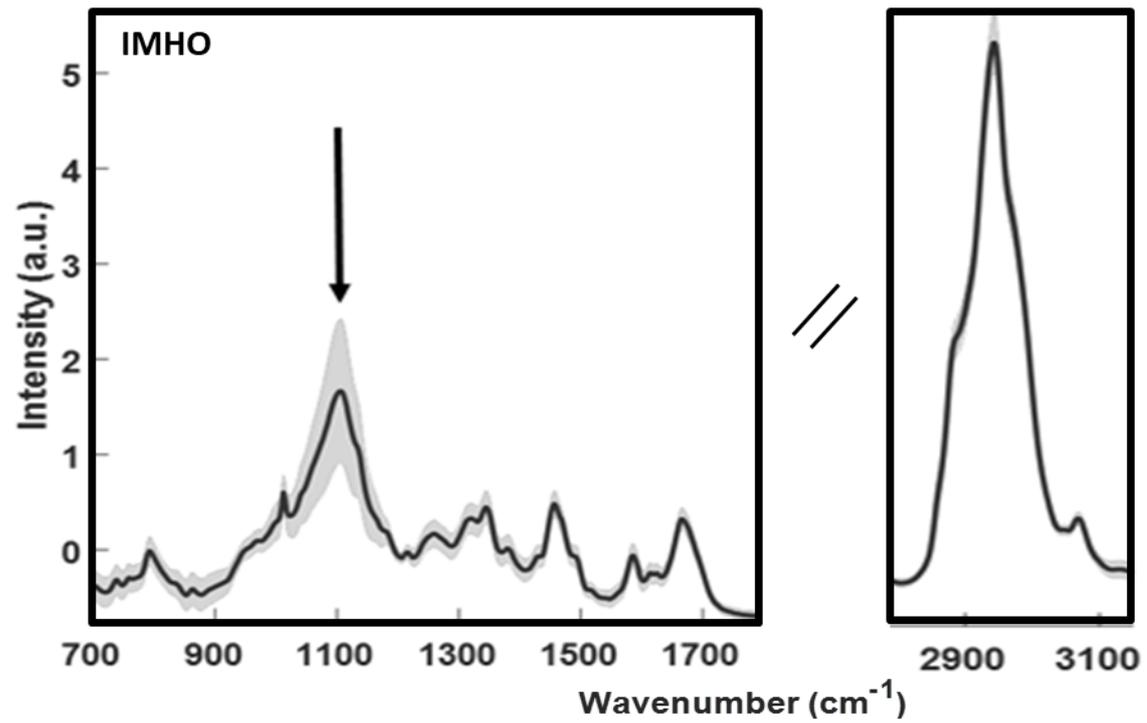
### **3.3. Highlighting the spectral differences between the two measurement campaigns**

The introduction of Raman spectroscopy into the clinic is severely hampered by the strong Raman signature of the glass slides that are routinely used in haematology department to produce blood smears. A key point of the IMHO and M3S measurements campaigns was the acquisition of Raman spectra on lymphocytes under clinical conditions, i.e. to work on fresh blood smears spread on glass slides. As part of the M3S measurement campaign, experimental optimizations were carried out. Indeed, the smear spreading method has been automated and modifications have been made at the microscope level to optimize the optical set-up.

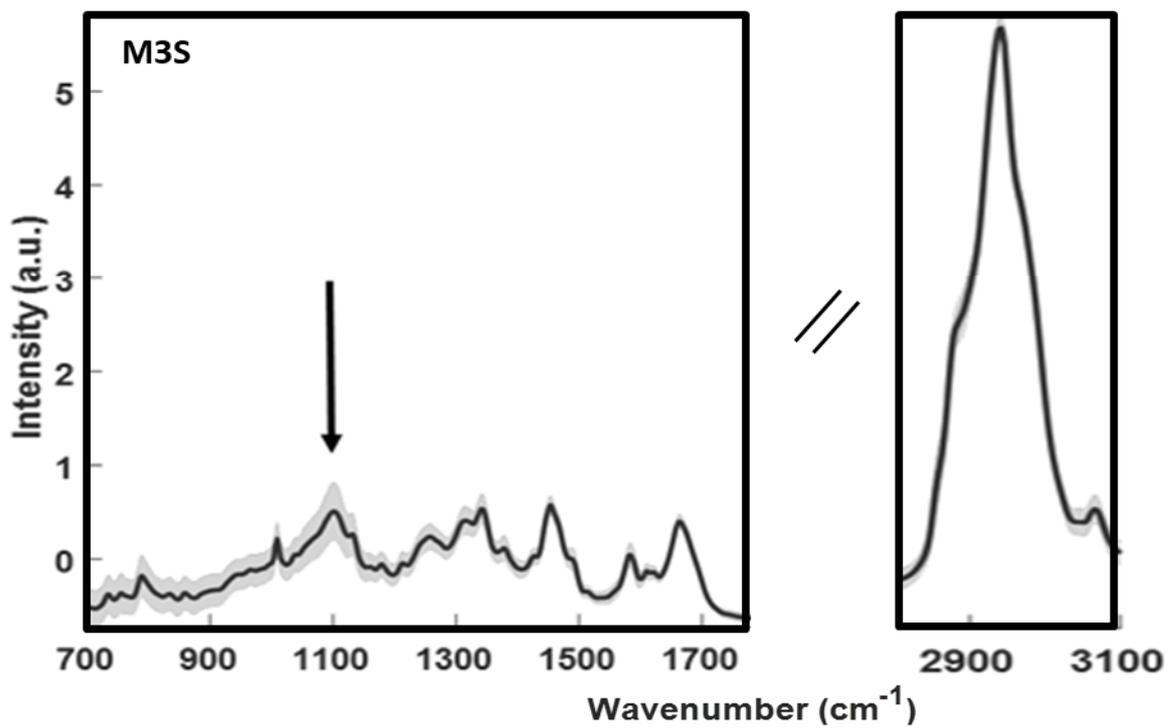
Figures 1a and 1b show the averages of the SNV normalized spectra acquired on lymphocytes[13] during the IHMO and M3S campaigns, respectively. The spectra baseline has been corrected using a polynomial function of order 2. Even reduced using a laser source with excitation at 532 nm [16, 25, 26], a significant contribution of glass, characterized by the 1050-1150  $\text{cm}^{-1}$  region, is still visible on these spectra (localized by a black arrow). A greater

contribution of glass can be observed for the IHMO spectra (Figure 1a) compared to the M3S spectra (Figure 1b) due to the confocality of the respective employed objectives.

These observations are confirmed by PCA[27] applied to explore the sources of variance into the IHMO and M3S datasets. As can be seen on the two first PCs (Figure 2a) the main source of variability is due to glass. The cumulated variance in the datasets achieved 99,5% with these two first PCs. Figure 2b shows the corresponding score plot with the 95% confidence ellipses of each campaign. It can be noticed that PC1 separates the data according to the measurement campaigns. Even if some similarities exist between the two datasets, visualized by the overlapping part of the score plot, they spread in opposite directions proving that the datasets contain spectral interferences limiting their comparison.

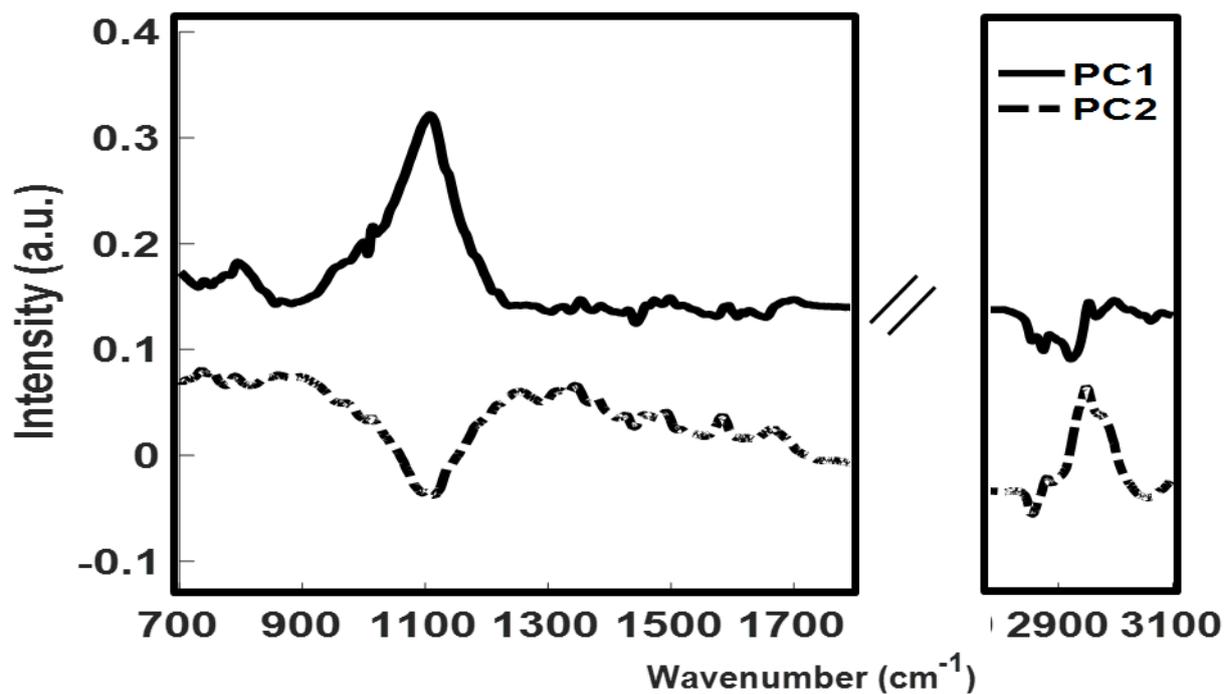


(a)

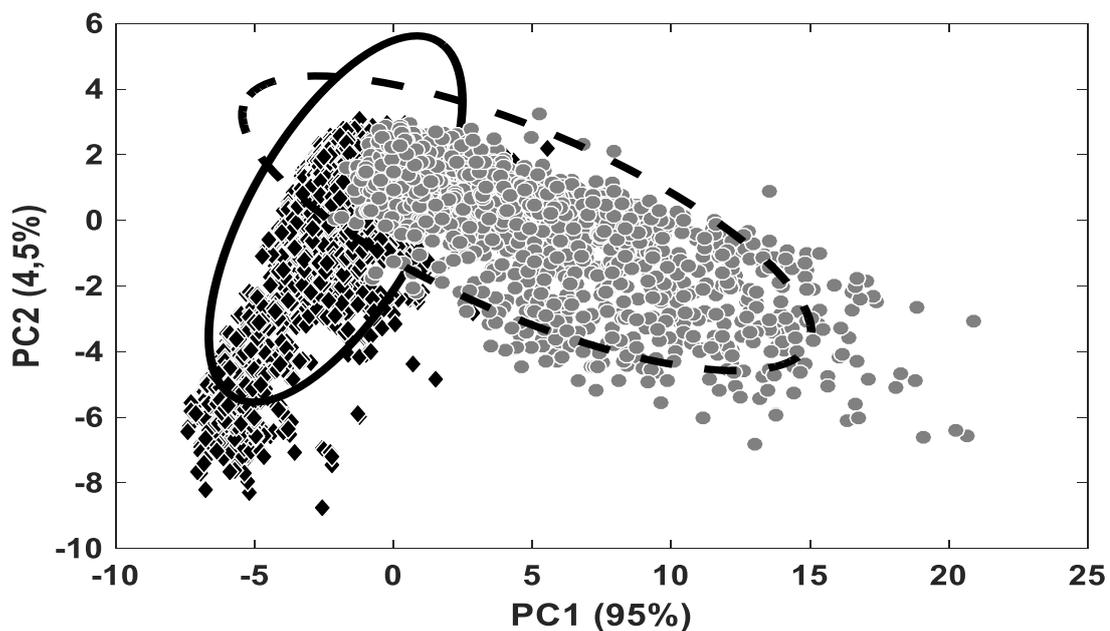


(b)

Figure 1- Comparison between IHMO and M3S spectral data. Mean (solid black line) and standard deviation (shaded areas) of the SNV normalized Raman spectra acquired on lymphocytes during the (a) IHMO and (b) M3S campaigns. The black arrows identify the spectral regions characteristic of the glass contribution.



(a)



(b)

Figure 2 – Results of PCA applied to the SNV normalized Raman spectra acquired during the IHMO and M3S measurement campaigns. (a) PC1 and PC2 loadings. (b) PC1 vs PC2 score plot of IHMO (black) and M3S (grey) spectral data. 95% confidence ellipse are represented in solid and dashed lines for IMHO and M3S datasets respectively.

These interferences can originate from changes not only in instrumental set-up but also in spreading procedure between the two campaigns. Various studies have already shown that different sample handling, preparation, storage or experimental changes can induce variations in Raman spectra[23, 24].

A possible solution to avoid glass contribution would have been to use substrates without Raman signal in the analyzed spectral range (700 to 3170  $\text{cm}^{-1}$ ), such as calcium fluoride ( $\text{CaF}_2$ ), quartz or barium fluoride ( $\text{BaF}_2$ ) as commonly used in a large majority of published studies[28]:[29-31]. However, these substrates are not practical for large-scale applications in clinical routine because of their cost. The use of glass slide was thus a requirement of IHMO and M3S project specifications, which were drawn up in the context of application in clinical settings.

#### **3.4. Incomplete correction by Pre-Processing 1**

During the IHMO project, the efficient Pre-processing 1 was developed in order to remove glass contribution[13], however the poor results presented in section 3.2 are suspected to be due to an instability of Pre-processing 1 to correct for variance due to experimental changes during the M3S project. The correction efficiency of Pre-processing 1 was therefore evaluated on the spectra of the two measurement campaigns.

As can be seen on Figures 3a and 3b, the spectral contribution of glass in the 1050-1150  $\text{cm}^{-1}$  region is neutralized within the datasets.

This observation is confirmed by the application of PCA on these data. Indeed, PC1 and PC2 presented on Figure 4a do not exhibit the characteristic glass band in the 1050-1150  $\text{cm}^{-1}$  region. However, as can be seen on the score plot considering PC1 and PC2 presented in Figure 4b, PC2 separates data according to the measurement campaigns. Despite a greater overlap of ellipses, the data remain separable due of the residual heterogeneity after Pre-processing 1.

These poor results may be due to the sequential design of Pre-processing 1, inducing an accumulation of estimation errors which amplify the small spectral differences existing between the two campaigns due to different instrumental and experimental conditions, and to its individual spectrum-based correction, which introduces an inter-spectrum variability.

These results demonstrate the inadequate correction of Pre-Processing 1, preventing the transferability of IHMO models on M3S. In order to overcome this problem, Pre-Processing 2 was developed.

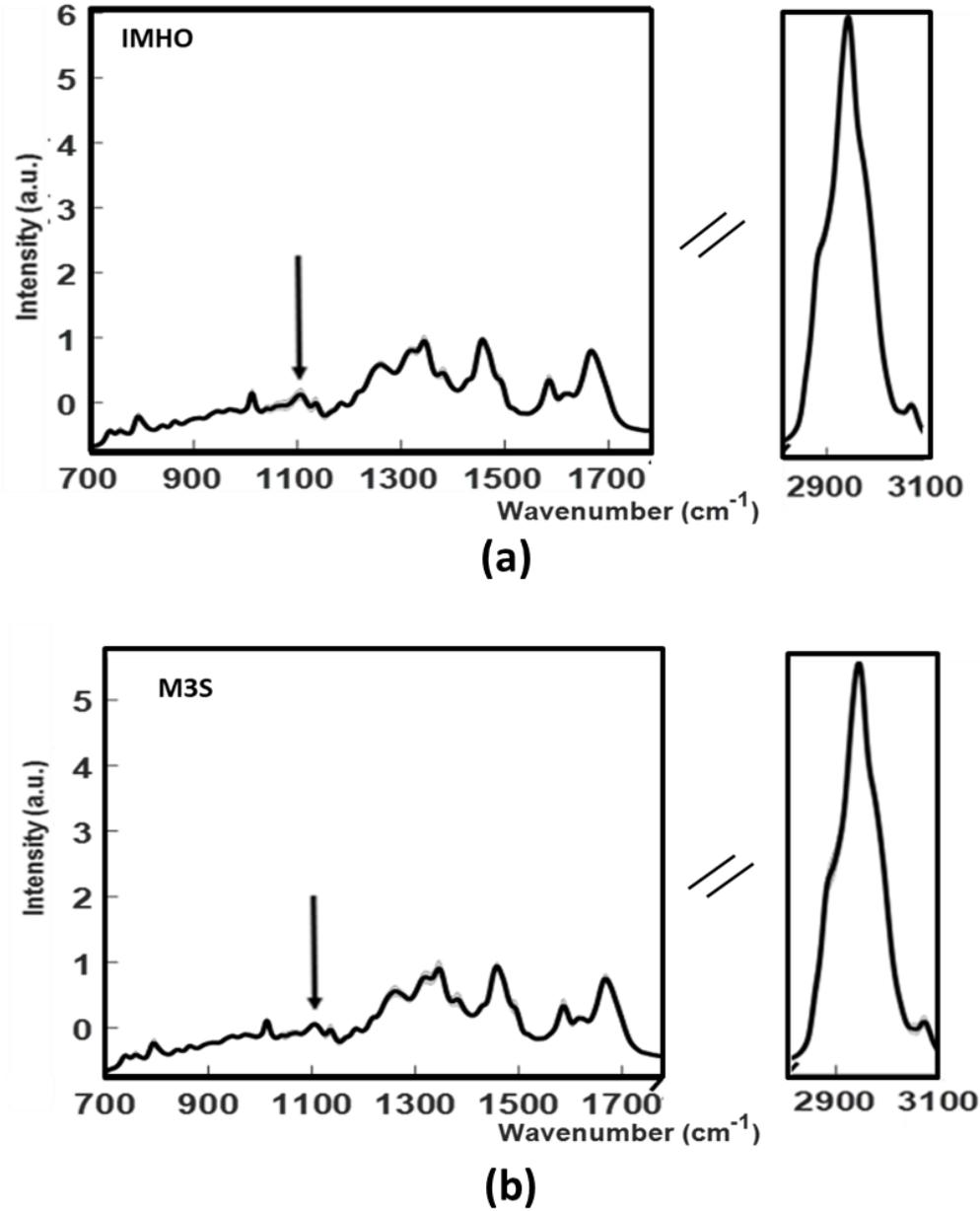
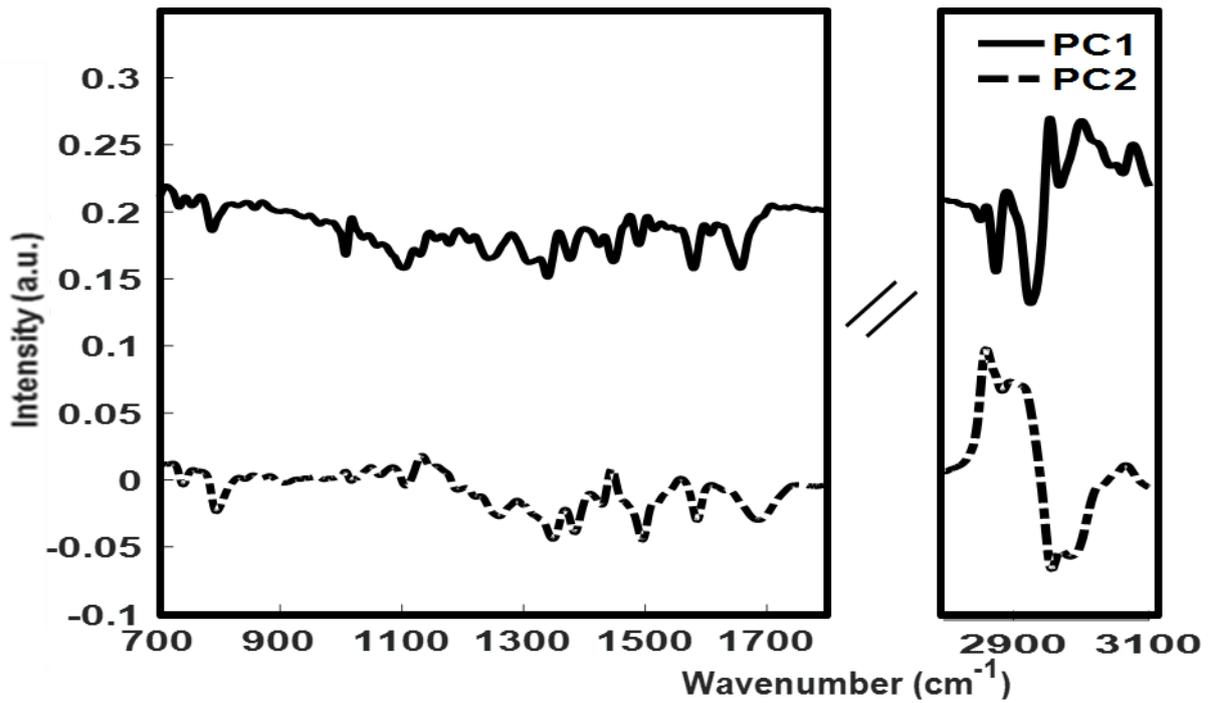
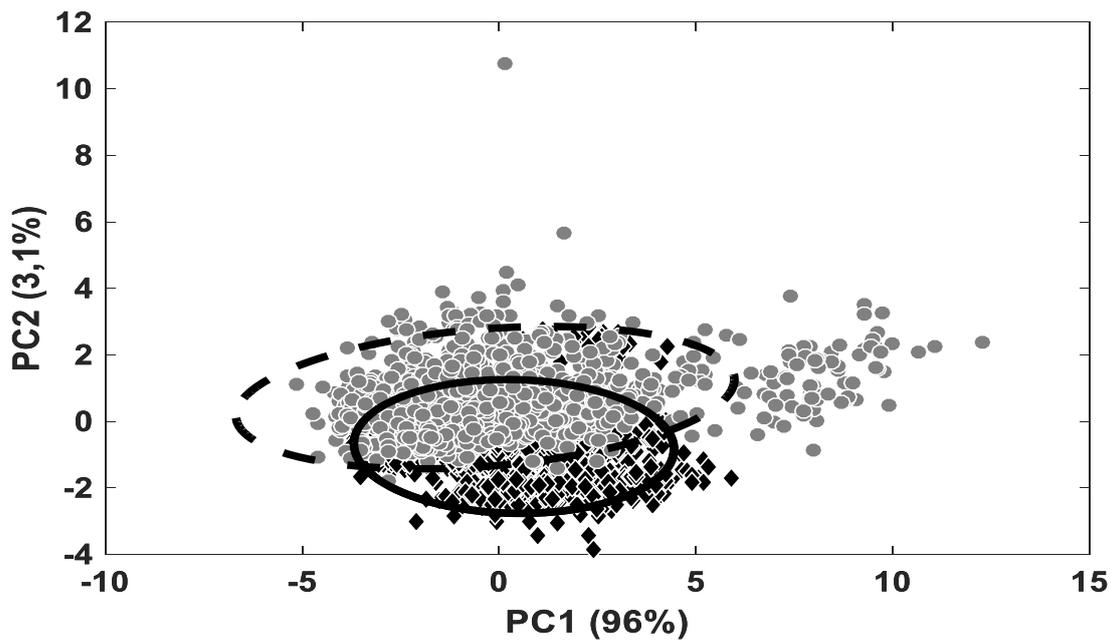


Figure 3 – Evaluation of Pre-processing 1. Mean (solid black line) and standard deviation (shaded areas) of pre-processed Raman spectra acquired on lymphocytes during the (a) IHMO and (b) M3S campaigns. The black arrows identify the spectral regions characteristic of the glass contribution.



(a)



(b)

Figure 4 – PCA applied on the data pre-processed by Pre-Processing 1. (a) PC1 and PC2 loadings. (b) PC1 vs PC2 score plot of IHMO (black) and M3S (grey) spectral data. 95% confidence ellipse are represented in solid and dashed lines for IHMO and M3S datasets respectively.

### 3.5. High performance of Pre-Processing 2

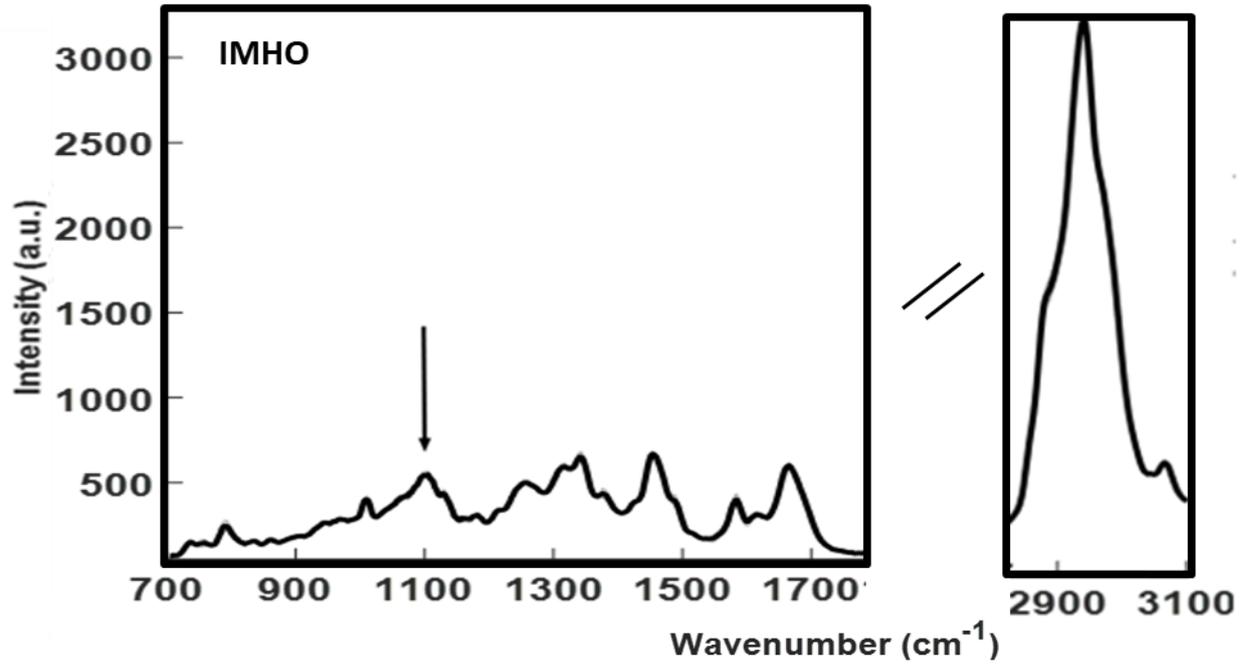
The spectra of the two campaigns were subjected to Pre-Processing 2. Contrary to Pre-Processing 1 which removes glass signal, Pre-processing 2 aims at neutralizing the variance of glass signal, while preserving its features, as can be seen on Figure 5. The neutralization of the glass signal variability is confirmed by the application of PCA on these data. Indeed, no characteristic glass band can be seen on PC1 and PC2 presented on Figure 6a.

The efficiency of Pre-Processing 2 to remove the spectral responses generated by the modification of experimental and instrumental conditions for the M3S campaign can be evaluated by the PCA score plot using the two first PCs (Figure 6b). It appears an overlapping between the IHMO and M3S datasets illustrated by mixed up and compact 95% confidence ellipses.

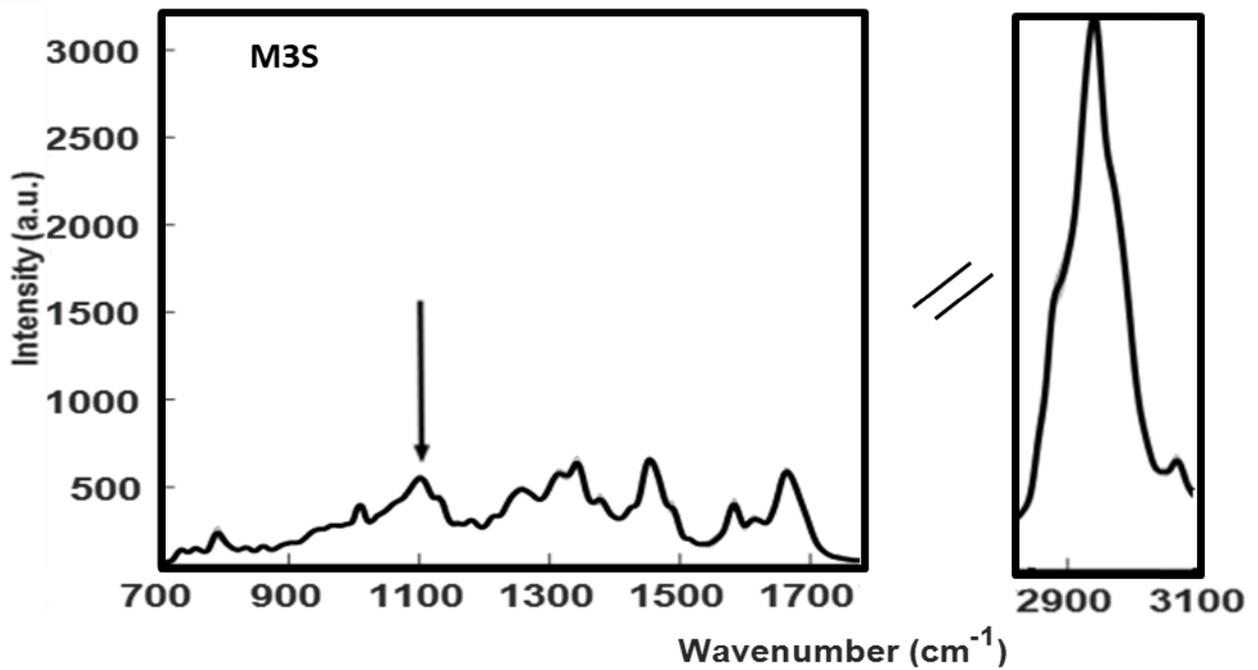
This observation can be explained by the properties of EMSC. First, EMSC is a data-driven method. Indeed, it uses the same reference spectrum computed from the entire dataset to guide the corrections of the spectra. Second, it is a model-based method, i.e. quantifying and separating simultaneously the different undesired variability sources in the spectra [12]. Because of these two properties, Pre-processing 2 treats all the data jointly from all the undesired effects at once using a batch algorithm, resulting in a minimization of the modelization error. Consequently, the data pre-processed by this method are known to be more homogeneous than with other iterative polynomial correction methods[16]. Furthermore, the fast execution speed of EMSC pre-treatment is another important advantage for our clinical application[16].

Taken together, these results are in accordance with previous studies demonstrating the efficiency of EMSC to pre-process Raman spectra acquired on biological samples deposited on aluminum and CaF<sub>2</sub> slides[32] and also to efficiently neutralize glass signal variability from Raman spectra acquired on cytological samples[16].

Further classification models will thus be more reliable since constructed exclusively on useful biomolecular information of interest. Indeed, the choice of appropriate pre-processing steps is crucial for the construction of an efficient chemometric chain that includes the training of a supervised classification model[33].



(a)



(b)

Figure 5 – Evaluation of Pre-processing 2. Mean (solid black line) and standard deviation (shaded areas) of pre-processed Raman spectra acquired on lymphocytes during the (a) IHMO and (b) M3S campaigns. The black arrows identify the spectral regions characteristic of the glass contribution.

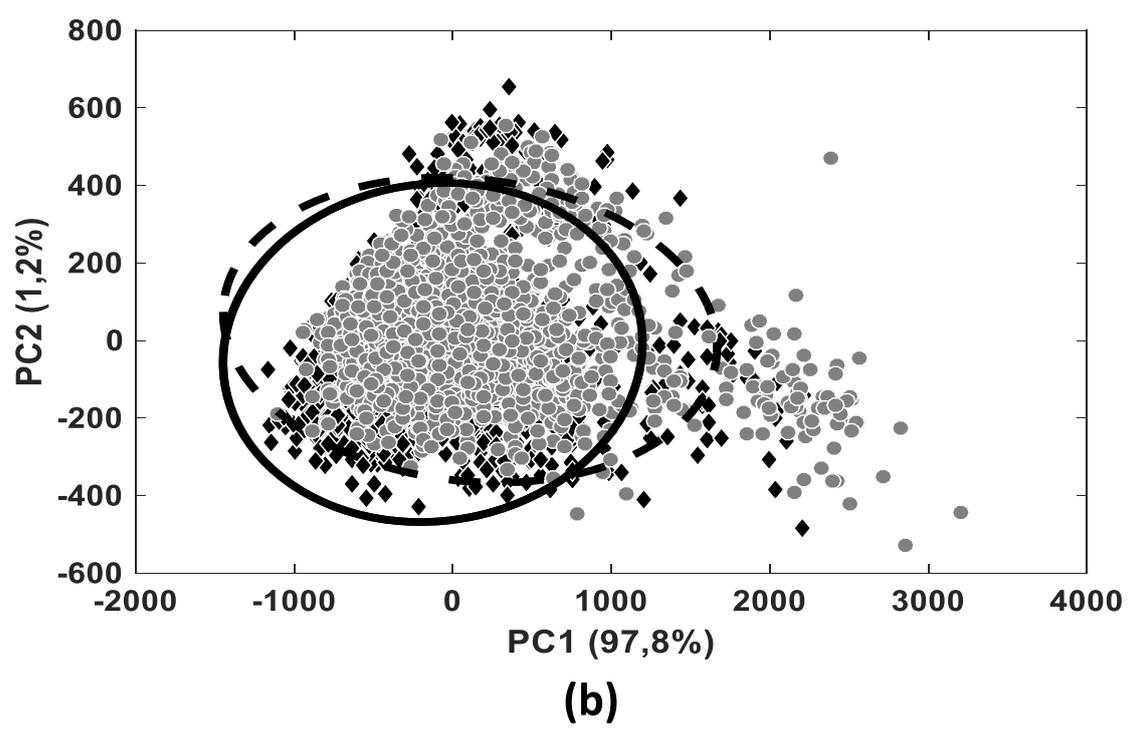
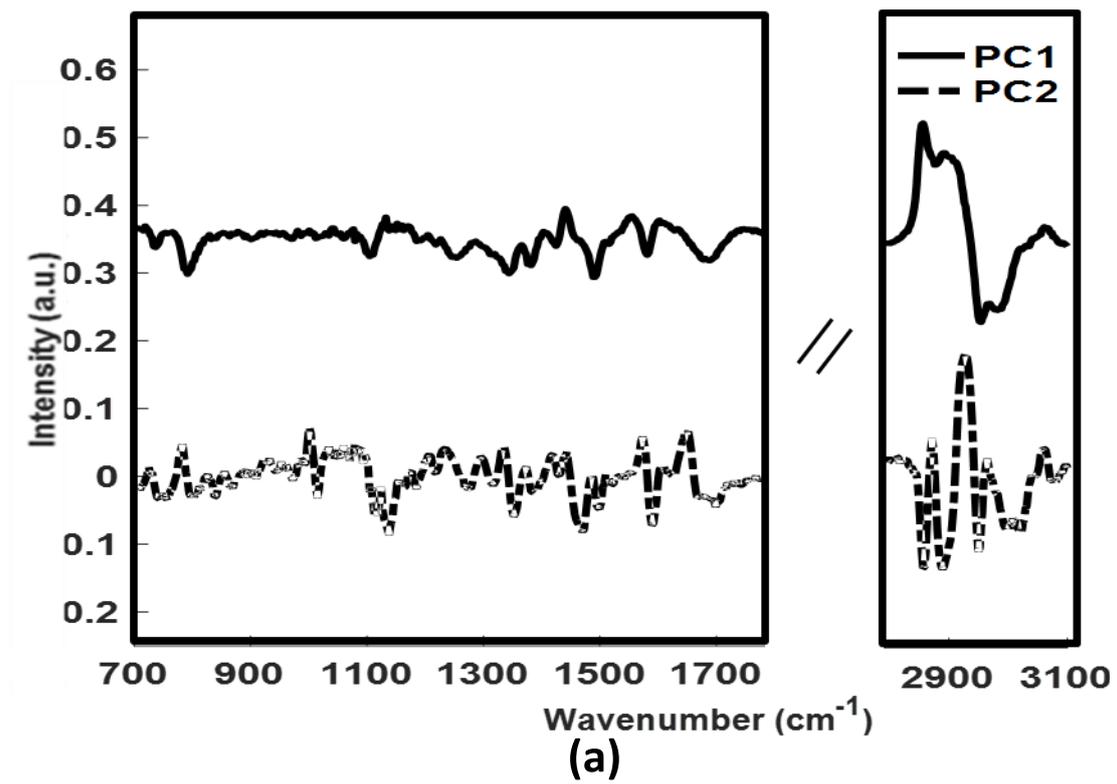


Figure 6 – PCA applied on the data pre-processed by Pre-Processing 2. (a) PC1 and PC2 loadings. (b) PC1 vs PC2 score plot of IHMO (black) and M3S (grey) spectral data. 95% confidence ellipses are represented in solid and dashed lines for IHMO and M3S datasets respectively.

### **3.6. Efficient pre-processing is a necessary but not sufficient condition to construct transferable models**

Previously, it was shown by PCA that Pre-processing 2 permitted to correct and homogenize Raman data from two measurement campaigns with different experimental conditions.

To support the previous results and assess if the pre-treatment strategy alone is sufficient to achieve transferable models, a predictive model was constructed on the IHMO dataset (pre-processed by Pre-processing 2) by PLS-DA coupled with feature selection and leave-one-patient-out cross validation as described in section 2.6. Table 2 presents the performance of this PLS-DA model for the classification of healthy and CLL patients in internal and external validation, and in blind test on M3S dataset. Although good performance is obtained for internal validation, this model lacks accuracy in external validation, and results fall when blindly tested on the M3S dataset.

The same experiment was realized by reversing IHMO and M3S datasets, i.e. the model was constructed on the M3S data and blindly tested on the IHMO data. Similar results were obtained as presented in Table 2, with the same similar loss of performance from the internal validation to the external validation on the M3S data and the blind test on the IHMO dataset.

These results demonstrated that pre-treatment alone is not sufficient to build a high-performance model despite effective correction and homogenization of the two given sets. The models being trained on a dataset acquired during a unique measure campaign, a possible interpretation of these results is that these models are specific to a particular experimental condition and fail when they predict the state of data acquired with a different experimental condition. However, whatever the model, the performance decrease is similar between the external validation set and the blind set. This argument can thus be rejected. The low number of patients composing the training set is a more pertinent argument, which will be addressed in the next section.

*Table 2 - Performance comparison of PLS-DA in internal and external validation and blind test sets for models trained exclusively on IMHO or M3S datasets pre-processed by Pre-processing 2.*

	<b>Sensitivity</b>	<b>Specificity</b>
<b>Model 1: Training with IMHO dataset</b>		
Internal validation on IMHO dataset	83%	94%
External validation on IMHO dataset	65%	75%
Test on M3S data set	82%	59%
<b>Model 2: Training with M3S dataset</b>		
Internal validation on M3S dataset	80%	87%
External validation on M3S dataset	73%	71%
Test on IMHO data set	81%	65%

### **3.7. Efficient pre-processing and high number of patients are necessary and sufficient conditions to construct transferable models**

In order to increase the number of patients, the data sets from the two measurement campaigns were used to construct a prediction model. Table 3 presents the performance of PLS-DA in classifying healthy and B-CLL patients after mixing M3S and IMHO data. The results show that PLS-DA model constructed using the data treated by Pre-processing 1 has a high prediction error both during training and validation steps. Better performance was obtained on the data transformed using Pre-processing 2.

Of course, the predictive results presented in Table 3 may be improved by optimizing the supervised classification strategy. However, this is the objective of a forthcoming article describing in detail how the choices made at each step of the supervised classification strategy impact the discrimination between Raman spectra acquired on lymphocytes of B-CLL and healthy patients.

In this study, PLS-DA was chosen because it is an efficient technique [10, 34, 35] commonly used in spectroscopy. Furthermore, it is a method easier to optimize than other classic supervised classification methods, such as SVM or Random Forest, since PLS-DA depends on only one parameter, which is the number of latent variables.

*Table 3 - Performance comparison of PLS-DA in internal and external validation for models trained using simultaneously the IMHO and M3S datasets, in function of the pre-processing protocol.*

	<b>Sensitivity</b>	<b>Specificity</b>
<b>Pre-processing 1</b>		
Internal validation	79%	63%
External validation	80%	58%
<b>Pre-processing 2</b>		
Internal validation	86%	83%
External validation	84%	81%

#### **4. CONCLUSION**

This article demonstrates that changes in experimental and instrumental conditions have a direct impact on the Raman signals acquired on healthy and B-CLL lymphocytes. It is therefore important to correctly choose a pre-processing method to homogenize the data sets and construct relevant transferable models. Two different pre-processing strategies were studied: a step-wise one working on each spectrum individually, and a parallel and joined one based on EMSC. PCA and PLS-DA analyses suggest that the EMSC based pre-processing is effective in homogenizing Raman spectra acquired with different experimental and instrumental conditions. However, pre-processing alone is not sufficient to build high-performance models. It is essential to have a high number of patients included in the study in order to improve classification results.

## **AUTHOR INFORMATION**

Corresponding Author

\* Piot Olivier, BioSpecT EA n°7506, Faculty of Pharmacy, University of Reims Champagne-Ardenne, 51 rue Cognacq-Jay, 51096 Reims Cedex, France. Email address [olivier.piot@univ-reims.fr](mailto:olivier.piot@univ-reims.fr)

## **Acknowledgements**

The authors would like to acknowledge the Agence Nationale de la Recherche (ANR) and the European Community for financial support.

- [1] M. Hallek, B.D. Cheson, D. Catovsky, F. Caligaris-Cappio, G. Dighiero, H. Dohner, P. Hillmen, M.J. Keating, E. Montserrat, K.R. Rai, Guidelines for the diagnosis and treatment of chronic lymphocytic leukemia: a report from the International Workshop on Chronic Lymphocytic Leukemia updating the National Cancer Institutes Working Group 1996 guidelines, *Blood*, 111 (2008) 5446-5456.
- [2] L. Zhang, M.A. Freitas, J. Wickham, M.R. Parthun, M.I. Klisovic, G. Marcucci, J.C. Byrd, Differential expression of histone post-translational modifications in acute myeloid and chronic lymphocytic leukemia determined by high-pressure liquid chromatography and mass spectrometry, *Journal of the American Society for Mass Spectrometry*, 15 (2004) 77-86.
- [3] T. Ichimura, K. Fujita, S. Kawata, T.M. Watanabe, T. Yanagida, H. Fujita, Visualizing cell state transition using Raman spectroscopy, *PLoS One*, 9 (2014) e84478.
- [4] H.G. Schulze, S.O. Konorov, N.J. Caron, J.M. Piret, M.W. Blades, R.F.B. Turner, Assessing differentiation status of human embryonic stem cells noninvasively using Raman microspectroscopy, *Analytical chemistry*, 82 (2010) 5020-5027.
- [5] J.W. Chan, D.K. Lieu, T. Huser, R.A. Li, Label-free separation of human embryonic stem cells and their cardiac derivatives using Raman spectroscopy, *Analytical chemistry*, 81 (2009) 1324-1331.
- [6] M. Poplineau, A. Trussardi-Régnier, T. Happillon, J. Dufer, M. Manfait, P. Bernard, O. Piot, F. Antonicelli, Raman microspectroscopy detects epigenetic modifications in living Jurkat leukemic cells, *Epigenomics*, 3 (2011) 785-794.
- [7] T. Ichimura, K. Fujita, H. Machiyama, T. Yamaguchi, T.M. Watanabe, H. Fujita, Non-label immune cell state prediction using Raman spectroscopy, *Scientific reports*, 6 (2016) 37562.
- [8] M. Chen, N. McReynolds, E.C. Campbell, M. Mazilu, J.o. Barbosa, K. Dholakia, S.J. Powis, The use of wavelength modulated Raman spectroscopy in label-free identification of T lymphocyte subsets, natural killer cells and dendritic cells, *PLoS One*, 10 (2015) e0125158.
- [9] A. Weselucha-Birczynska, M. Koziński, J. Czepiel, M. Birczynska, Raman microspectroscopy tracing human lymphocyte activation, *Analyst*, 138 (2013) 7157-7163.
- [10] A.J. Hobro, Y. Kumagai, S. Akira, N.I. Smith, Raman spectroscopy as a tool for label-free lymphocyte cell line discrimination, *Analyst*, 141 (2016) 3756-3764.
- [11] H.J. Byrne, M. Baranska, G.J. Puppels, N. Stone, B. Wood, K.M. Gough, P. Lasch, P. Heraud, J. Sulé-Suso, G.D. Sockalingum, Spectroscopy for the next generation: Quo vadis?, *Analyst*, 140 (2015) 2066-2073.
- [12] N.K. Afseth, A. Kohler, Extended multiplicative signal correction in vibrational spectroscopy, a tutorial, *Chemometrics and Intelligent Laboratory Systems*, 117 (2012) 92-99.
- [13] T. Happillon, V. Untereiner, A. Beljebbar, C. Gobinet, S. Daliphard, P. Cornillet-Lefebvre, A. Quinquenel, A. Delmer, X. Troussard, J. Klossa, Diagnosis approach of chronic lymphocytic leukemia on unstained blood smears using Raman microspectroscopy and supervised classification, *Analyst*, 140 (2015) 4465-4472.
- [14] A. Savitzky, M.J.E. Golay, Smoothing and differentiation of data by simplified least squares procedures, *Analytical chemistry*, 36 (1964) 1627-1639.
- [15] R.J. Barnes, M.S. Dhanoa, S.J. Lister, Standard normal variate transformation and detrending of near-infrared diffuse reflectance spectra, *Applied Spectroscopy*, 43 (1989) 772-777.
- [16] L.T. Kerr, B.M. Hennelly, A multivariate statistical investigation of background subtraction algorithms for Raman spectra of cytology samples recorded on glass slides, *Chemometrics and Intelligent Laboratory Systems*, 158 (2016) 61-68.
- [17] D.R. Hardoon, S. Szedmak, J. Shawe-Taylor, Canonical correlation analysis: An overview with application to learning methods, *Neural computation*, 16 (2004) 2639-2664.
- [18] H. Abdi, V. Guillemot, A. Eslami, D. Beaton, Canonical correlation analysis, *Encyclopedia of Social Network Analysis and Mining*, (2017) 1-16.

- [19] M. Barker, W. Rayens, Partial least squares for discrimination, *Journal of Chemometrics*, 17 (2003) 166-173.
- [20] V. Gaydou, M. Polette, C. Gobinet, C. Kileztky, J.-F. Angiboust, M. Manfait, P. Birembaut, O. Piot, Vibrational Analysis of Lung Tumor Cell Lines: Implementation of an Invasiveness Scale Based on the Cell Infrared Signatures, *Analytical chemistry*, 88 (2016) 8459-8467.
- [21] A. Palermo, M. Fosca, G. Tabacco, F. Marini, V. Graziani, M.C. Santarsia, F. Longo, A. Lauria, R. Cesareo, I. Giovannoni, Raman Spectroscopy Applied to Parathyroid Tissues: A New Diagnostic Tool to Discriminate Normal Tissue from Adenoma, *Analytical chemistry*, (2017).
- [22] S. Guo, T. Bocklitz, U. Neugebauer, J.r. Popp, Common mistakes in cross-validating classification models, *Anal. Methods*, (2017).
- [23] D.S. Read, A.S. Whiteley, Chemical fixation methods for Raman spectroscopy-based analysis of bacteria, *Journal of microbiological methods*, 109 (2015) 79-83.
- [24] K. Mlynáriková, O. Samek, S. Bernatová, F. Růžička, J. Ježek, A. Hároniková, M. Šiler, P. Zemánek, V. Holá, Influence of culture media on microbial fingerprints using Raman spectroscopy, *Sensors*, 15 (2015) 29635-29647.
- [25] H.J. Butler, L. Ashton, B. Bird, G. Cinque, K. Curtis, J. Dorney, K. Esmonde-White, N.J. Fullwood, B. Gardner, P.L. Martin-Hirsch, Using Raman spectroscopy to characterize biological materials, *Nature protocols*, 11 664-687.
- [26] L. Mikoliunaite, R.D. Rodriguez, E. Sheremet, V. Kolchuzhin, J. Mehner, A. Ramanavicius, D.R.T. Zahn, The substrate matters in the Raman spectroscopy analysis of cells, *Scientific reports*, 5 (2015).
- [27] H. Hotelling, Analysis of a complex of statistical variables into principal components, *Journal of educational psychology*, 24 (1933) 417.
- [28] R. Vanna, P. Ronchi, A.T.M. Lenferink, C. Tresoldi, C. Morasso, D. Mehn, M. Bedoni, S. Picciolini, L. Terstappen, F. Ciceri, Label-free imaging and identification of typical cells of acute myeloid leukaemia and myelodysplastic syndrome by Raman microspectroscopy, *Analyst*, 140 (2015) 1054-1064.
- [29] F. Draux, P. Jeannesson, A. Beljebbar, A. Tfayli, N. Fourre, M. Manfait, J. Sulé-Suso, G.D. Sockalingum, Raman spectral imaging of single living cancer cells: a preliminary study, *Analyst*, 134 (2009) 542-548.
- [30] A. Ramoji, U. Neugebauer, T. Bocklitz, M. Foerster, M. Kiehntopf, M. Bauer, J.r. Popp, Toward a spectroscopic hemogram: Raman spectroscopic differentiation of the two most abundant leukocytes from peripheral blood, *Analytical chemistry*, 84 (2012) 5335-5342.
- [31] A.D. Meade, F.M. Lyng, P. Knief, H.J. Byrne, Growth substrate induced functional changes elucidated by FTIR and Raman spectroscopy in in vitro cultured human keratinocytes, *Analytical and bioanalytical chemistry*, 387 (2007) 1717-1728.
- [32] K.H. Liland, A. Kohler, N.K. Afseth, Model-based pre-processing in Raman spectroscopy of biological samples, *J. Raman Spectrosc*, 47 (2016) 643-650.
- [33] T. Bocklitz, A. Walter, K. Hartmann, P. Rösch, J. Popp, How to pre-process Raman spectra for reliable and stable models?, *Analytica Chimica Acta*, 704 (2011) 47-56.
- [34] G. McLaughlin, K.C. Doty, I.K. Lednev, Discrimination of human and animal blood traces via Raman spectroscopy, *Forensic science international*, 238 (2014) 91-95.
- [35] W. Liu, H. Wang, J. Du, C. Jing, Raman microspectroscopy of nucleus and cytoplasm for human colon cancer diagnosis, *Biosensors and Bioelectronics*, 97 (2017) 70-74.

**Article #2 : Mise en œuvre d'une stratégie de classification de données Raman collectées dans différentes conditions cliniques : application au diagnostic de la leucémie lymphocytaire chronique.**

---

**Implementation of a classification strategy of Raman data collected in different clinical conditions: application to the diagnosis of chronic lymphocytic leukemia**

**M. Féré, C. Gobinet, L.H Liu, A. Beljebbar, V. Untereiner, D. Gheldof, M. Chollat, J. Klossa, B. Chatelain, O. Piot**

Article soumis dans le journal Elsevier : Analytical Chemistry.

---

## **Préambule à l'article 2 :**

### **Contexte**

Nous avons vu dans l'article précédent que le prétraitement des données Raman était une étape primordiale afin de rendre des données multicentriques homogènes et analysables. Au vu du nombre croissant des données et de leurs complexités, les techniques de chimiométrie et de machine Learning sont devenues des outils cruciaux pour extraire l'information pertinente. Le principe de ces outils est de découvrir de manière autonome des corrélations significatives dans un jeu de donnée entre des cibles (labels) et des variables (les spectres Raman). Ces corrélations permettent de créer un modèle mathématique (un classifieur) et de prédire un événement inconnu à partir de nouvelles données. Mais pour construire des modèles de prédictions « robustes » et performants à des fins diagnostiques, il est important de considérer certains critères comme le choix de la stratégie d'apprentissage, l'optimisation et stabilité de ces modèles. Dans le cadre de la spectroscopie, ces critères manquent cruellement de consensus et sont encore aujourd'hui mal évalués. En spectroscopie Raman, la performance des modèles de prédictions réside donc à la fois dans la qualité et la richesse des données spectrales, mais également dans les outils statistiques utilisés pour les analyser.

### **Objectif**

L'objectif de cette partie a été d'évaluer différentes stratégies d'entraînement, de validation et d'optimisation afin de construire des modèles de prédiction utilisables pour des finalités cliniques. Dans cet article, divers aspects sont discutés, comme les méthodes de validations croisées, les méthodes de modélisation statistiques pour construire les modèles de prédiction et le choix de seuil de décision en fonction des besoins du clinicien. Nous avons proposé une stratégie de classification basée sur la méthode de prédiction ensembliste adaptative, c'est-à-dire utilisant la combinaison prédictive de plusieurs modèles optimisés afin de créer un seul classifieur. L'ensemble des modèles a été entraîné avec le même algorithme de classification. Nous avons établi une stratégie d'entraînement et de validation permettant de créer des modèles stables et insensibles au sur-apprentissage, et adaptés aux exigences cliniques.

## Matériels et méthodes

Dans le cadre de cette étude, des données Raman de 61 patients sains et de 79 patients atteints de LLC ont été utilisées. Ces données proviennent de deux campagnes de mesures acquies dans des conditions d'acquisition différentes. Pour chaque patient, un frottis a été préparé directement à l'hôpital de recrutement dans des conditions cliniques classiques. La première campagne comprenait 25 patients sains et 45 patients atteints de LLC, recrutés au centre hospitalier de Reims Champagne-Ardenne (RCA-HC). La deuxième campagne comprenait 36 patients sains et 34 patients LLC recrutés au centre hospitalier du Mont Godinne-Namur (MGN-HC). Pour la campagne IHMO, un spectre Raman a été acquis sur le noyau de 2 596 cellules saines et 4 257 cellules LLC avec un temps d'acquisition fixé à deux accumulations de 10 secondes, correspondant à 6 853 spectres Raman acquis. Pour le projet M3S, trois spectres Raman ont été acquis en trois positions différentes sur le noyau de 1804 cellules saines et 3128 cellules LLC, correspondant à 14796 spectres Raman, chacun acquis pendant une accumulation de 10 secondes. Pour les deux campagnes de mesures environ 100 cellules ont été analysées par patients.

Les données Raman ont subi un test qualité afin d'éliminer les spectres aberrants et de mauvaise qualité. Ces données ont ensuite subi un prétraitement par EMSC afin d'homogénéiser les deux campagnes de mesure. La création des modèles de prédiction a été réalisée à l'aide de trois algorithmes de classification : Partial Least Squares - Discriminant Analysis (PLS-DA), Support Vector Machine (SVM) and Random Forest (RF). Afin de choisir les meilleurs paramètres pour ces algorithmes et d'optimiser les classifieurs, nous avons utilisé l'algorithme de double cross-validation répétée (rdCV). La rdCV est une stratégie de rééchantillonnage afin de développer et d'optimiser des modèles de classification. Elle est composée de deux boucles imbriquées : une boucle interne permettant de trouver les paramètres optimaux pour les algorithmes de classification et une boucle externe permettant de construire plusieurs modèles à partir de différents jeux d'entraînement sélectionnés aléatoirement et des paramètres choisis dans la boucle interne. Ainsi 100 modèles ont été construits et testés sur un jeu de test indépendant. Chaque modèle optimisé prédit l'état physiopathologique (LLC ou sain) de chaque patient testé et la décision finale est obtenue par un vote majoritaire de la prédiction des 100 modèles.

## Résultats

L'ensemble des résultats n'a montré aucune différence significative au niveau des performances entre les différentes méthodes de validation croisée testées dans la boucle interne de rdCV. Les trois algorithmes supervisés (SVM, RF et PLS-DA) ont montré également des performances moyennes similaires lorsqu'elles sont appliquées 100 fois sur les données Raman. La combinaison entre la conjugaison des résultats de prédiction de plusieurs classifieurs (prédiction ensembliste), l'application de seuils adaptatifs de décision pour chacun des modèles et leurs optimisations, ont conduit à une sensibilité de 95% et une spécificité de 85%, soit une augmentation de 14% pour la sensibilité et de 8% pour la spécificité par rapport à la moyenne des modèles uniques.

## Conclusion

La combinaison de la spectroscopie Raman avec un programme de prétraitement efficace et une solide analyse statistique, a permis d'établir un diagnostic précis de la LLC, basé sur une prédiction ensembliste de 100 modèles indépendants et optimisés. Dans notre approche, nous avons essayé de solutionner les problèmes d'apprentissage, de validation et d'optimisation en travaillant sur plusieurs points tel que les problèmes de sur-apprentissage, du choix de la stratégie de validation, de la stabilité des modèles et du choix du seuil de décision. Il a été montré que la méthode par vote majoritaire est beaucoup plus efficace que la stratégie des modèles uniques. Comme cette stratégie comporte des seuils adaptables au contexte clinique, il a été possible de montrer que le diagnostic pouvait être orienter selon les besoins du clinicien, en favorisant soit la sensibilité, soit la spécificité.

# Implementation of a classification strategy of Raman data collected in different clinical conditions: application to the diagnosis of chronic lymphocytic leukemia

M. Féré<sup>1</sup>, C. Gobinet<sup>1</sup>, L.H Liu<sup>1</sup>, A. Beljebbar<sup>1</sup>, V. Untereiner<sup>2</sup>, D. Gheldof<sup>3</sup>, M. Chollat<sup>4</sup>, J. Klossa<sup>4</sup>, B. Chatelain<sup>3</sup>, O. Piot<sup>1,2</sup>

<sup>1</sup>University of Reims Champagne-Ardenne, BioSpecT EA 7506, Faculty of Pharmacy, Reims, France

<sup>2</sup>Cellular and Tissular Imaging Platform PICT, Faculty of Pharmacy, University of Reims Champagne-Ardenne, Reims, France

<sup>3</sup>Catholic University of Louvain, CHU UCL Namur, Namur Thrombosis and Hemostasis Center, Hematology Laboratory, Yvoir, Belgium

<sup>4</sup>TRIBVN, Châtillon, France

---

**ABSTRACT:** The literature is rich in proof-of-concept studies demonstrating the potential of Raman spectroscopy for disease diagnosis. However, few studies are conducted in a clinical context to demonstrate its applicability in current clinical practice and workflow. Indeed, this translational research remains far from the patient's bedside for several reasons. First, samples are often cultured cell lines. Second, they are prepared on non-standard substrates for clinical routine. Third, a unique supervised classification model is usually constructed using inadequate cross-validation strategy. Finally, the implemented models maximize classification accuracy without taking into account the clinician's needs. In this paper, we address these issues through a diagnosis problem in real clinical conditions, i.e. the diagnosis of chronic lymphocytic leukemia from fresh unstained blood smears spread on glass slides. From Raman data acquired in different experimental conditions, a repeated double cross validation strategy was combined with different cross-validation approaches, a consensus label strategy and adaptive thresholds able to adapt to the clinician's needs. Combined with validation at the patient level, classification results were improved compared to traditional strategies.

---

## INTRODUCTION

Raman spectroscopy is a label-free biophotonic technique, applicable to the analysis of complex biological samples, such as biofluids, cells and tissues. Raman spectroscopy allows accessing the global molecular composition of the biological samples, without affecting their integrity. With this technology, it is possible to detect biochemical changes caused by various diseases such as cancer<sup>1-3</sup> or metabolic pathologies<sup>4-7</sup>.

The literature is rich of proof of concept studies demonstrating that Raman spectroscopy has a high potential to improve disease diagnosis<sup>3,8-11</sup> or predict its progression<sup>12,13</sup> for a better personalization of patient care. However, beyond the proofs of concept, few studies are realized in a clinical context in order to demonstrate the applicability of Raman spectroscopy in current clinical practice and workflow. Indeed, this translational research remains far from the patient's bedside for several reasons.

For example, samples are often cultured cell lines<sup>14-16</sup> or prepared on non-standard substrates<sup>17,18-20</sup>.

Furthermore, concerning the construction of prediction models, the literature is highly heterogeneous on the choices of the training-validation strategy and of the supervised classification algorithm<sup>8,10,21-24</sup>. Inappropriate choices can lead to overfitting, i.e. models with poor generalization properties on unknown data.

Moreover, in most studies, a single model is built. However, the performance of this model is depending on the selected training and validation sets, and can thus highly fluctuate in function of these sets.

Finally, the implemented models are generally static and cannot be adapted to the clinician's needs, i.e., either to prioritize sensitivity or specificity, or to have balanced sensitivity and specificity.

In our approach, we have tried to address these different points by establishing a training, validation and optimization strategy to create stable and not subject to over-fitting prediction models, which are able to adapt to clinical demands. For this, we worked in the context of chronic lymphocytic leukaemia (CLL). Raman data were collected under different experimental conditions on unstained blood smears, spread manually or automatically on standard glass slides, during two different measurement campaigns. In our study, we used an innovative approach based on the strategy of repeated double cross validation (rdCV) combined with validation at the patient level, in order to limit overfitting as much as possible. We tested also different cross-validation approaches (leave-one-patient-out cross-validation<sup>25</sup>, K-fold cross-validation<sup>26</sup> (KFCV) and Monte-Carlo cross-validation<sup>27</sup> (MCCV)) to evaluate their effects on classification performance. For the problem of model stability, we have developed a solution using the consensus label strategy that makes a decision using a combination of models, unlike unique model prediction where performances fluctuate according to the used training set. In order to adapt the final prediction to the clinical context, our method has adaptive thresholds that allow the clinician to either promote sensitivity or have a balance between sensitivity and specificity.

# MATERIALS AND METHODS

## Patients

In this study, one group of 61 healthy patients and one group of 79 untreated B-CLL patients, with a Matutes score over 3 and a stage A in the Binet classification, were formed from two different measurement campaigns.

The first one was achieved in the years 2010-2011 during the ANR TecSan IHMO project and included 25 healthy and 45 B-CLL patients recruited at the Reims Champagne-Ardenne hospital center (RCA-HC).

The second campaign, performed in the years 2015-2016 during the EU CIP ICT PSP M3S project, included 36 healthy and 34 B-CLL patients recruited at Mont Godinne-Namur hospital center (MGN-HC). Both campaigns were realized with patient consent and approved by local ethics committees.

## Sample preparation

For each patient, one smear was prepared directly in the recruitment hospital, by spreading a blood drop deposited on a conventional glass slide in order to work in classical clinical conditions. At RCA-HC, a manual spreading is used, while an automatic spreading with an automated blood-smearing device (HemaPrep) is preferred at MGN-HC, both without prior chemical treatment. Each patient's blood is systematically analysed at the hospital by flow cytometry in order to know with certainty the pathophysiological label of the patient (healthy or CLL).

## Acquisition of Raman spectra

Raman data were acquired with a multimodal device developed by the TRIBVN company (Châtillon, France) combining a conventional microscope (ECLIPSE FN1, Nikon SA, Champigny-sur-Marne, France) and a Raman spectrometer (HORIBA FRANCE SAS, France).

The microscope was equipped with a motorized XYZ stage (Ludl Electronic Products Ltd, New-York, USA) and two dry lenses (Nikon): i) a 40x lens (NA 0.6) used to localize the position of approximately eighty lymphocytes on each smear, ii) a high magnification lens (100x/NA 0.9 in IHMO project and 150x/NA 0.95 in M3S project) used for Raman acquisitions on lymphocytes.

The Raman spectrometer was composed of a 532 nm laser excitation source (Toptica Photonics, Munich, Germany) delivering a power of 13 mW on the sample, a holographic grating of 1200 g/mm and a -70°C cooled-Pelletier CCD (Charge Coupled Device) detector (Andor Technology, South Windsor, CT, USA) of 1024 x 256 pixels and a 100  $\mu\text{m}$  confocal hole.

This set-up leads to a XY spatial resolution of 1  $\mu\text{m}$ , an axial resolution of 2  $\mu\text{m}$ , a spectral range from 700 to 3170  $\text{cm}^{-1}$  and a spectral resolution of 4  $\text{cm}^{-1}$ .

For the IHMO campaign, one Raman spectrum was acquired on the nucleus of 2596 healthy and 4257 B-CLL cells with an acquisition time fixed to two accumulations of 10 seconds, corresponding to 6853 acquired Raman spectra.

For the M3S project, three Raman spectra were acquired at three different positions on the nucleus of 1804 healthy and 3128 B-CLL cells, corresponding to 14796 Raman spectra, each

acquired during one accumulation of 10 seconds. For this study, the 3 spectra of each cell were averaged.

Each cell is thus characterized by its nucleus Raman spectrum and its physiopathological label (healthy or B-CLL) used in further supervised classification.

### **Quality tests**

The classification results are highly dependent on the quality of training data. Only high quality Raman spectra should be retained for the construction of robust and optimal models<sup>28</sup>. Four quality tests were designed in order to quantify the contribution of various disturbing factors<sup>26</sup>.

Signal-to-noise ratio (SNR) was computed by dividing the standard deviation of the first derivative of the lymphocyte signal in the 2800-3150  $\text{cm}^{-1}$  range by the standard deviation of the first derivative of the noise in the 1800-2200  $\text{cm}^{-1}$  range. Spectra with a SNR smaller than a predefined threshold were removed from further analysis. A nominal threshold of two has been determined as optimal.

Lymphocytes may be close to or superimposed on red blood cells, inducing Raman spectra contaminated by a red blood cell Raman signature. The lymphocyte signal to hemoglobin signal ratio (LSHSR) was computed as the ratio between the maximum intensity of the lymphocyte signal in the 1656-1720  $\text{cm}^{-1}$  range and the maximum intensity of the hemoglobin signal in the 1540-1656  $\text{cm}^{-1}$  range. Spectra with a LSHSR smaller than a predefined threshold were removed from the analysis. A nominal threshold of 1 has been determined as optimal.

Due to the use of a CCD detector, acquired Raman spectra can be contaminated by spikes generated by cosmic rays<sup>28</sup>. A spectrum is discarded if at least one of its intensities is superior to the whole data mean spectrum plus five times their standard deviation.

Due to cell fluorescence, Raman spectra can be saturated. Saturation is characterized by several consecutive wavenumbers having the same recorded maximum intensity allowed by the CCD. Spectra presenting such saturation are removed from the analysis.

The application of these four quality tests has discarded approximately 416 spectra (8%) from the M3S project and 822 spectra (12%) from the IHMO project.

### **Spectral data pre-processing**

To make data from the two campaigns comparable for further classification, a specific pre-processing was carried out. It consists in the application of a Savitzky-Golay smoothing<sup>29</sup> using a window length of 9 points and a third order polynomial function in order to reduce uncorrelated noise in the Raman spectra. The second step consists in the application of a collective method based on Extended Multiplicative Signal Correction (EMSC)<sup>30</sup> using the mean spectrum  $\hat{S}$  of the entire dataset as the reference spectrum to guide all corrections. Using an EMSC model, the baseline and the glass signal were neutralized at once using a fourth order polynomial function and a mean Raman spectrum of glass<sup>31</sup>, respectively. Finally, the spectra were normalized around the reference spectrum  $\hat{S}$ .

### **Datasets for numerical processing**

The pre-processed Raman data were divided into two sets. The first one (Dataset 1) is composed of 41 healthy patients (15 from IHMO and 26 from M3S campaigns) and 59 B-CLL patients (35 from IHMO and 24 from M3S campaigns). This dataset is used for feature selection and

optimization of supervised classification models as described below. The second set (Dataset 2) is composed of 20 healthy patients (10 from IHMO and 10 from M3S campaigns) and 20 B-CLL patients (10 from IHMO and 10 from M3S campaigns). This independent test set is blindly tested on these optimized classification models in order to assess their performance.

### Feature selection

In this work, a canonical correlation analysis<sup>32</sup> as applied to Dataset 1 in order to achieve a supervised feature selection aiming at identifying the discriminant variables between Raman spectra acquired on lymphocytes of healthy and CLL patients.

In order further to decrease the number of selected variables, a supervised data dimension reduction was developed in order to remove redundant selected variables. The first step consisted in the computation of the Fisher-score at each selected wavenumber. The second step computed the wavenumber correlation coefficient matrix  $\mathbf{R}_\lambda$  from the dataset. In a third step, the elements of  $\mathbf{R}_\lambda$  being smaller than a threshold fixed to 0.7 were removed. Each line of  $\mathbf{R}_\lambda$  was thus composed of highly correlated wavenumbers. In the fourth step, each line was reduced to its wavenumber having the highest Fisher-score computed in the first step, i.e. the most discriminant wavenumber. The last step computed the unique wavenumbers composing  $\mathbf{R}_\lambda$ , i.e. removed repeated wavenumbers. At the end of this fast procedure, all the spectral information is summarized in its most representative and discriminant wavenumbers.

### Supervised classification algorithms

The aim of this study was to optimize a model able to automatically diagnose B-CLL patients from healthy ones based on their lymphocyte Raman signature. In this project, three different supervised classification algorithms were tested for their known high performance in vibrational spectroscopy: Partial Least Squares - Discriminant Analysis (PLS-DA)<sup>33,34</sup>, Support Vector Machine (SVM)<sup>35-37</sup> and Random Forest (RF)<sup>35,38</sup>.

PLS-DA is a linear method used to model the covariance structure between data and labels by latent variables. This method has been proved particularly effective for collinear data composed of more features than observations, typically observed in Raman spectroscopy studies<sup>39,40</sup>. This method is parametrized by the number of latent variables  $n_{lv}$ . In this study, this parameter can take the following values:  $n_{lv} \in \{1, 2, \dots, 40\}$ .

SVM is a non-linear method designed to implicitly map data into a high dimensional feature space using a kernel function in order to separate classes with a hyperplane maximizing their margin. In this study, a Gaussian RBF kernel parameterized by  $\gamma$  was chosen because of its widespread popularity. The SVM model estimation was achieved by the  $\nu$ -SVM algorithm implemented by the LIBSVM library<sup>41</sup>. In this study, these parameters were chosen among the following values:  $\gamma \in \{10^{-8}, 10^{-7}, \dots, 10^3\}$  and  $\nu \in \{0.1, 0.2, \dots, 0.9\}$ .

RF is a non-linear method based on the construction of a multitude  $n_{dt}$  of decision trees. Each tree, with a predefined depth  $d$ , is built using a random selection of the data and a predefined number of features randomly selected. The final RF decision is based on majority voting of all the decision trees. In this study, we selected these parameters in:  $n_{dt} \in \{10, 30, \dots, 200\}$  and  $d \in \{10, 20, \dots, 100\}$ .

### Repeated double cross-validation for classifier optimization

Classifier construction and its parameter optimization is realized by a training step necessary to construct predictive models. Among the different existing training strategies, a repeated double

cross-validation (rdCV)<sup>42,43</sup> was used in this work because of its known reliable model prediction performance estimation compared to a simple cross-validation (CV) which can overestimate the model's performance<sup>44</sup>. Furthermore, as a dataset must always be carried out at the highest hierarchical level in order to properly evaluate classification models<sup>45</sup>, data were considered at the patient level in our study (not at the spectrum level). After the quality test, each patient corresponds to a set of about 76 cells associated with their Raman spectra.

The rdCV consists in two nested loops, as represented in Figure 1. The inner loop performs a grid search model tuning in order to optimize the parameters associated with each classification algorithm. At each iteration of the inner loop, the parameters of the classification algorithm are fixed to predefined values  $\{i, j\}$  ( $j$  is empty for PLS-DA), and an inner cross-validation loop procedure (simple CV), coupled to feature selection, is run in order to estimate the most discriminant and uncorrelated wavenumbers and to evaluate the mean accuracy of the model constructed with this set of parameters and applied on an internal validation set. As suggested in literature<sup>45</sup>, feature selection is included in the cross-validation loop in order to prevent model from overfitting. At the end of this inner loop, a new classification model is constructed on the entire training set reduced to the selected wavenumbers using the best performing parameters, i.e. those maximizing the mean accuracy.

The outer loop repeats  $T = 100$  times the inner loop for different training sets. Indeed, at each iteration of the outer loop, data from Dataset 1 are randomly divided into two sets, i.e. a training set composed of 70% of patients is used to optimize the classifier during the inner loop, while the 30% remaining patients compose the external validation set used to independently evaluate the predictive performance of this optimized model in term of sensitivity and specificity. At the end of this outer loop, 100 optimal models estimated on different training sets and predicting the physiopathological state of a patient, i.e. healthy or B-CLL, are obtained.

Among the existing cross-validation strategies, three popular methods were tested in the inner loop, i.e. the methods named leave-one-patient-out-validation (LOPOCV), K-fold cross-validation (KFCV) and Monte-Carlo cross-validation (MCCV). In LOPOCV, the  $N$  patients are divided into  $N$  one-patient-folds, inducing  $N$  iterations. At each iteration, the training set is composed of  $N - 1$  patients and the validation set is composed of the remaining patient. In KFCV,  $K$  folds of patients are randomly constructed, inducing  $K$  iterations. In this study, the classical value  $K = 10$  was chosen. At each iteration, the training set is composed of  $K - 1$  folds and the validation set is composed of the remaining fold. At each iteration of the MCCV, the validation set is composed of  $M$  patients randomly selected. The remaining  $N - M$  patients compose the training set. In this work, the number  $M$  of patients selected for the validation set and the number of repetitions of MCCV were chosen equal to 10 and 100, respectively.

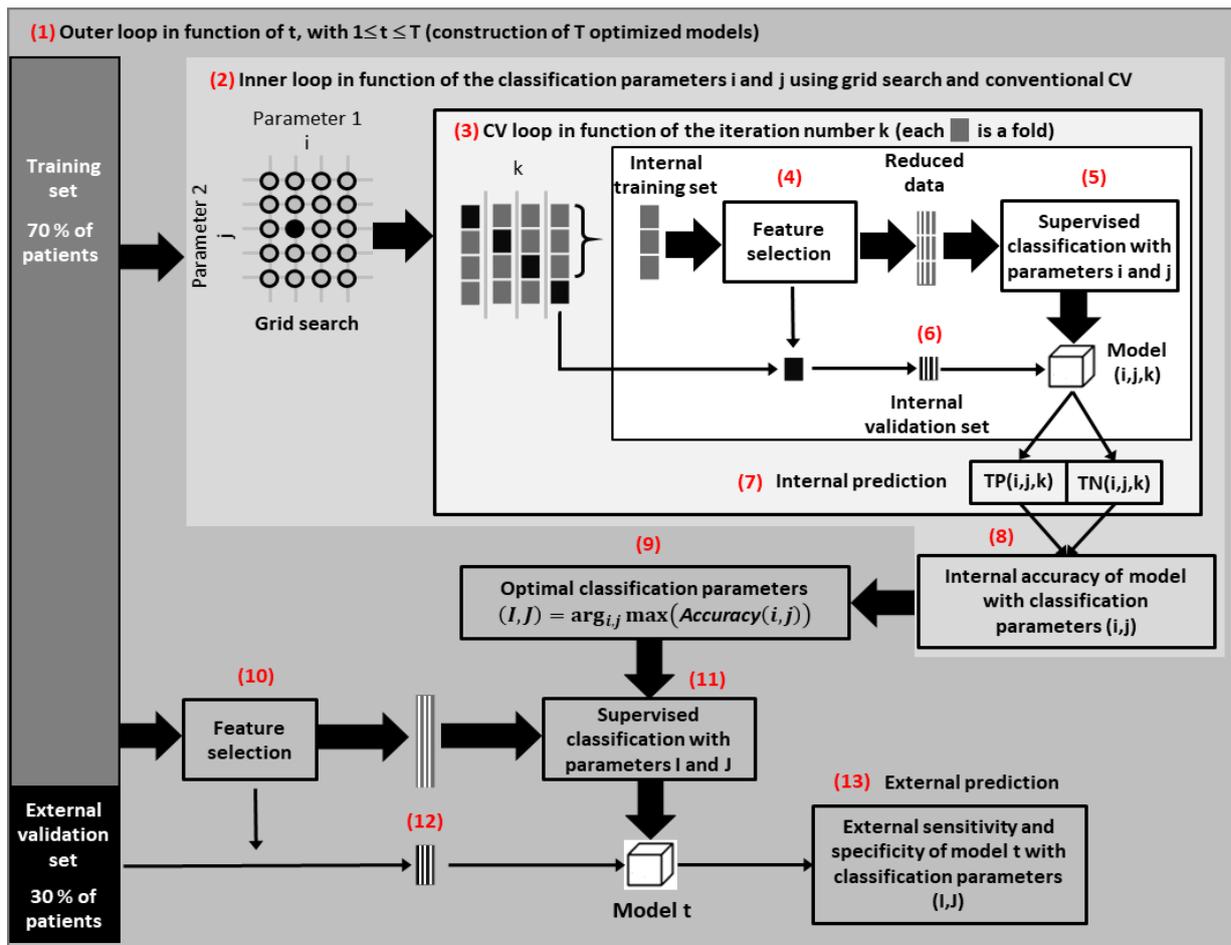


Figure 1 - Flowchart of the application on Dataset 1 of rdCV which is a combination of an inner loop composed of a conventional CV method (LOPOCV, KFCV or MCCV), a supervised feature selection and a supervised classification algorithm (PLS-DA, SVM or RF) in order to train a classifier and optimize its parameters, and an outer loop aiming at constructing several models based on different training sets. (1) At each iteration of the outer loop, patients from Dataset 1 are randomized and divided into a training set composed of 70% of patients and an external validation set composed of 30% of patients. (2) At each iteration of the inner loop, a set of classification parameters is selected as part of a grid search. (3) Using these parameters, a conventional CV is run. (4) At each step of the CV loop, the internal training set feeds the supervised feature selection in order to identify the most discriminant and decorrelated features. (5) On this reduced data, a supervised classification model is trained using the parameters selected at step (2). (6) The internal validation set is reduced to the features estimated at step (4). (7) These data are injected into the model constructed at step (5) in order to estimate its generalization performance measured by the number of true positive (TP) and true negative (TN) patients. (8) At the end of the CV loop, TP and TN are used to compute the accuracy of the supervised classification algorithm using the parameters selected at step (2). (9) At the end of the inner loop, i.e. at the end of the grid search, the optimal classification parameters are estimated as those maximizing the accuracy computed at step (8). (10) The entire training set is now used to identify a unique set of the most discriminant and decorrelated features by the supervised feature selection. (11) The resulting reduced training set is used to compute a unique classification model using the optimal parameters estimated at step (9). (12) The external validation data are reduced to the features estimated at step (10). (13) They are injected into the model constructed at step (11) in order to estimate its generalization performance in term of sensitivity and specificity. Steps (1) to (13) are repeated  $T = 100$  times, resulting in the construction of 100 optimized models.

### **Patient decision threshold**

After the pre-processing steps, each patient is represented approximately by 76 Raman spectra, each one being representative of one lymphocyte. A patient is classified as B-CLL by a classification model if the proportion of its cells classified as B-CLL is greater than a patient decision threshold  $\tau$ , and as healthy otherwise. In the following,  $\tau$  has been first fixed arbitrarily to 50%. The impact of the choice of this threshold will be discussed in section “Adaptive patient decision threshold in function of clinical objective”.

### **Blind test of the optimized classifiers**

The last step consists in testing the 100 models on the independent test set (see figure 1) represented by Dataset 2 that has not been used in the rdCV. Each optimized model predicts the physiopathological state of each test patient, i.e. healthy or B-CLL. For a patient, the final decision is obtained using majority voting of the prediction of the 100 optimized models.

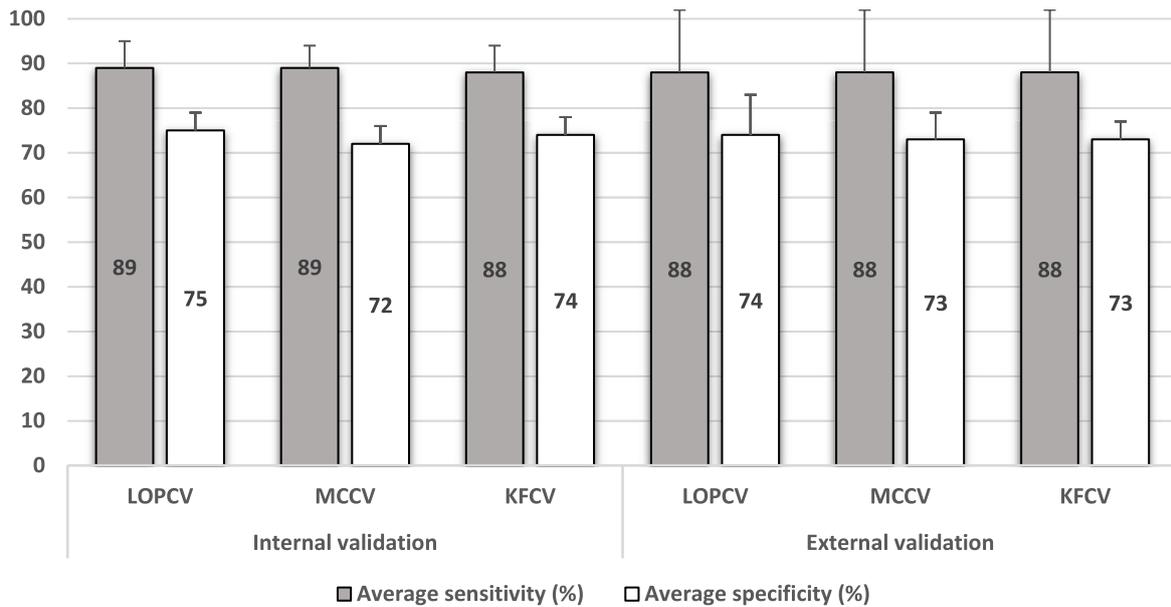
## **RESULTS AND DISCUSSION**

### **Evaluation of the performance of the different cross-validation methods**

One important element of the development of a supervised classification scheme is the choice of a CV method. Most of spectroscopic studies are performed using a simple CV method, such as LOPOCV<sup>25</sup>, KFCV<sup>26</sup> and MCCV<sup>27</sup>. However, several studies<sup>46,47</sup> have demonstrated the efficiency of repeated CV in term of a better estimate of the performance.

That is the reason why rdCV was used in our study. As previously explained in the section “Materials and methods”, its inner loop includes a CV method, such as LOPOCV, KFCV and MCCV, which were evaluated in this work. Performance of a CV strategy relying more on data statistics than on the choice of the supervised classification algorithm, this study on the inner CV of rdCV was led using PLS-DA exclusively. The mean and standard deviation of sensitivities and specificities estimated by each CV method on internal and external validation sets are summarized in Figure 2. Whatever the validation set, no significant difference can be observed between the CV methods. Furthermore, each CV method gives exactly the same results for internal and external validation sets, with a mean sensitivity around 88% and mean specificity around 74%. As expected, standard deviations of sensitivity and specificity are higher for the external validation set than for the internal one. The similarity between the performances of cross-validation methods can be explained by the high number of repetitions  $T = 100$  of the external loop of rdCV.

These results are in accordance with a previous study<sup>46</sup> which showed similar performance of LOPOCV and KFCV with  $k = 10$ . However, LOPOCV and MCCV are more time-consuming than KFCV which is also known to offer a good compromise between variance and bias of predictive models<sup>28</sup>. Therefore, KFCV was applied in the rest of the study.

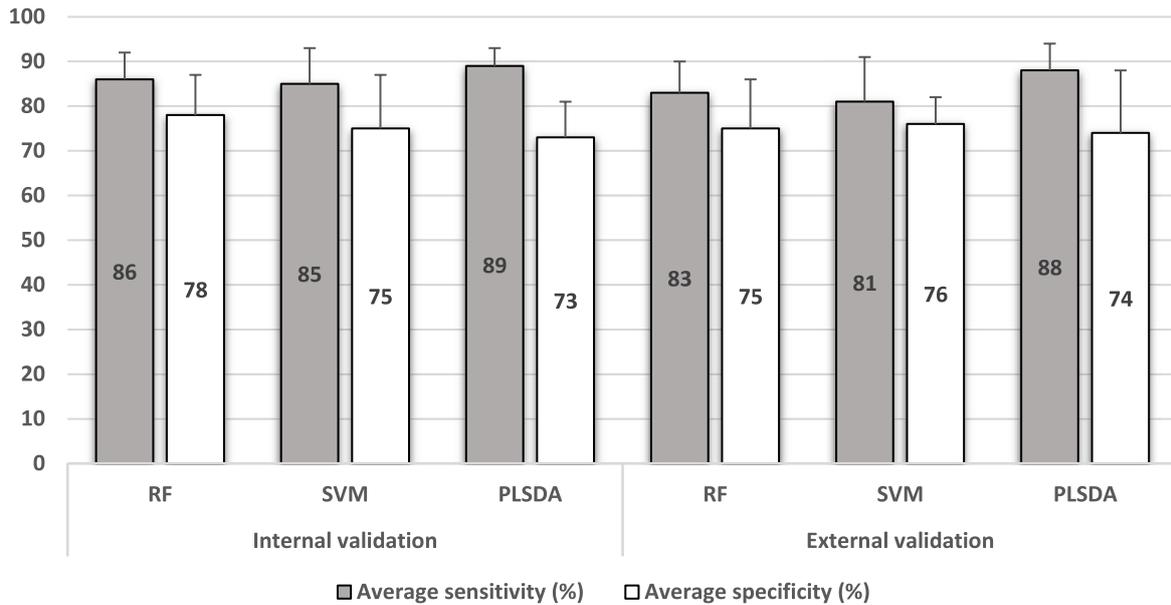


**Figure 2 - Comparison of different cross-validation techniques (LOPCV, KFCV, MCCV), for the inner loop of rdCV using PLS-DA, expressed in terms of mean sensitivity and specificity computed on the internal and external validation sets.**

### Assessment of the classifying capability of Raman spectroscopy

Three supervised algorithms (SVM, RF and PLS-DA) based on different principles were applied to the dataset. As presented in Figure 3, these three methods have the same mean performance when applied 100 times, for both the internal validation set, with mean sensitivity and specificity around 87% and 75% respectively, and the external validation set, with mean sensitivity and specificity around 84% and 75% respectively. Note also that the standard deviation is smaller for PLS-DA than other algorithms. In literature, non-linear supervised classification methods, such as RF and SVM, have been shown better than linear methods, such as PLS-DA, when data are not linearly separable<sup>48,49</sup>. Thus, our results could show that the B-CLL/healthy classification is a linearly separable problem.

Moreover, PLS-DA depends on a unique parameter, i.e. the number of latent variables. Its optimization is thus simpler than SVM and RF. Taken together, these results show that PLS-DA is a valuable method for the classification of Raman spectra acquired on B-CLL and healthy lymphocytes in clinical conditions. It has to be noticed that the similar results obtained on both internal and external validation sets prove that no overfitting occurs during data training thanks to the rdCV strategy<sup>42,50</sup>.



**Figure 3 - Comparison of different supervised classification algorithms (RF, SVM, PLSDA), using KFCV for the inner loop of rdCV, expressed in terms of mean sensitivity and specificity computed on the internal and external validation sets.**

### **Adaptive patient decision threshold in function of clinical objective**

The previously presented results were obtained using a patient decision threshold  $\tau$  of 50% that is considered as the default value for classifying patients based on Raman spectra cell. However, this arbitrary value is questionable since it may influence the determination of the optimized classification model at each iteration of the outer loop. It is thus important to evaluate the impact of this parameter on the classification results.

A naïve approach would be to fix  $\tau$  to the same value for all iterations of the outer loop. However, the training dataset changing at each outer loop iteration, a more objective strategy is to adaptively select the optimal  $\tau$  value at each iteration in order to maximize diagnostic performances linked to the clinical objective. In this study, two different clinical objectives have been considered: i) balanced sensitivity and specificity (objective 1) in order to have the best compromise between sensitivity and specificity, ii) sensitivity maximization (objective 2) in order to reduce the risk of misdiagnosing B-CLL patients.

Technically, a Receiver Operating Characteristic (ROC) curve was constructed by varying  $\tau$  for each model estimated at each inner loop iteration. Concretely, for a high value of  $\tau$ , all patients are classified as healthy, inducing a sensitivity of 0% and a specificity of 100%. A progressive reduction of  $\tau$  leads to an increase of sensitivity and a decrease of specificity. For a low value of  $\tau$ , all patients are classified as CLL, inducing a sensitivity of 100% and a specificity of 0%. Figure 4a depicts an example of ROC curves for a model estimated at an inner loop. If objective 1 is considered, an increase of specificity from 77% to 84% and a small decrease of sensitivity from 85% to 84% are obtained when the threshold of the black curve is slightly increased from 50% to the optimized value of 53%.

The point of ROC curves that, i) minimizes the distance between sensitivity and specificity for objective 1 or ii) maximizes the sensitivity for objective 2, defines the best  $\tau$  value and the best

accuracy associated to each model of the inner loop. At the end of each inner loop, the parameters of PLS-DA and threshold  $\tau$  of the model with the highest accuracy are selected. As illustrated on Figure 4b, the optimal threshold  $\tau$  is highly variable with values ranging from 35% to 85%. This high variability is due to the random selection of the training set at each outer loop iteration. This result demonstrates importance of the optimization of the threshold  $\tau$  in order to adapt the best classification model selection to each training set.

As can be seen on Figure 5, this result can be generalized to the 100 models estimated during outer loop. The proposed adaptive patient decision threshold procedure is efficient since: i) a rebalancing is visible between the sensitivity and the specificity (around 80%) for objective 1, and ii) a significant increase of sensitivity (around 95%) can be observed for objective 2, both for the internal and external validation sets, compared to the results obtained using a 50% threshold (see Figure 3).

Due to the ROC curve definition, the choice of a threshold inducing an increase of sensitivity will surely induce a decrease of specificity, and vice versa<sup>51</sup>. Indeed, if we compare the results for PLS-DA in Figure 3 and the results in Figure 5 for internal and external validation for objective 1, the specificity increased from around 74% to 80%, while the sensitivity decreased from around 89% to 80%, compared to the results obtained with a 50% threshold. In addition, for objective 2, the increase in sensitivity from around 89% to 95% is followed by a sharp decrease in specificity from around 74% to 54%. These results demonstrate the importance of parameterization and the possibility of adapting these parameters according to the clinician needs.

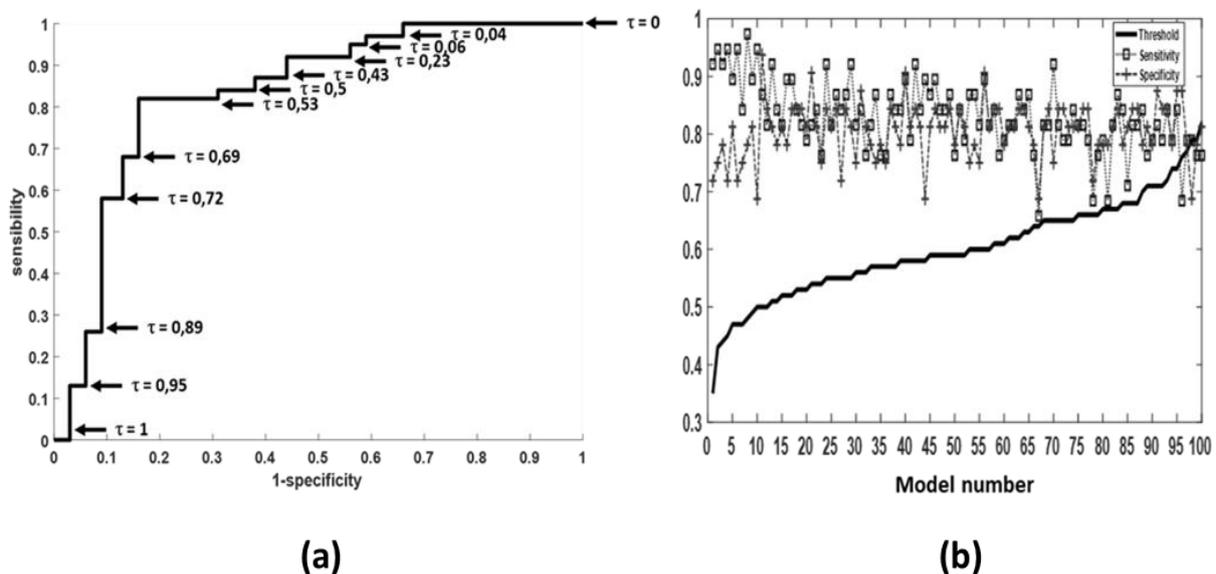
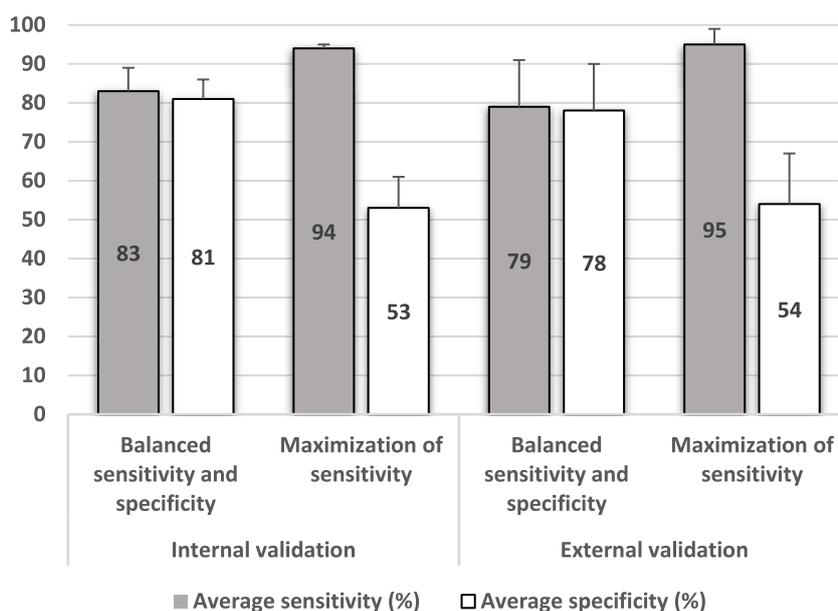


Figure 4 - (a) Two examples of ROC curves computed on internal validation set by varying the patient decision threshold  $\tau$  for two models estimated at two different inner loops of rdCV. (b) Variability of the optimal threshold  $\tau$ , of sensitivity and of specificity in function of the model number of the outer loop. The models are sorted according to the increasing value of  $\tau$ .



**Figure 5 – Mean sensitivity and specificity computed on internal and external validation sets after optimal optimization of threshold  $\tau$  for two different objectives, balanced sensitivity and specificity, and sensitivity maximization**

### **Improvement of classification predictive performance using label consensus**

Most biomedical studies related to vibrational spectroscopy evaluated the predictive performance of a unique optimized model from an independent test set unseen during the training phase<sup>52-54</sup>. For objective 1 (balanced sensitivity and specificity), we have evaluated independently the predictive performance of each of the 100 optimized models on an independent test set (Dataset 2) composed of 40 patients (20 healthy and 20 B-CLL patients). These results are summarized in Figure 6a as mean and standard deviation of sensitivity and specificity. The results show a similar mean sensitivity and specificity to those obtained during the training and validation (see Figure 5). This coherence of predictive performance is characteristic of a non-overfitting of the classifiers. However, the individual performances of the 100 models varied from 56% to 94% for sensitivity and from 43% to 95% for specificity, showing the great influence of the selection of the training set on the performance of the classifiers. Some models being less efficient, the strategies consisting in the training of a unique model or in the averaging of several models are thus unsuccessful approaches.

In our study, a different approach based on the label consensus was proposed. The principle is based using a set of individual models whose predictions were combined by majority voting. In the case of an unknown patient, each of the 100 optimized models realizes a prediction and assigns a label (healthy or B-CLL) to the patient. The predictions of the 100 models are then merged and the algorithm chooses the class label that receives the most votes. Figure 6b shows the performance obtained with this strategy, i.e. a sensitivity of 95% and a specificity of 85%. Compared to the average of all models (Figure 6a), label consensus led to increases of 14% for sensitivity and 8% for specificity.

By looking more closely at the results, label consensus applied at the patient level on the test set resulted in the misclassification of one healthy patient and three B-CLL patients. Two of these decisions are unclear since less than 65% of the optimized models wrongly classified this healthy patient and one of these B-CLL patients. On the contrary, the two other B-CLL patients are clearly misclassified by more than 85% of the models.

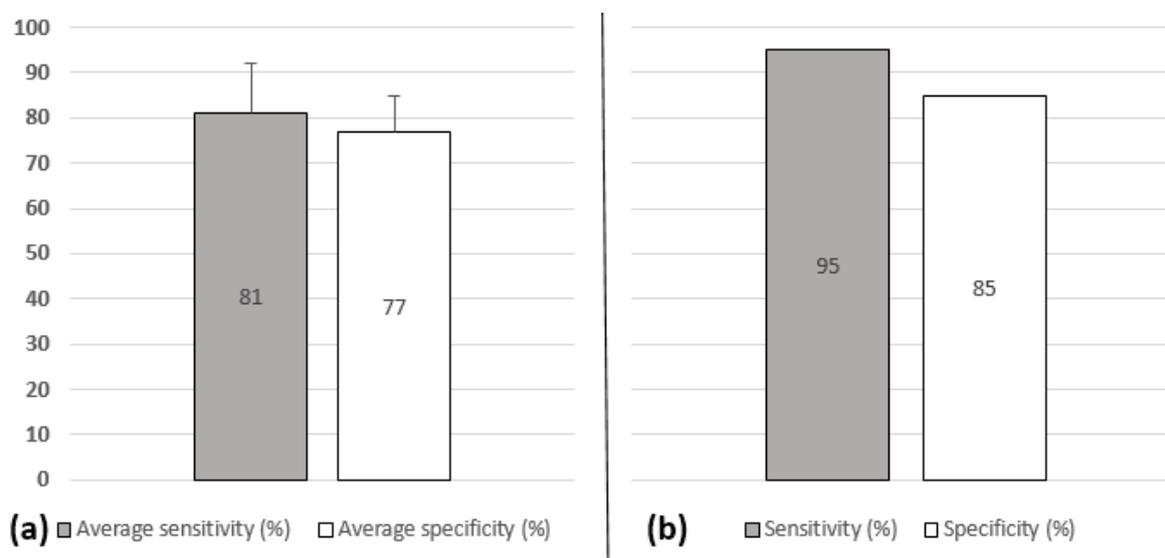


Figure 6 - Comparison sensibility and specificity for independent models (a) and label consensus (b)

### Assignment of the discriminant Raman vibrations between healthy and B-CLL patients

As can be seen on Figure 7, the mean Raman spectra of healthy and B-CLL patients after the spectral data pre-processing step are highly correlated and no evident difference is visible. In order to highlight the subtle discriminant features between these two groups of spectra, a supervised feature selection algorithm described in section “Feature selection” was used at each rdCV loop. Each of the 100 models was thus optimized on selected spectral features to differentiate the B-CLL group from the healthy group. The 50 most common variables among the 100 runs are shown in Figure 7 as black vertical bands.

The spectral bands selected by the algorithm are attributed in Table 1 and mainly correspond to nucleic acid vibrations<sup>55</sup>, i.e. 745  $\text{cm}^{-1}$  for adenine, thymine and cytosine respiration and 1577  $\text{cm}^{-1}$  for guanine and adenine. Other bands are associated with RNA and DNA phosphate stretching bands at 813  $\text{cm}^{-1}$ , 840  $\text{cm}^{-1}$  and 1100  $\text{cm}^{-1}$ . The predominance of these bands is related to the fact that the spectra were acquired on the nucleus of the cells. Other typical contributions can be found mainly due to proteins such as vibrations within the Amide I and Amide III bands at 1660  $\text{cm}^{-1}$  and 1252  $\text{cm}^{-1}$  respectively. The band at 1006  $\text{cm}^{-1}$  originates from the aromatic amino acid phenylalanine. Cytochrome c bands<sup>55,56</sup> are visible at 745  $\text{cm}^{-1}$ , 755  $\text{cm}^{-1}$ , 1134  $\text{cm}^{-1}$ .

These observations show that DNA, protein and cytochrome c bands can be biochemical markers that discriminate between healthy and diseased states. These results are in accordance with literature which has shown that Raman spectroscopy is capable of distinguishing healthy cells from cancer cells by DNA condensation or protein modifications<sup>15,57</sup>. Furthermore, in the

case of chronic lymphoid leukemia, pathological lymphocytes have a modified chromatin<sup>58,59</sup> (clumped chromatin). Cytochrome c plays a central role in cellular apoptosis<sup>60</sup> and in the cellular respiration cycle<sup>60</sup>. Variations in cytochrome c bands could reveal a metabolic acceleration of leukemic cells that divide anarchically or show dysfunction at the apoptosis level.

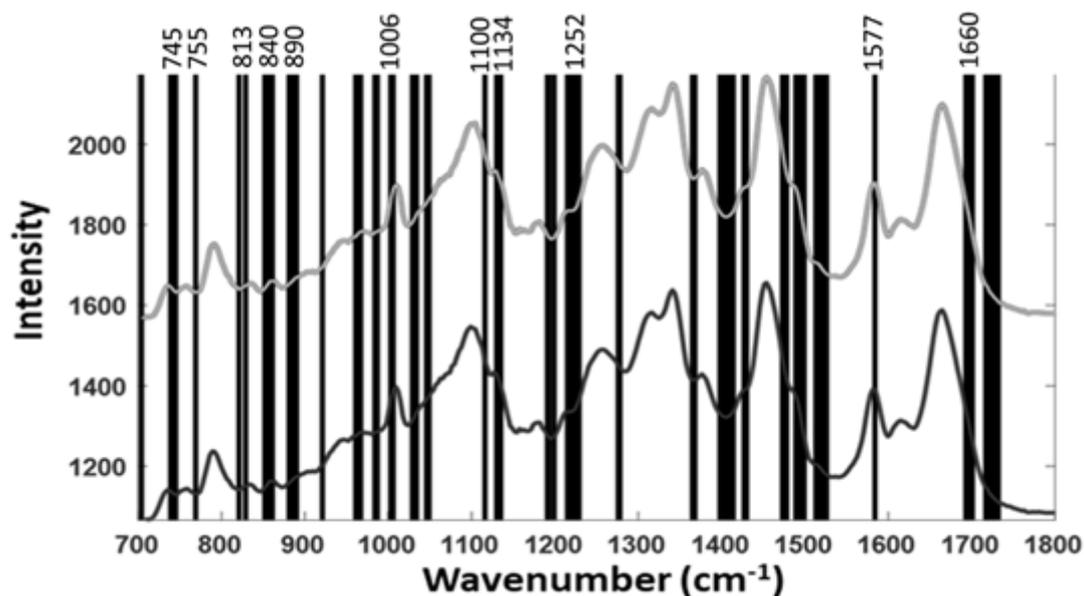


Figure 7 - Average spectra of healthy (grey curve) and B-CLL (black curve) patients after spectral data pre-processing step, and their discriminant wavenumbers represented by black vertical bands.

Table 1- Position and assignment of characteristic bands to differentiate healthy from CLL patients<sup>26,56,61,62</sup>

Raman shift (cm <sup>-1</sup> )	Attributions
745	A,T,C, Trp (sym br), Cytochrome c
755	Cytochrome c
833	PO2 asym str (DNA), Tyr
929	DNA bk, C-C str
1006	Phe (phenyl ring)
1120	PO2 bk, C-N str, Cytochrome c
1134	Cytochrome c
1206	Phe, Amide III
1252	Amide III

<b>1370</b>	A,T,G, Cytochrome c
<b>1420</b>	CH def, CH2 def
<b>1485</b>	A,G, CH def
<b>1520</b>	A
<b>1577</b>	A,G, Phe, Cytochrome c
<b>1660</b>	Amide I

Abbreviations: A, adenine; C, cytosine; G, guanine; T, thymine; Tyr, tyrosine; Trp, tryptophan; Phe, phenylalanine; sym, symmetric; asym, asymmetric; br, breathing; str, stretching; def, deformation; bk, backbone.

## CONCLUSION

This paper demonstrated that the combination of Raman spectroscopy with an effective supervised classification strategy is able to accurately diagnose chronic lymphocytic leukemia. First, the use of a repeated double cross validation strategy is efficient whatever the considered cross-validation technique and supervised classification algorithm, due to the high number of repetitions inducing an exhaustive exploration of the training set. Second, we introduced an adaptive patient decision threshold based on the number of cells per patient which is necessary to decide from the patient classification. In fact, the choice of the threshold is automatically driven by the clinical objective, i.e. balanced sensitivity and specificity, or maximization of sensitivity, or maximization of specificity. Finally, label consensus exploited the diversity of trained classifiers to aggregate their predictions and improve classification results on an independent test set.

## REFERENCES

- (1) Huang, Z.; McWilliams, A.; Lui, H.; McLean, D. I.; Lam, S.; Zeng, H. *International journal of cancer* **2003**, *107*, 1047-1052.
- (2) Talari, A. C. S.; Movasaghi, Z.; Rehman, S.; Rehman, I. U. *Applied Spectroscopy Reviews* **2015**, *50*, 46-111.
- (3) Kong, K.; Kendall, C.; Stone, N.; Notingher, I. *Advanced drug delivery reviews* **2015**, *89*, 121-134.
- (4) Vuiblet, V.; Fere, M.; Bankole, E.; Wynckel, A.; Gobinet, C.; Birembaut, P.; Piot, O.; Rieu, P. *Scientific reports* **2016**, *6*, 33045.
- (5) Vuiblet, V.; Nguyen, T. T.; Wynckel, A.; Fere, M.; Van-Gulick, L.; Untereiner, V.; Birembaut, P.; Rieu, P.; Piot, O. *Analyst* **2015**, *140*, 7382-7390.
- (6) Sharma, N.; Takeshita, N.; Ho, K. Y. *Clinical endoscopy* **2016**, *49*, 404.
- (7) Rohleder, D.; Kiefer, W.; Petrich, W. *Analyst* **2004**, *129*, 906-911.
- (8) Khan, S.; Ullah, R.; Khan, A.; Ashraf, R.; Ali, H.; Bilal, M.; Saleem, M. *Photodiagnosis and photodynamic therapy* **2018**, *23*, 89-93.
- (9) Pinto, J. *Cancer Classification in Human Brain and Prostate Using Raman Spectroscopy and Machine Learning*. University of Waterloo 2017.
- (10) Teh, S. K.; Zheng, W.; Ho, K. Y.; Teh, M.; Yeoh, K. G.; Huang, Z. *British journal of cancer* **2008**, *98*, 457.
- (11) Austin, L. A.; Osseiran, S.; Evans, C. L. *Analyst* **2016**, *141*, 476-503.
- (12) Ellis, D. I.; Goodacre, R. *Analyst* **2006**, *131*, 875-885.
- (13) Hobro, A. J.; Konishi, A.; Coban, C.; Smith, N. I. *Analyst* **2013**, *138*, 3927-3933.
- (14) Crow, P.; Barrass, B.; Kendall, C.; Hart-Prieto, M.; Wright, M.; Persad, R.; Stone, N. *British journal of cancer* **2005**, *92*, 2166.
- (15) Chan, J. W.; Taylor, D. S.; Zwerdling, T.; Lane, S. M.; Ihara, K.; Huser, T. *Biophysical journal* **2006**, *90*, 648-656.
- (16) Manago, S.; Zito, G.; De Luca, A. C. *Optics & Laser Technology* **2018**, *108*, 7-16.
- (17) Vanna, R.; Ronchi, P.; Lenferink, A. T. M.; Tresoldi, C.; Morasso, C.; Mehn, D.; Bedoni, M.; Picciolini, S.; Terstappen, L.; Ciceri, F. *Analyst* **2015**, *140*, 1054-1064.
- (18) Draux, F.; Jeannesson, P.; Beljebbar, A.; Tfayli, A.; Fourre, N.; Manfait, M.; Sulé-Suso, J.; Sockalingum, G. D. *Analyst* **2009**, *134*, 542-548.
- (19) Ramoji, A.; Neugebauer, U.; Bocklitz, T.; Foerster, M.; Kiehntopf, M.; Bauer, M.; Popp, J. r. *Analytical chemistry* **2012**, *84*, 5335-5342.
- (20) Meade, A. D.; Lyng, F. M.; Knief, P.; Byrne, H. J. *Analytical and bioanalytical chemistry* **2007**, *387*, 1717-1728.
- (21) Crow, P.; Stone, N.; Kendall, C. A.; Uff, J. S.; Farmer, J. A. M.; Barr, H.; Wright, M. P. J. *British journal of cancer* **2003**, *89*, 106.
- (22) Del Mistro, G.; Cervo, S.; Mansutti, E.; Spizzo, R.; Colombatti, A.; Belmonte, P.; Zucconelli, R.; Steffan, A.; Sergio, V.; Bonifacio, A. *Analytical and bioanalytical chemistry* **2015**, *407*, 3271-3275.
- (23) Liu, W.; Wang, H.; Du, J.; Jing, C. *Biosensors and Bioelectronics* **2017**, *97*, 70-74.
- (24) Larraona-Puy, M.; Ghita, A.; Zoladek, A. B.; Perkins, W.; Varma, S.; Leach, I. H.; Koloydenko, A.; Williams, H.; Notingher, I. *Journal of biomedical optics* **2009**, *14*, 054031.
- (25) Lyng, F. M.; Faolain, E. O.; Conroy, J.; Meade, A. D.; Knief, P.; Duffy, B.; Hunter, M. B.; Byrne, J. M.; Kelehan, P.; Byrne, H. J. *Experimental and molecular pathology* **2007**, *82*, 121-129.
- (26) Happillon, T.; Untereiner, V.; Beljebbar, A.; Gobinet, C.; Daliphard, S.; Cornillet-Lefebvre, P.; Quinquenel, A.; Delmer, A.; Troussard, X.; Klossa, J. *Analyst* **2015**, *140*, 4465-4472.

- (27) Graça, G.; Moreira, A. S.; Correia, A. J. V.; Goodfellow, B. J.; Barros, A. S.; Duarte, I. F.; Carreira, I. M.; Galhano, E.; Pita, C.; do Céu Almeida, M. *Analytica Chimica Acta* **2013**, *764*, 24-31.
- (28) Bocklitz, T.; Walter, A.; Hartmann, K.; Rösch, P.; Popp, J. *Analytica Chimica Acta* **2011**, *704*, 47-56.
- (29) Savitzky, A.; Golay, M. J. E. *Analytical chemistry* **1964**, *36*, 1627-1639.
- (30) Afseth, N. K.; Kohler, A. *Chemometrics and Intelligent Laboratory Systems* **2012**, *117*, 92-99.
- (31) Kerr, L. T.; Hennelly, B. M. *Chemometrics and Intelligent Laboratory Systems* **2016**, *158*, 61-68.
- (32) Hardoon, D. R.; Szedmak, S.; Shawe-Taylor, J. *Neural computation* **2004**, *16*, 2639-2664.
- (33) Ming, L. C.; Gangodu, N. R.; Loh, T.; Zheng, W.; Wang, J.; Lin, K.; Zhiwei, H. *Oncotarget* **2017**, *8*, 49443.
- (34) Barker, M.; Rayens, W. *Journal of Chemometrics* **2003**, *17*, 166-173.
- (35) Maguire, A.; Vega-Carrascal, I.; Bryant, J.; White, L.; Howe, O.; Lyng, F. M.; Meade, A. D. *Analyst* **2015**, *140*, 2473-2481.
- (36) Neugebauer, U.; Bocklitz, T.; Clement, J. H.; Krafft, C.; Popp, J. *Analyst* **2010**, *135*, 3178-3182.
- (37) Cortes, C.; Vapnik, V. *Machine learning* **1995**, *20*, 273-297.
- (38) Breiman, L. *Machine learning* **2001**, *45*, 5-32.
- (39) Gaydou, V.; Polette, M.; Gobinet, C.; Kileztky, C.; Angiboust, J.-F.; Manfait, M.; Birembaut, P.; Piot, O. *Analytical chemistry* **2016**, *88*, 8459-8467.
- (40) Palermo, A.; Fosca, M.; Tabacco, G.; Marini, F.; Graziani, V.; Santarsia, M. C.; Longo, F.; Lauria, A.; Cesareo, R.; Giovannoni, I. *Analytical chemistry* **2017**.
- (41) Chang, C.-C.; Lin, C.-J. *ACM transactions on intelligent systems and technology (TIST)* **2011**, *2*, 27.
- (42) Filzmoser, P.; Liebmann, B.; Varmuza, K. *Journal of Chemometrics* **2009**, *23*, 160-171.
- (43) Varmuza, K.; Filzmoser, P. *Current Applications of Chemometrics. Ed. Khanmohammadi, M., Nova Science Publishers, Hauppauge, NY, USA* **2014**, 15-32.
- (44) Trevor, H.; Robert, T.; Jh, F.; New York, NY: Springer, 2009.
- (45) Guo, S.; Bocklitz, T.; Neugebauer, U.; Popp, J. r. *Anal. Methods* **2017**.
- (46) Molinaro, A. M.; Simon, R.; Pfeiffer, R. M. *Bioinformatics* **2005**, *21*, 3301-3307.
- (47) Kim, J.-H. *Computational statistics & data analysis* **2009**, *53*, 3735-3745.
- (48) Sattlecker, M.; Bessant, C.; Smith, J.; Stone, N. *Analyst* **2010**, *135*, 895-901.
- (49) Bergner, N.; Bocklitz, T.; Romeike, B. F. M.; Reichart, R.; Kalff, R.; Krafft, C.; Popp, J. r. *Chemometrics and Intelligent Laboratory Systems* **2012**, *117*, 224-232.
- (50) Schoeller, D. A.; Westerterp, M. *Advances in the Assessment of Dietary Intake*; CRC Press, 2017.
- (51) Florkowski, C. M. *The Clinical Biochemist Reviews* **2008**, *29*, S83.
- (52) Ramos, I. R.; Meade, A. D.; Ibrahim, O.; Byrne, H. J.; McMEnamin, M.; McKenna, M.; Malkin, A.; Lyng, F. M. *Faraday discussions* **2016**, *187*, 187-198.
- (53) Haifler, M.; Pence, I.; Sun, Y.; Kutikov, A.; Uzzo, R. G.; Mahadevan, J.; Jansen, A.; Patil, C. A. *Journal of biophotonics* **2018**.
- (54) Hlaing, M. M.; Dunn, M.; Stoddart, P. R.; McArthur, S. L. *Vibrational Spectroscopy* **2016**, *86*, 81-89.
- (55) Manago, S.; Mirabelli, P.; Napolitano, M.; Zito, G.; De Luca, A. C. *Journal of biophotonics* **2017**.
- (56) Hobro, A. J.; Kumagai, Y.; Akira, S.; Smith, N. I. *Analyst* **2016**, *141*, 3756-3764.
- (57) Poplineau, M.; Trussardi-Régnier, A.; Happillon, T.; Dufer, J.; Manfait, M.; Bernard, P.; Piot, O.; Antonicelli, F. *Epigenomics* **2011**, *3*, 785-794.

- (58) Peterson, L. C.; Bloomfield, C. D.; Sundberg, R. D.; Gajl-Peczalska, K. J.; Brunning, R. D. *The American journal of medicine* **1975**, *59*, 316-324.
- (59) Oscier, D.; Else, M.; Matutes, E.; Morilla, R.; Strefford, J. C.; Catovsky, D. *British journal of haematology* **2016**, *174*, 767-775.
- (60) Hüttemann, M.; Pecina, P.; Rainbolt, M.; Sanderson, T. H.; Kagan, V. E.; Samavati, L.; Doan, J. W.; Lee, I. *Mitochondrion* **2011**, *11*, 369-381.
- (61) Pully, V. V.; Lenferink, A. T. M.; Otto, C. *Journal of Raman Spectroscopy* **2011**, *42*, 167-173.
- (62) Uzunbajakava, N.; Lenferink, A.; Kraan, Y.; Willekens, B.; Vrensen, G.; Greve, J.; Otto, C. *Biopolymers: Original Research on Biomolecules* **2003**, *72*, 1-9.

#### **AUTHOR INFORMATION**

Corresponding Author

\* Piot Olivier, BioSpecT EA n°7506, Faculty of Pharmacy, University of Reims Champagne-Ardenne, 51 rue Cognacq-Jay, 51096 Reims Cedex, France. Email address [olivier.piot@univ-reims.fr](mailto:olivier.piot@univ-reims.fr)

#### **Acknowledgements**

The authors would like to acknowledge the Agence Nationale de la Recherche (ANR) and the European Community for financial support.

# **CHAPITRE 5 : APPLICATION À L'ECHELLE CELLULAIRE**

## Contexte

Comme vu dans le préambule le système immunitaire est constitué de différents types de cellules tels les lymphocytes B, T et les Natural Killer (NK), qui fournissent une barrière efficace contre les éléments potentiellement pathogènes. Un changement dans la numération sanguine de ces cellules peut indiquer la présence d'une pathologie, d'une bactérie, d'une virose, d'un parasite ou d'un cancer.

Actuellement, les cellules du système immunitaire sont détectées par différentes méthodologies qui peuvent être plus ou moins agressives et qui peuvent altérer l'état natif des cellules étudiées. L'analyse par imagerie morphologique, couramment utilisée, passe par une étape de fixation et une coloration chimique avant l'observation au microscope, ce qui peut altérer leur morphologie. De plus une fois colorés, ces échantillons ne peuvent plus être utilisés pour une autre modalité. L'imagerie morphologique reste limitée en sensibilité, car les lymphocytes B et T sont étroitement apparentés et morphologiquement très similaires. La cytométrie en flux utilise des anticorps fluorescents pour marquer les cellules d'intérêts. Cette technique peut potentiellement altérer la composition biochimique des cellules à cause des contraintes biomécaniques auxquelles elles sont confrontées lors de la préparation et de l'analyse. La spectroscopie Raman permettrait l'utilisation et la manipulation de cellules immunitaires n'ayant subi aucune modification et restant dans leur état presque natif. De plus, après l'analyse les échantillons peuvent être réutilisés pour une autre modalité.

Dans le chapitre précédent, il a été montré que la spectroscopie Raman couplée à des algorithmes de prétraitements et d'analyse de données, peut-être une alternative ou une méthode complémentaire efficace par rapport aux techniques conventionnelles, ceci afin de poser un diagnostic de la LLC à l'échelle du patient dans le cas des stades avancés. Malgré ces résultats encourageants cette méthodologie ne permet pas de détecter les cas de LLC peu avancés où le nombre de cellules cancéreuses reste minoritaire. Pour aborder ce point, il est nécessaire de déterminer le type et le statut de chacune des cellules sondées par la spectroscopie Raman. Pour cela, il est primordial de déterminer une méthode d'identification (« Gold standard ») sur les mêmes frottis cellulaires que ceux utilisés pour l'analyse Raman.

## Objectif

Le but de ce chapitre est de développer un classifieur Raman permettant d'identifier les sous-types lymphocytaires (B vs T, NK) et de détecter les cellules tumorales (B tumoraux). Deux stratégies de labellisation (gold standard) ont été suivies :

La première méthodologie a consisté en l'immunocoloration des frottis a posteriori, c'est-à-dire après avoir réalisé les acquisitions des spectres Raman, une coloration des cellules par immunohistochimie est réalisée au moyen d'anticorps spécifiques afin d'identifier les sous-types cellulaires. Dans le cadre de cette stratégie, seule la population minoritaire a été marquée afin de minimiser l'erreur liée à un immunomarquage faible. Ainsi le marqueur de CD20+ a permis d'identifier les lymphocytes B sur des patients non atteints de LLC et CD2+ a permis de marquer les lymphocytes T et NK sur les patients atteints de LLC. La seconde stratégie repose sur l'acquisition de spectres Raman de populations de lymphocytes issue de tri cellulaire avec un taux de pureté supérieur à 95%.

### 1. Stratégie de labellisation par immunohistochimie

Dans cette stratégie de marquage a posteriori, le délai entre la préparation du frottis et le marquage est un paramètre critique. Nous avons pu mettre au point un protocole d'immunocoloration (CD2 ou CD20) selon le type cellulaire ciblé sur des cellules fixées par un mélange de méthanol et acétone.

#### 1.1. Mesures des spectres Raman

Pour cette étude, nous avons repris les échantillons analysés dans le cadre de la campagne de mesures M3S. Pour chaque patient, les échantillons ont été fixés dans un mélange 50% MeOH – 50% Acétone juste après l'analyse Raman, puis rapidement immunocolorées selon le protocole décrit en annexe 1. Cette immunocoloration a permis aux cliniciens d'attribuer un label à chaque cellule analysée, permettant de les identifier sur les frottis sanguins.

Pour rappel, toutes les mesures spectrales Raman ont été recueillies à l'aide d'une diode laser 532 nm fournissant 13 mW de puissance à l'échantillon avec un objectif 150x/NA 0,95, et avec une résolution spectrale de  $4 \text{ cm}^{-1}$  grâce à un réseau holographique de 1200 lignes/mm. Les spectres ont été enregistrés avec un temps d'intégration de 10 secondes et 1 accumulation en 3 points distincts au niveau du noyau cellulaire séparés d'environ  $1 \mu\text{m}$ .

Une cohorte de 81 patients, soit 33 sains et 48 LLC, a été analysée. Pour chacun des patients, entre 60 et 100 cellules différentes ont été ciblées, soit un total de 6897 cellules. Pour les sous-ensembles B et NK+T, un total de 5161 lymphocytes ont été analysés, soit 2100 lymphocytes NK+T et 3061 lymphocytes B. Pour les sous-ensembles B sains et B LLC, un total de 3836 lymphocytes ont été analysés, soit 650 cellules B saines et 3186 cellules B LLC. Seules 650 cellules B saines ont été analysées en raison de la difficulté à isoler une concentration élevée de ces lymphocytes sur un frottis sanguin de patients non-LLC (population minoritaire). Les analyses Raman ainsi que l'immunomarquage ont été effectués le même jour.

## **1.2. Prétraitement des mesures Raman**

Le prétraitement des cellules a été effectué selon le « prétraitement 2 », comprenant un lissage de Savitsky Golay et une correction par EMSC, le tout guidé par un spectre moyen de référence. La correction spectrale du verre a été effectuée à l'aide d'une matrice d'interférence du verre.

## **1.3. Similarité des signatures Raman entre les sous-groupes lymphocytaires**

Dans un premier temps, les moyennes des spectres Raman de chaque groupe ont été comparées afin d'identifier visuellement d'éventuelles différences. Sur la Figure 40, aucune différence entre les spectres n'est à noter que ce soit sur les basses fréquences de 700 à 1800  $\text{cm}^{-1}$  ou les hautes fréquences de 2800 à 3100  $\text{cm}^{-1}$ , entre les données spectrales des différents groupes. Dans le but de mettre en évidence des différences spectrales plus subtiles qui permettraient de distinguer chaque groupe, une méthode de classification non supervisée, basée sur l'analyse en composantes principales (ACP) a été utilisée.

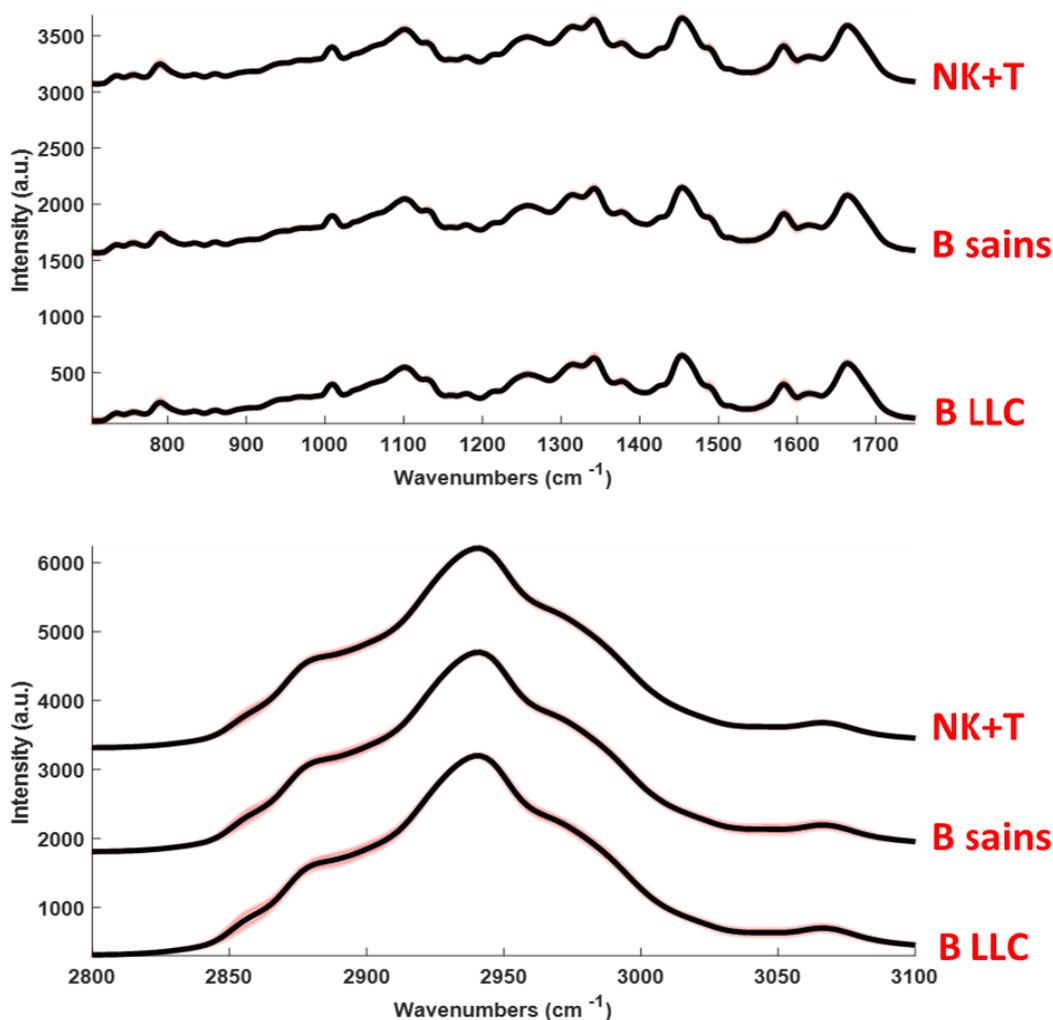


Figure 40 - Moyenne et écart-type des différents sous-type lymphocytaire

Dans ce travail, l'analyse en composantes principales a été appliquée sur les données labellisées provenant de la campagne M3S et comprenant les différents sous-types lymphocytaires NK+T, B sains et B LLC. La Figure 41 présente la projection selon les trois premières composantes principales des scores pour les trois sous-groupes de lymphocytes. Comme le montre la figure, il est possible d'observer un regroupement compact du nuage de points. Ces trois premières composantes principales représentent environ 99% de la variance cumulée, soit 95% pour PC1, 3% pour PC2 et 2% pour PC3.

Les différents types cellulaires n'ont pas pu être classés avec l'ACP sur la base des trois premières composantes principales. En effet, l'ensemble des données provenant des trois sous-groupes lymphocytaires se chevauchent considérablement. Ces résultats reflètent que les signatures Raman des groupes B sains, B LLC et NK+T possèdent de grandes similitudes.

Dans la littérature, certaines études [147-149] ont montré que l'ACP est une méthode exploratoire des données efficace pour mettre en évidence les différents sous-groupes lymphocytaires. Dans notre étude, les conditions expérimentales sont totalement différentes afin de se rapprocher le plus possible des conditions cliniques de routine (frottis sanguins séchés à l'air et sur lame de verre).

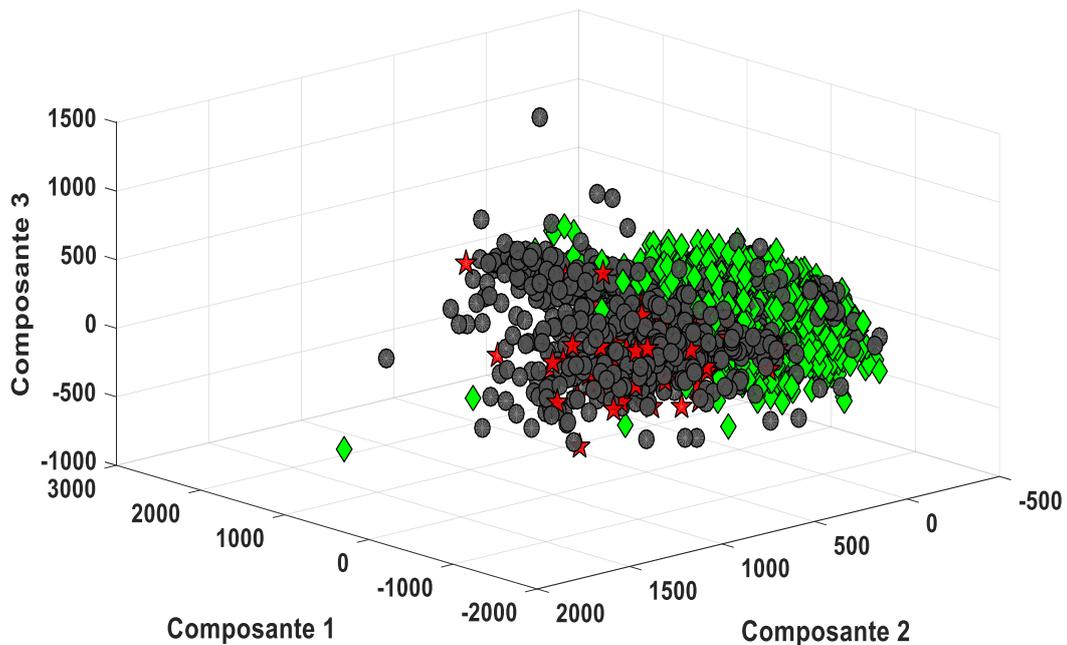


Figure 41 - Projection 3D de l'ensemble des scores provenant de l'ACP pour les trois sous-groupes de lymphocytes

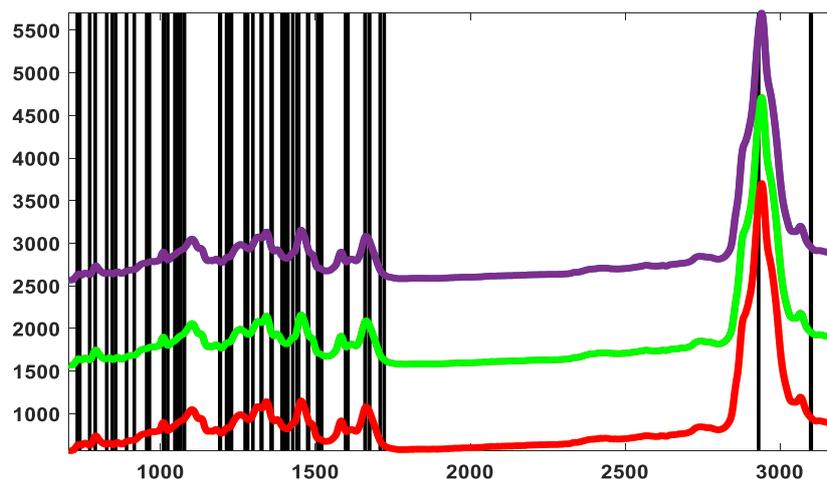
#### 1.4. Développement de modèles de reconnaissance lymphocytaire et paramètres d'optimisation

Dans le chapitre 3, il a été montré qu'il est essentiel de choisir soigneusement la stratégie d'entraînement afin d'optimiser au mieux les modèles, éviter le sur-apprentissage (over-fitting) ou limiter la robustesse du modèle en utilisant des petits jeux d'entraînement. Dans cette étude, compte tenu de la taille de l'ensemble de données, une stratégie de validations croisées répétées a été utilisée (rdCV) et les classes ont été automatiquement équilibrées. Les données d'apprentissage pour chacun des modèles ont été construites en utilisant 70 % de l'ensemble de données, tandis que 30 % des données ont servi pour la validation. Les classifications ont été effectuées au total 100 fois avec un tirage de données aléatoires sur la formation et la validation pour chaque itération de la boucle externe de rdCV.

Les paramètres pour chacun des modèles multivariés ont ensuite été optimisés en choisissant les paramètres qui ont permis d'obtenir la meilleure performance pour chacune des classifications. Chaque algorithme a été optimisé en fonction de ses paramètres respectifs (nombre de variables latentes pour PLS-DA ou combinaison de paramètres pour SVM et RF). Deux stratégies ont été testées : i) simultanée (C1) qui consiste à construire des modèles qui sont capables de prédire simultanément les 3 classes recherchées NK+T, B sains et B LLC, ii) séquentielle (C2) qui consiste à entraîner indépendamment deux types de modèles NK+T vs B et B sain vs B LLC.

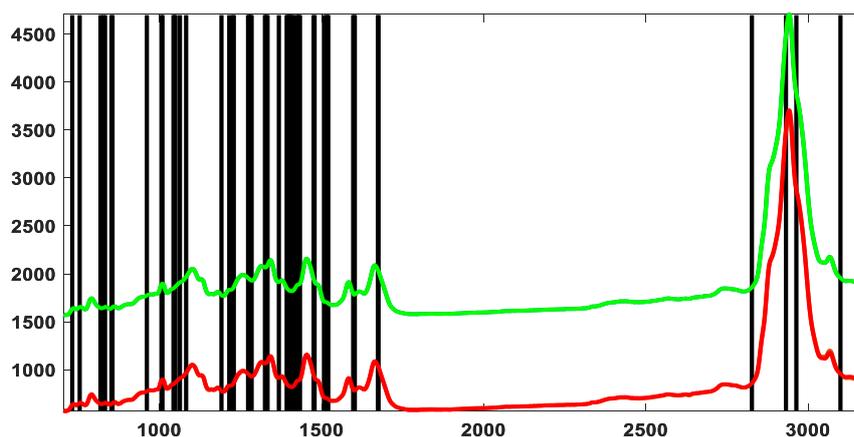
#### **1.4.1. Discrimination des sous-types de lymphocytes**

Pour développer une méthode de séparation des groupes de lymphocytes NK+T, B sains et B LLC. Nous avons utilisé la même méthodologie que précédemment, c'est-à-dire maximiser les différences et la séparabilité entre les groupes par la méthode de réduction de dimensions supervisée décrite dans le « matériel et méthode ». Celle-ci a été appliquée sur l'ensemble des données, à l'aide des labels cellulaires afin d'identifier les nombres d'ondes différenciant les sous-groupes. Cette méthode de réduction de dimensions supervisée a permis la simplification des spectres en utilisant seulement quelques variables discriminantes représentant la variance maximale entre les groupes pour chaque modèle optimisé. Les raies Raman présentant une différence significative entre les sous-ensembles de cellules sont visualisées par les barres verticales noires sur les figures 42, 43 et 44. Dans le cadre de la stratégie C1, c'est-à-dire l'élaboration de modèles tri-classes, les 50 premières longueurs d'onde les plus représentées sur l'ensemble des 100 modèles optimisés de la rdCV ont été sélectionnées et représentées sur la Figure 42.



*Figure 42 - Spectre moyen pour les lymphocytes NK+T (bleu), B sain (vert) et B LLC. Région d'intérêt, composée de 50 nombres d'onde les plus caractéristiques, permettant de différencier les sous-groupes lymphocytaires.*

La stratégie C2 a consisté à développer des modèles séquentiels dans le but de séparer deux classes à chaque étape, dans un premier temps des modèles permettant de séparer les groupes NK+T et B sont construits afin d'isoler les lymphocytes B d'intérêt, puis d'autres modèles sont entraînés afin de réaliser la séparation B sains et B LLC. Les figures 43 et 44 montrent les nombres d'onde sélectionnés pour les distinctions NK+T versus B et B sains versus B LLC.



*Figure 43 - Spectre moyen pour les lymphocytes NK+T (rouge) et B (vert). Région d'intérêt, composée de 50 nombres d'onde les plus caractéristiques, permettant de différencier les sous-groupes lymphocytaires*

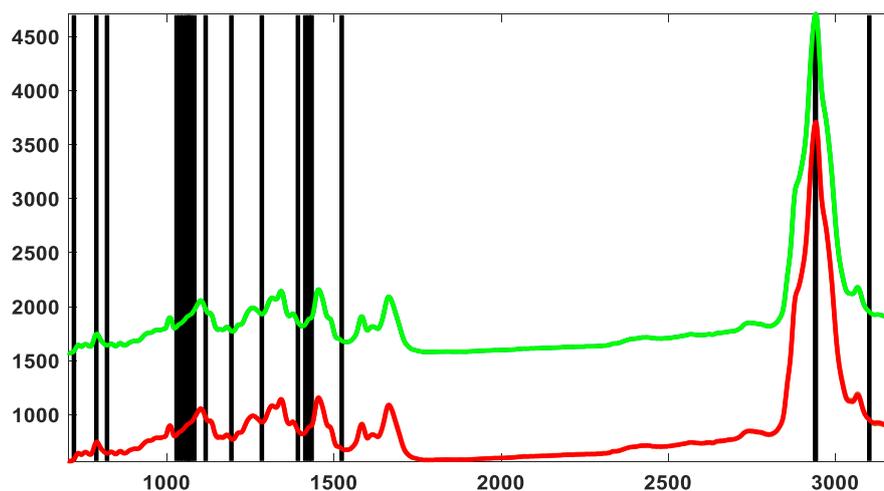


Figure 44 - Spectre moyen pour les lymphocytes B sains (rouge) et LLC (vert). Région d'intérêt, composée de 50 longueurs d'onde les plus caractéristiques, permettant de différencier les sous-groupes lymphocytaires.

En observant les zones sélectionnées par traitement statistique, il est possible d'accéder à certaines informations sur les différences de composition chimique entre les populations lymphocytaires. Il a été possible grâce à la littérature [149, 150], d'attribuer pour la stratégie C2, les raies spectrales les plus discriminantes sélectionnées. Les raies spectrales permettant de séparer les cellules NK+T et B peuvent être identifiées dans le tableau 5.

Tableau 5 - Tentative d'attribution des bandes spectrales permettant de séparer les lymphocytes NK+T et B

Raman shift (cm-1)	Groupements
645	torsion C-C : tyrosine
671	étirement C-S : cystéine
725	respiration ring : adénine
800	étirement symétrique O-P-O
1004	respiration ring : phénylalanine
1097	étirement symétrique O-P-O
1129	étirement C-C : lipides

1209	étirement phénylalanine / tyrosine / C-N
1259	amide III
1304	adénine / amide III
1345	chaîne polynucléotidique
1378	thymine / adénine / guanine
1455	déformation CH <sub>2</sub> : lipides
1585	adénine / guanine
1665	amide I
2835	protéines / lipides
2874	

Les raies spectrales identifiées permettant de séparer les lymphocytes B sain et B LLC, sont attribuées principalement aux bandes spectrales liées à l'ADN et aux lipides/protéines. Pour l'ADN, nous retrouvons les bandes 796 cm<sup>-1</sup>, 1106 cm<sup>-1</sup>, 1380 cm<sup>-1</sup> et 1585 cm<sup>-1</sup>, ceci pourrait indiquer une différence de concentration d'ADN entre les cellules B LLC et les cellules B saines. Dans la région haute fréquence, la zone à 2922 cm<sup>-1</sup> et 2954 cm<sup>-1</sup> peut être attribuée à l'étirement CH des lipides/protéines et CH<sub>3</sub> respectivement, indiquant une variation d'intensité entre les deux groupes. Les bandes 1429-1468 cm<sup>-1</sup>, 2874 cm<sup>-1</sup> et 2835 cm<sup>-1</sup> sont corrélées aux protéines et aux lipides. La bande à 1007 cm<sup>-1</sup> est liée à la phénylalanine, un léger décalage en fréquence de cette bande peut être observé.

#### 1.4.2. Évaluation de la stratégie C1

Afin d'évaluer les performances de la spectroscopie Raman à distinguer les différents sous-groupes lymphocytaires, la méthode de réduction de dimensions a été associée à la méthode de validation croisée rdCV et aux trois algorithmes de classification SVM, RF et PLSDA. Cent modèles ont été générés à partir des spectres Raman provenant des cellules caractérisées à posteriori par immunocoloration.

La Figure 45 présente la moyenne et l'écart-type des taux de classifications pour chacune des méthodes SVM, RF et PLS-DA sur les 100 runs de la boucle externe de rdCV. Le nombre total de cellules dans le jeu d'entraînement et le jeu de validation est présenté dans le tableau 6.

*Tableau 6 - Nombre total de cellules dans le jeu d'entraînement et le jeu de validation*

<b>Echantillon</b>	<b>Set d'entraînement</b>	<b>Validation externe</b>
<b>NK+T</b>	1470 cellules	630 cellules
<b>B sains</b>	455 cellules	195 cellules
<b>B LLC</b>	1600 cellules	723 cellules

Afin d'évaluer les performances de chaque algorithme de classification, la performance globale ou « accuracy » en anglais a été calculé en faisant la moyenne des taux de prédictions pour chaque algorithme de classification. Ainsi en validation interne, RF et PLS-DA possèdent la performance globale la plus haute avec un pourcentage d'environ 57%. SVM est un peu moins performant avec un taux de 50%. En validation externe, la performance globale avoisine les 55% pour les trois méthodes. Malgré que la PLS-DA possède la meilleure performance pour estimer la classe NK+T avec un taux moyen interne et externe de 57 % contre 35% pour les deux autres classifieurs, les performances générales sont très mauvaises que ce soit pour SVM, RF ou PLS-DA. Malgré l'utilisation de trois méthodes de classification différentes allant de la méthode linéaire à des méthodes permettant d'évoluer dans des espaces de données plus complexes, il a possible de constater des performances globales médiocres. Ces résultats démontrent la complexité des données Raman liées aux lymphocytes et viennent appuyer les résultats de l'ACP selon lesquelles la composition biochimique et la signature Raman des différents groupes de lymphocytes sont très similaires, ce qui complexifie le problème de discrimination. Tout ceci montre que la stratégie simultanée tri-classe n'est donc pas adaptée dans le cadre de notre étude.

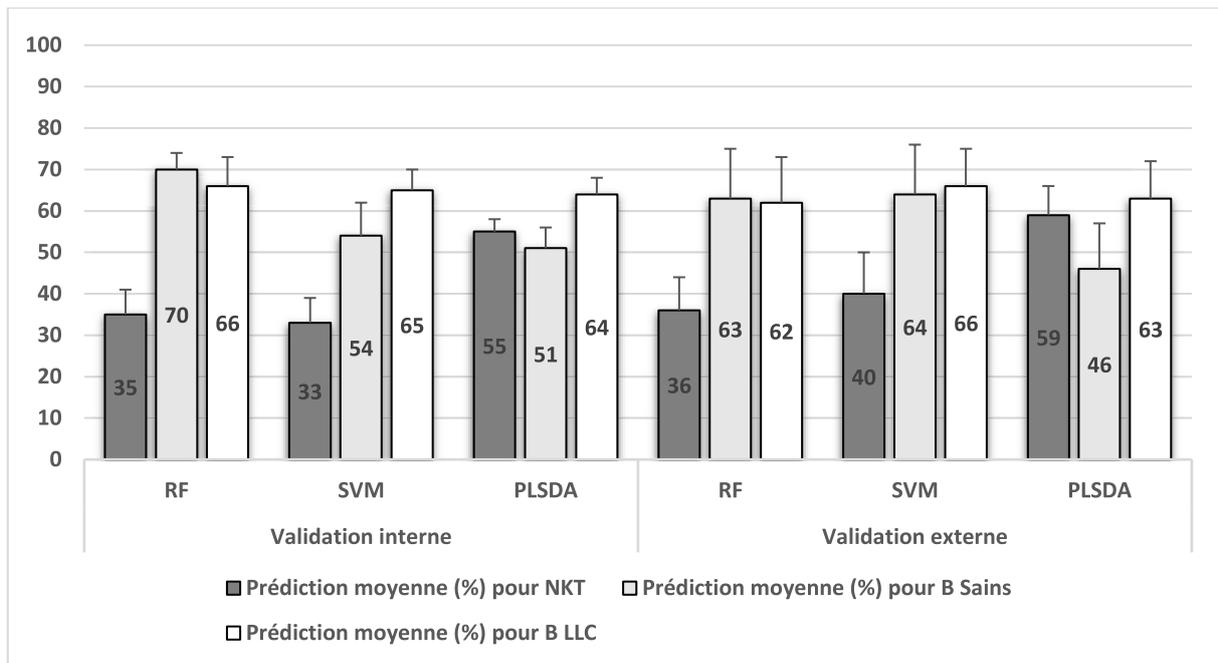


Figure 45 - Comparaison des différents algorithmes de classification pour le modèle tri-classe

### 1.4.3. Evaluation de la stratégie C2

Le problème de classification simultanée multi groupes présentée précédemment peut être résolu en utilisant une stratégie séquentielle, c'est-à-dire de décomposer le problème en sous-problème binaire plus facile à résoudre. Dans le cas de notre étude, afin d'obtenir une discrimination des sous-groupes, des modèles ont été construits sur la base de la stratégie de classification actuelle, c'est-à-dire la méthode rdCV et l'optimisation de 100 modèles. Ainsi l'approche a été décomposée en construisant deux types de modèles de prédiction indépendants : dans un premier temps, discrimination de cellules NK+T versus B puis dans un second temps, discrimination des lymphocytes B sains versus LLC. Les données ont été réduites grâce à la méthode de réduction supervisée qui a permis de sélectionner les zones discriminantes pour chaque étape de classification. Le nombre de cellules en entraînement et en validation est présenté dans le tableau 7. Cette approche a permis d'isoler par étapes chaque sous-groupe lymphocytaire, de la plus générale (NK+T versus B) à la plus spécifique (B sain versus B LLC). La méthode de validation croisée rdCV associée au 10-K-Fold a permis d'optimiser la performance et la robustesse des différents modèles construits ainsi que la précision de la prédiction.

Tableau 7 - Nombre de cellules NK+T, B et B sains et B LLC en entraînement et validation externe

<b>Nombre de cellules NK+T et B en entraînement et validation</b>		
<b>Echantillon</b>	<b>Set d'entraînement</b>	<b>Validation externe</b>
<b>NK+T</b>	1470 cellules	630 cellules
<b>B</b>	1489 cellules	637 cellules
<b>Nombre de cellules B sains et B LLC en entraînement et validation</b>		
<b>B sains</b>	455 cellules	195 cellules
<b>B LLC</b>	1600 cellules	723 cellules

#### 1.4.4. Validation interne

Les sensibilités et spécificités moyennes ont été calculées par paires de groupes et représentées sur les Figures 46 et 47. Comme précédemment, afin d'évaluer les performances globales de chaque algorithme de classification, nous avons moyenné les sensibilités et spécificités de chaque algorithme. Ainsi, les performances globales pour RF sont de 72% dans les classes NK+T versus B et de 70% pour B sains versus B LLC, pour SVM de 79% dans les classes NK+T versus B et 73% dans les classes B sains versus B LLC et pour PLS-DA de 74% dans les classes NK+T versus B et 71% dans les classes B sains versus B LLC. À la vue de ces résultats, en validation interne, le SVM offre de meilleures performances afin de différencier les classes NK+T versus B, suivi de près par PLS-DA puis RF. Pour le cas des classes B sains versus B LLC, les trois algorithmes présentent des performances quasi similaires.

#### 1.4.5. Validation externe

En validation externe, les performances globales par algorithmes de classification sont pour RF de 69% dans les classes NK+T versus B et B sains versus B LLC, pour SVM de 67% dans les classes NK+T versus B et 73% dans les classes B sains versus B LLC et pour PLS-DA de 73% dans les classes NK+T versus B et 70% dans les classes B sains versus B LLC (Figure 46 et Figure 47). Pris ensemble, ces résultats montrent de meilleures performances pour la PLS-DA afin de différencier les classes NK+T versus B. En ce qui concerne les classes B sains versus B LLC, SVM et PLS-DA possèdent les meilleures performances. En comparant les

lymphocytes NK+T aux lymphocytes B et les lymphocytes B sains avec lymphocytes B LLC, il est possible de voir que SVM a montré des performances légèrement supérieures à PLS-DA pour distinguer B sains de B LLC. PLS-DA reste relativement performante pour distinguer à la fois NK+T/B et B sains/B LLC.

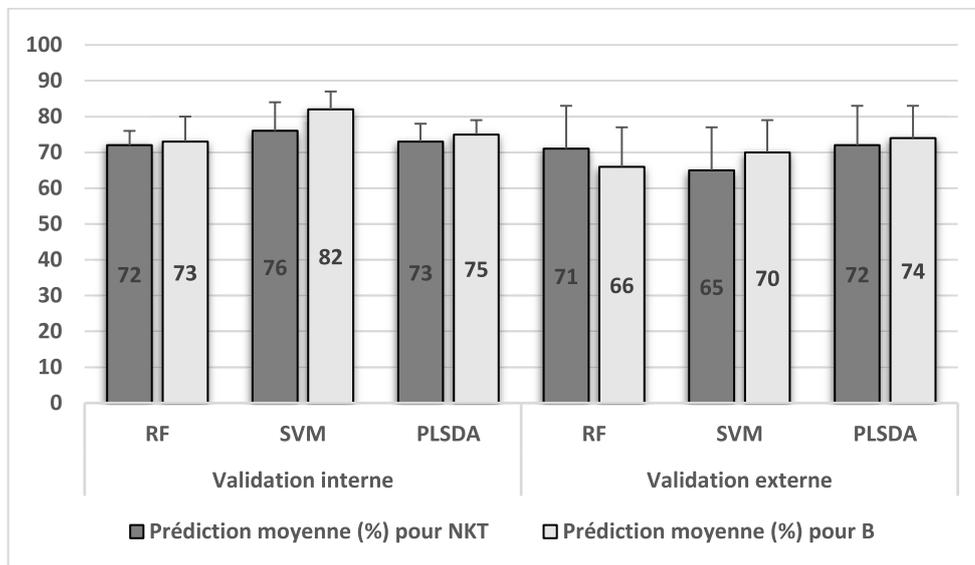


Figure 46 - Comparaison des différents algorithmes de classification pour le modèle bi-classe NK+T vs B

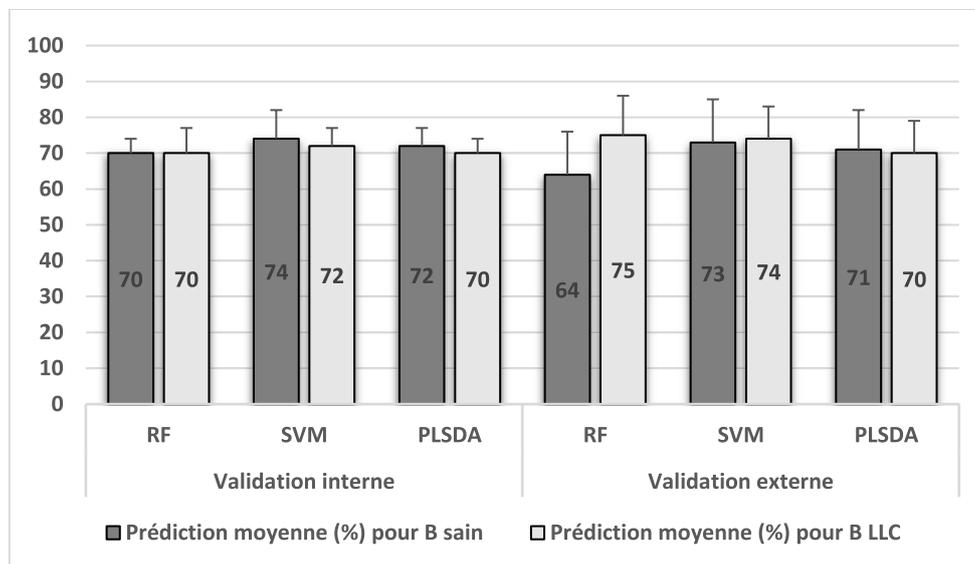


Figure 47 - Comparaison des différents algorithmes de classification pour le modèle bi-classe B sains vs B LLC

Certaines études ont montré le potentiel de la spectroscopie Raman à discriminer les cellules cancéreuses [151], en hématologie, il a été montré que il était possible de discriminer les type des cellules immunitaires [149, 152-159], mais également leur état physiologique [160,

161]. D'autres études ont permis de mettre en avant les modifications biochimiques liées à l'activation ou non des lymphocytes [153]. Pour des données aussi complexes que les spectres Raman de lymphocytes, il est important d'être judicieux dans la préparation des données, le choix de l'algorithme de classification et de la stratégie d'optimisation, ceci est la clé pour l'élaboration de modèles robustes. Il est également important de choisir un gold standard fiable et reproductible. Dans cette partie, il a été démontré pour trois types d'algorithmes, PLS-DA, SVM et RF, qu'il était possible de résoudre un problème de classification multiclasse difficile, lié à la discrimination des sous-classes lymphocytaires chez un patient donneur sain ou atteint de la LLC. Chaque modèle de prédiction cellulaire a été créé grâce à une labellisation précise des lymphocytes.

Dans le cadre de cette étude, deux stratégies ont été utilisées : une simultanée (C1) et une séquentielle (C2). Tous les modèles issus de la stratégie simultanée (C1) se sont montrés relativement mauvais pour séparer les différentes classes simultanément que ce soit avec l'algorithme linéaire PLS-DA ou les algorithmes non linéaires SVM et RF. La stratégie séquentielle a montré des performances beaucoup plus importantes que la méthode C1. Pour cette stratégie, en validation interne, Le SVM offre de meilleures performances afin de différencier les classes NK+T versus B, suivi de près par PLS-DA puis RF. Pour le cas des classes B sains versus B LLC, les trois algorithmes présentent des performances quasi similaires. En validation externe SVM a montré des performances légèrement supérieures à PLS-DA pour distinguer B sain de B LLC. PLS-DA reste relativement performant pour distinguer à la fois NK+T/B et B sains/B LLC. RF a montré des performances globales inférieures à SVM et PLS-DA. Nous pouvons en conclure qu'une stratégie binaire séquentielle couplée à une labellisation par immunohistochimie et un algorithme d'apprentissage approprié, constitue une approche prometteuse pour entraîner des modèles et distinguer les sous-groupes lymphocytaires malgré la complexité du problème.

Toutefois, un problème peut être encore soulevé concernant la fiabilité et la reproductibilité de l'immunocoloration comme gold-standard. En effet, des erreurs liées à une coloration défectueuse du frottis ou une mauvaise labellisation des cellules par le clinicien peuvent influencer sur l'efficacité des modèles construits. Des alternatives peuvent être envisagées comme le tri par billes magnétiques ou le tri par cytométrie en flux inverse. Ces approches consistent à marquer les cellules non souhaitées pour les éliminer et ne garder que les cellules d'intérêt. Dans le cadre de cette thèse, nous avons opté pour un tri magnétique car c'est une méthode simple et rapide à mettre en place.

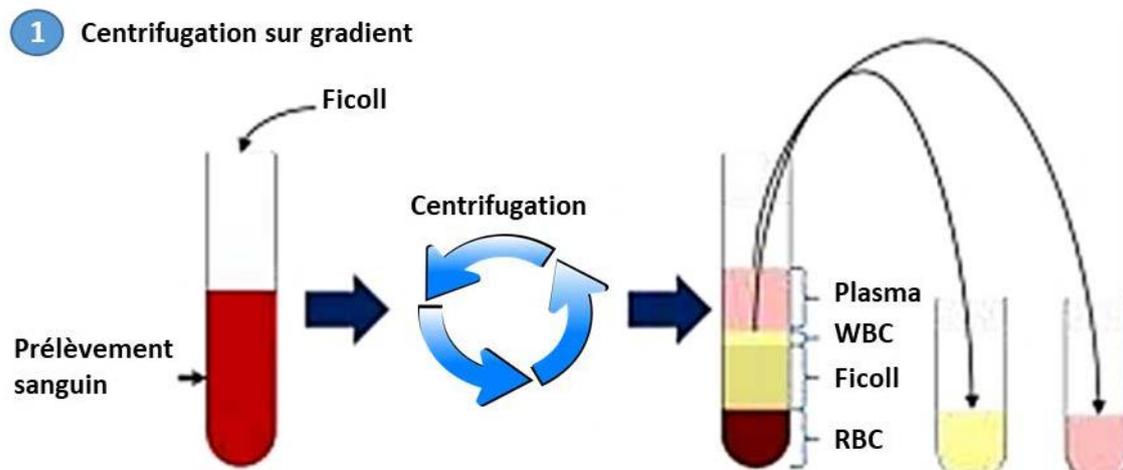
## 2. Résultats complémentaires : études pilotes

### 2.1. Tri cellulaire magnétique

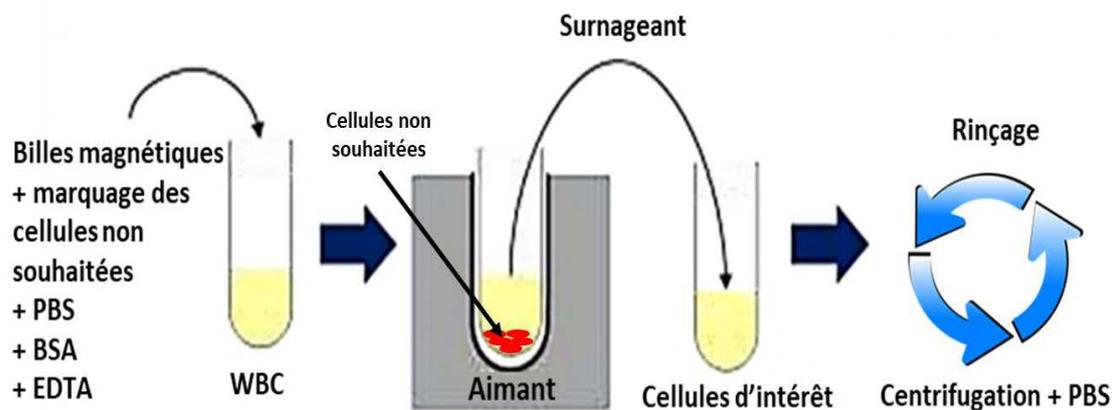
L'avantage du tri cellulaire est qu'il permet d'obtenir plus de données Raman sur les cellules B des patients sains difficiles à recueillir en aveugle sur un frottis sanguin. Grâce à cette technique, il a été possible d'extraire des lymphocytes T avec un niveau de pureté allant jusqu'à 96%, les cellules B ont été obtenues à un niveau de pureté allant jusqu'à 98% pour les B LLC et 94% pour les B sains. Trois spectres Raman ont été acquis sur le noyau de 50 à 100 cellules, provenant de 11 donneurs sains et 7 LLC. Pour chaque cellule, les trois spectres Raman ont été moyennés. Au total, 1005 spectres moyens ont été recueillis soit 195 lymphocytes NK, 408 T, 292 B sains et 110 B LLC.

#### 2.1.1. Acquisition des spectres sur les tris cellulaires magnétiques

Sur les cellules triées, la production de données se fait en deux étapes : la préparation des lames au CHU MGD et l'acquisition des données (images et spectres) à l'URCA. Après ajout de Ficoll (polysaccharide) le prélèvement sanguin est centrifugé ce qui permet de récupérer le plasma et les leucocytes par densité. Dans le tube contenant les leucocytes, nous enlevons ensuite les cellules non pertinentes par ajout de billes magnétiques porteuses d'anticorps. Les cellules d'intérêt sont finalement rincées par centrifugation en utilisant du PBS puis une partie est étalée sur deux lames. Les lames sont vernies avec un mélange 50/50 plasma du patient/PBS. La figure 48 décrit le processus. Les cellules restantes sont passées en cytométrie de flux pour déterminer le taux de pureté obtenu. Une lame est envoyée à l'URCA l'autre est contrôlé en coloration MGG directement au CHU MGD.



## 2 Tri cellulaire



## 3 Etalement

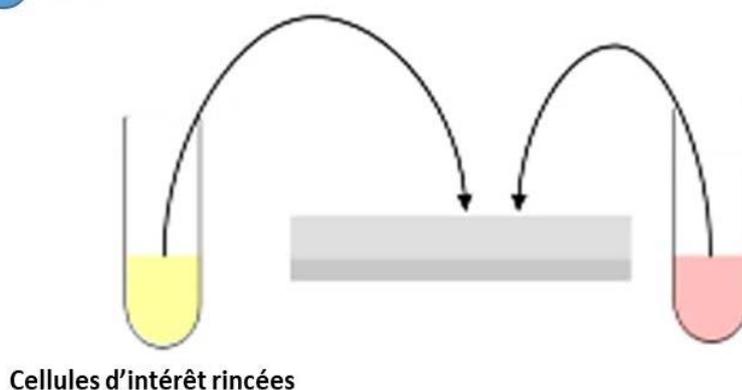


Figure 48 - Protocol du tri magnétique des différents sous-types cellulaires

### 2.1.2. Élaboration de modèles de prédiction

Au vu du nombre limité de cellules pour ce test préliminaire, nous avons utilisé la sélection de variables supervisée couplée à une leave-one-patient-out-crossvalidation (LOOPCV) afin d'optimiser les modèles de prédictions. La méthode séquentielle C2 a été couplée à l'algorithme PLS-DA afin de créer des modèles de classification. Le processus a été lancé 100 fois avec un choix aléatoire des données d'entrée entre chaque tour du processus. Les sensibilités et les spécificités moyennes ont ensuite été calculées pour chaque classe de lymphocytes et sont résumées dans le tableau 8. Pour chaque étape, les résultats préliminaires montrent des performances de prédiction encourageantes qui varient de 75% à 80% pour chaque classe. Sur l'ensemble des deux modèles, la classe NK+T possède des performances moyennes légèrement inférieures aux autres avec 75%.

*Tableau 8 - Performances des différentes pour chaque paire de lymphocytes*

<b>Tris magnétiques (méthode séquentielle)</b>			
<b>Etape 1</b>		<b>Etape 2</b>	
<b>T, NK</b>	<b>B</b>	<b>B sains</b>	<b>B LLC</b>
75 % ( $\pm 10$ )	82 % ( $\pm 5$ )	79 % ( $\pm 7$ )	77 % ( $\pm 8$ )

Malgré un petit jeu de donnée, ces résultats préliminaires semblent vraiment très prometteurs. Afin d'avoir une évaluation beaucoup plus pertinente, il est indispensable d'augmenter le nombre de patients et de cellules analysées. Une seconde chose importante serait d'évaluer la transférabilité des modèles construits grâce aux cellules triées par tri magnétique sur les frottis sanguins. En effet, il a été montré dans ce mémoire que les conditions expérimentales et la préparation des échantillons pouvaient influencer fortement sur le signal Raman. Donc, il se peut que l'utilisation de cellules triées comme base pour entraîner des modèles de classification ne soit pas approprié. Malgré cela, le prétraitement par EMSC pourrait éventuellement permettre une homogénéisation des données afin de rendre transférables les données de tri magnétique.

## **2.2. Numération spectrale semi-supervisée par algorithme mémétique**

Dans le cadre de cette étude pilote, les proportions des sous-populations de lymphocytes (T, B et NK) pour chaque patient ont été déterminées à l'aide de l'analyse MFC. Les échantillons de sang ont été prélevés dans des tubes K2 EDTA et le MFC a été exécuté sur l'appareillage FACS Canto-II (Becton Dickinson). Les marqueurs CD19, CD20, CD5, kappa, lambda, CD10, CD3, CD4, CD4, CD8 et CD56 ont été utilisés afin d'identifier les différents types de lymphocytes. L'acquisition et l'analyse des données ont été effectuées à l'aide du logiciel DiVa (Becton Dickinson). Pour chaque échantillon analysé en Raman sur la plate-forme M3S, la numération des lymphocytes B, T et NK a été déterminée.

Afin d'estimer les proportions de sous-populations de lymphocytes à l'aide des signatures spectrales, une méthode de classification semi-supervisée basée sur un algorithme mémétique (CAM) [162] a été utilisée. Les algorithmes mémétiques sont des méthodes d'hybridation entre les algorithmes métaheuristiques (colonies de fourmis, essaims

particulaires, algorithmes génétiques...), et les algorithmes de recherche locale comme le K-means qui permet une convergence plus rapide de l'algorithme.

La méthode utilisée dans cette étude est le couplage entre un algorithme génétique (AG) qui a servi à réaliser la sélection de variables et un K-Means pour regrouper les différents types de cellules entre elles et accéléré la convergence de l'AG. Le principe général de l'algorithme repose sur quatre étapes fondamentales. La première consiste à générer une population 1 de  $P = 1000$  chromosomes où 1 chromosome est composé de 70 gènes, correspondant à 70 nombres d'ondes tirées aléatoirement. Parmi ces  $P$  chromosomes,  $P/2$  parents sont sélectionnés. La seconde étape consiste à réaliser des croisements aléatoires (cross-over) entre couple de chromosomes parents, amenant à la construction de chromosomes enfants constitués de parties des deux parents. A l'issue de cette étape, une nouvelle population 2 est engendrée contenant  $P/2$  chromosomes parents et  $P/2$  chromosomes enfants. Dans la troisième étape, chacun des chromosomes de la population 2 subit une mutation aléatoire de 5% de ses gènes. Cette étape a pour but de limiter une convergence prématurée de l'algorithme. La quatrième étape consiste à réaliser une recherche locale par KM afin d'accélérer la convergence de l'algorithme génétique. Les quatre étapes sont répétées  $M$  fois. A l'issue de la  $M^{\text{ème}}$  population, le chromosome (70 longueurs d'onde) et les centroïdes optimaux sont retenus. Ces longueurs d'ondes et ces centroïdes serviront de point de départ afin de prédire de nouveaux échantillons. Ces méthodes ont montré leur efficacité dans plusieurs domaines de recherche tels que la reconnaissance de forme mais sont très peu utilisées dans le domaine des spectroscopies vibrationnelles [163].

Pour notre étude, cette méthode semi-supervisée a été guidée par la numération lymphocytaire de chaque patient. L'entraînement du modèle a été effectué sur 6 patients LLC ayant une hyperlymphocytose B supérieure à 70% et 1 patient LLC intermédiaire ayant un taux de lymphocytes B de 56%. Ce modèle a été testé sur 16 patients externes au jeu d'entraînement et les résultats ont été comparés à la numération de référence. Pour chaque cellule de chaque patient, l'algorithme renvoie un label de classe. Le tableau 9 présente les résultats obtenus après classification. Malgré quelques divergences sur certains patients, les résultats présentent une tendance intéressante.

À l'heure actuelle, la méthodologie possède une faiblesse notable qui est le nombre de cellules analysées par patient en Raman. En effet, le programme est guidé par la numération qui lui sert de référence pour construire le modèle de prédiction, or cette numération est réalisée sur un nombre de plusieurs milliers de cellules et l'estimation en Raman est faite sur seulement

quelques centaines de cellules. Cette différence de cellules prises en compte peut engendrer une erreur importante entre la prédiction et la réalité. Il est donc primordial d'augmenter le nombre de patient et de cellules analysées par patient.

In fine, cette méthodologie pourrait devancer largement les autres techniques de labellisation. En effet, cette méthode possède une double qualité, elle est entièrement automatisée, la labellisation des cellules s'effectue de façon autonome sans intervention humaine ce qui permettrait de réduire de façon considérable le pourcentage d'erreur lié à une mauvaise labellisation du clinicien. De plus, l'algorithme est capable d'estimer une numération lymphocytaire grâce aux spectres Raman et pourrait permettre au clinicien de poser également un diagnostic.

Tableau 9 - Résultats issus de l'algorithme génétique

Numéro patient	Theorique			Estimé			
	pcT	pcBT	pcNK	pcT	pcBT	pcNK	
Patient 2	3,40	95,50	1,00	34,41	52,33	13,26	Entraînement
Patient 3	11,60	82,50	4,60	1,99	97,61	0,40	
Patient 4	21,90	75,40	2,20	3,53	94,70	1,77	
Patient 5	37,70	56,00	5,90	48,10	32,91	18,99	
Patient 7	6,70	92,60	0,20	7,41	88,89	3,70	
Patient 8	16,00	78,80	4,00	18,75	79,30	1,95	
Patient 10	12,00	82,00	5,80	0,00	99,10	0,90	
Patient 13	5,80	93,00	1,10	0,32	92,74	6,94	
Patient 14	16,80	78,90	3,70	13,82	84,00	2,18	
Patient 15	8,10	89,50	1,90	7,62	82,96	9,42	
Patient 18	0,00	100,00	0,00	20,66	78,51	0,83	
Patient 19	4,40	94,30	1,00	10,71	89,29	0,00	
Patient 20	5,20	93,30	1,20	17,73	75,91	6,36	
Patient 21	0,00	100,00	0,00	59,79	31,62	8,59	
Patient 22	4,20	94,80	0,90	2,26	97,74	0,00	
Patient 23	15,70	81,10	3,10	1,48	93,49	5,03	
Patient 24	6,50	91,60	1,80	0,00	100,00	0,00	
Patient 25	11,00	86,00	3,00	47,60	37,64	14,76	
Patient 26	12,00	86,10	1,60	9,09	90,91	0,00	
Patient 27	16,10	74,80	8,40	0,00	100,00	0,00	
Patient 29	3,40	94,10	1,90	5,47	90,63	3,91	
Patient 30	11,00	85,20	3,30	7,51	90,51	1,98	
Patient 31	6,80	91,10	1,90	3,11	96,50	0,39	

# **CONCLUSIONS ET PERSPECTIVES**

## 1. Conclusions

Au cours de cette thèse, différents outils d'analyse ont été développés et testés afin d'être implantés dans un système multimodal automatisé de seconde ligne pour l'aide au diagnostic de la LLC. Cet ouvrage s'est articulé autour de deux grands axes.

Le premier axe, concerne la construction et l'optimisation d'un classifieur Raman pour le diagnostic de la LLC, en évaluant dans un premier temps l'impact du prétraitement sur la transférabilité de données Raman bicentriques acquises dans différentes conditions expérimentales et instrumentales, puis dans un second temps, l'élaboration d'une stratégie efficace de classification supervisée au niveau du patient, basée sur une méthode de prédiction par vote majoritaire.

Le second axe consiste en une étude pilote, portant sur l'élaboration de modèles de prédiction plus précis permettant une différenciation des sous-types lymphocytaires et l'identification des cellules cancéreuses et ainsi être capable de prédire des patients à un stade précoce de la maladie. Pour entrainer ces classifieurs, des méthodes de labellisation efficaces et reproductibles ont été mises au point (Annexe 1), ainsi nous avons pu montrer l'applicabilité de la spectroscopie Raman au niveau cellulaire.

Le couplage des méthodes chimiométriques courantes, des méthodes d'optimisation telle que rdCV et des méthodes de classification supervisée ont montré leur efficacité en ce qui concerne la prédiction de données Raman issue de lymphocytes sains et de lymphocytes tumoraux au niveau patient et au niveau cellulaire. Dans ce travail, nous avons démontré que des différences de conditions expérimentales et instrumentales ont un impact direct sur les signaux Raman acquis sur les lymphocytes sains et LLC. Il est donc important de choisir correctement la méthode de prétraitement afin d'homogénéiser les ensembles de données et de construire des modèles transférables. Afin de pouvoir utiliser des données issues de conditions expérimentales et instrumentales différentes, deux stratégies de prétraitement ont été testées. Dans ce cadre, il a été démontré que la stratégie de prétraitement par étape n'est pas une méthode adéquate afin de corriger ce type de données. Quant à la stratégie parallèle et conjointe basée sur l'EMSC, elle permet d'homogénéiser de façon efficace les spectres Raman. Cependant, le prétraitement à lui seul n'est pas suffisant pour construire des modèles de haute performance, il est essentiel d'inclure un nombre élevé de patients et déterminer une stratégie optimale de classification.

Contrairement à la stratégie des modèles uniques, la méthode par vote consensuelle a démontré son efficacité pour maximiser la capacité de prédiction, en utilisant un ensemble de classificateurs individuels dont les prédictions ont été combinées par un vote majoritaire. Dans le cadre de cette thèse, 100 modèles ont été optimisés par la méthode de double validation croisée répétée (rdCV-10-Fold-out) et en appliquant des seuils de prédiction pour chaque modèle. Pour distinguer l'état pathologique des patients, les résultats de cette stratégie ont donné 85 % pour la sensibilité et 95 % pour la spécificité.

En ce qui concerne la distinction des sous-types lymphocytaires grâce à la labellisation par immunocoloration, deux stratégies ont été utilisées. La stratégie séquentielle a montré des performances beaucoup plus intéressantes que la stratégie simultanée. Malgré des résultats prometteurs, l'immunocoloration reste un gold-standard possédant des problèmes de fiabilité et de reproductibilité. De plus, pour des modèles entraînés sur des lymphocytes issus de tris magnétiques, il est important d'évaluer le degré de transférabilité sur des frottis sanguins.

De manière générale, ces travaux de thèse ont permis de mettre en avant le potentiel de la microspectroscopie Raman comme outil d'aide au diagnostic de la LLC, lorsqu'elle est couplée à des analyses numériques multivariées, que la décision soit prise à l'échelle du patient ou l'échelle cellulaire.

## **2. Perspectives**

### **2.1. Construction d'une base de données multimodale**

Afin de comparer les données spectrales avec d'autres données cliniques et biologiques, une banque de données a été construite au cours du projet M3S, dans laquelle chaque signature spectrale a été liée à des données de patients et à un pronostic spécifique. L'ensemble des données acquises comprenait pour chaque cellule analysée, les spectres Raman, des images en intensité et des images en Phase Quantitative (QPI). L'imagerie QPI est une technique basée sur la mesure du décalage de phase de la lumière. En effet lorsque la lumière se propage à travers un échantillon, il induit à celle-ci un retard. Ce retard (ou décalage de phase) est mesuré par la caméra QPI, nous obtenons alors une image pour laquelle chaque valeur de pixel est la mesure du déphasage local de la lumière. Plus précisément, la valeur du pixel est liée à l'épaisseur physique et à l'indice de réfraction local de l'échantillon biologique. Ainsi l'imagerie QPI permet d'avoir accès à différents paramètres (densité, masse sèche...) qui sont des indicateurs de nombreux mécanismes cellulaires. Dans le cadre de la LLC, cette technologie combinée à une analyse d'image pourrait permettre de distinguer les lymphocytes d'intérêt

parmi les cellules sanguines, d'avoir à la fois des informations morphologiques et optiques des lymphocytes cancéreux et sains ou de faire la distinction entre les sous-types lymphocytaires.

L'objectif long terme de cette base de données serait de pouvoir détecter les premiers stades de la LLC et de poser un pronostic personnalisé de l'évolution de celle-ci chez des patients, grâce au croisement des informations qu'elle contient.

## **2.2.Enrichissement de la base de données spectrale et utilisation d'autres stratégies de classification**

Dans la continuité de ces travaux, il serait intéressant d'étendre l'approche à des patients présentant d'autres hémopathologies qui peuvent être confondues avec la LLC lors du diagnostic : i) le lymphome de la zone marginale, dont la morphologie cellulaire est proche de celle de la LLC et dont le phénotype moléculaire peut varier fortement, ii) le lymphome du manteau, dont la morphologie cellulaire peut être confondue avec celle d'une LLC qui commence à vieillir. La distinction se fait actuellement par le marqueur CD23 présent uniquement sur la LLC, mais il peut arriver que ce marqueur cesse d'être exprimé par les lymphocytes B LLC, iii) le lymphome folliculaire qui migre rarement dans le sang, mais dont la morphologie cellulaire peut être proche de celle du lymphome du manteau. Ainsi l'élaboration d'un prédicteur Raman permettant de distinguer ces trois pathologies serait donc un service rendu aux les cliniciens.

Une autre approche de fort intérêt serait de réaliser une étude rétrospective afin d'avoir un modèle capable de prédire l'évolution de la maladie au cours du temps. De plus, il serait intéressant de réaliser une étude pharmacocinétique pour voir la réponse de la cellule tumorale à un médicament. La technologie employée dans cette thèse pourrait être étendu à d'autres pathologies cellulaires telles que par exemple le diagnostic de la malaria [65, 66].

Afin d'améliorer les performances de classification, d'autres stratégies de prédiction pourraient être envisagées. Le vote majoritaire couplé à la combinaison de différents algorithmes de classification linéaires, non-linéaires [164-167] permettrait d'explorer les données sous différents points de vue et de fusionner leurs résultats afin de prendre la meilleure décision. Des études récentes [168-171] ont montré le fort potentiel des perceptron « fully connected » et des réseaux de neurones profonds convolutifs (CNN) à réduire fortement le besoin de prétraitement et pour mettre en évidence des régions spectrales importantes, qui sont deux étapes cruciales dans l'analyse des données en spectroscopie vibrationnelle, mais

comme nous avons pu le voir dans cette thèse, ce sont des étapes critiques, qui peuvent poser des problèmes et faire diminuer les performances du modèle final. Ces algorithmes surpassent les algorithmes de classification standard utilisés en chimiométrie. En CNN, différentes techniques peuvent être utilisés afin d'analyser des données Raman tel que le 1D-CNN [168, 171] qui utilise des filtres convolutifs à une dimension ou le 2D-CNN [172] qui utilise des filtres de convolution en deux dimensions.

# **BIBLIOGRAPHIE**

1. Eklouh-Molinier, C., et al., *Investigating the relationship between changes in collagen fiber orientation during skin aging and collagen/water interactions by polarized-FTIR microimaging*. *Analyst*. **140**(18): p. 6260-6268.
2. Essendoubi, M., et al., *Human skin penetration of hyaluronic acid of different molecular weights as probed by Raman spectroscopy*. *Skin Research and Technology*. **22**(1): p. 55-62.
3. Caspers, P.J., G.W. Lucassen, and G.J. Puppels, *Combined in vivo confocal Raman spectroscopy and confocal microscopy of human skin*. *Biophys J*, 2003. **85**(1): p. 572-80.
4. Vuiblet, V., et al., *Renal Graft Fibrosis and Inflammation Quantification by an Automated Fourier Transform Infrared Imaging Technique*. *Journal of the American Society of Nephrology*: p. ASN. 2015050601.
5. Vuiblet, V., et al., *Raman-based detection of hydroxyethyl starch in kidney allograft biopsies as a potential marker of allograft quality in kidney transplant recipients*. *Scientific Reports*. **6**.
6. Lyng, F.M., et al., *Vibrational spectroscopy for cervical cancer pathology, from biochemical analysis to diagnostic tool*. *Exp Mol Pathol*, 2007. **82**(2): p. 121-9.
7. Duraipandian, S., et al., *Simultaneous fingerprint and high-wavenumber confocal Raman spectroscopy enhances early detection of cervical precancer in vivo*. *Anal Chem*. **84**(14): p. 5913-9.
8. Pappu, R., et al. *Raman spectroscopic characterization of urine of normal and cervical cancer subjects*. in *Proc. of SPIE Vol.*
9. Shaikh, R., et al., *A comparative evaluation of diffuse reflectance and Raman spectroscopy in the detection of cervical cancer*. *Journal of biophotonics*.
10. Lopes, R.M., et al., *Diagnostic model based on Raman spectra of normal, hyperplasia and prostate adenocarcinoma tissues in vitro*. *Journal of Spectroscopy*. **25**(2): p. 89-102.
11. McGregor, H.C., et al., *Real-time endoscopic Raman spectroscopy for in vivo early lung cancer detection*. *Journal of biophotonics*.
12. Sharma, N., N. Takeshita, and K.Y. Ho, *Raman Spectroscopy for the Endoscopic Diagnosis of Esophageal, Gastric, and Colonic Diseases*. *Clinical endoscopy*. **49**(5): p. 404.
13. Isabelle, M., et al., *Raman Spectroscopy-Based Cancer Diagnostic Platform For Pathology Classification In Barretts Oesophagus And Its Integration Into Clinic*. *Diagnostic Pathology*. **1**(8).
14. Shetty, G., et al., *Raman spectroscopy: elucidation of biochemical changes in carcinogenesis of oesophagus*. *Br J Cancer*, 2006. **94**(10): p. 1460-4.
15. Vaqas, B., et al., *OS3. 6 Optical Biopsies in Neurosurgery: Raman Spectroscopy for the Real-time Identification of Tumours during Surgery*. *Neuro-Oncology*. **18**(suppl 4): p. iv8-iv8.
16. Goodacre, R., et al., *Biofluids and other techniques: general discussion*. *Faraday discussions*. **187**: p. 575-601.
17. Ellis, D.I., et al., *Illuminating disease and enlightening biomedicine: Raman spectroscopy as a diagnostic tool*. *Analyst*. **138**(14): p. 3871-3884.
18. Diem, M., et al., *Molecular pathology via IR and Raman spectral imaging*. *J Biophotonics*. **6**(11-12): p. 855-86.
19. Liu, Y., Y. Zhu, and Z. Li, *Application of Raman spectroscopy in Andrology: non-invasive analysis of tissue and single cell*. *Translational andrology and urology*. **3**(1): p. 125.
20. Mallidis, C., et al., *In situ visualization of damaged DNA in human sperm by Raman microspectroscopy*. *Hum Reprod*. **26**(7): p. 1641-9.
21. Schulze, H.G., et al., *Assessing differentiation status of human embryonic stem cells noninvasively using Raman microspectroscopy*. *Analytical chemistry*. **82**(12): p. 5020-5027.
22. Chan, J.W., et al., *Label-free separation of human embryonic stem cells and their cardiac derivatives using Raman spectroscopy*. *Analytical chemistry*, 2009. **81**(4): p. 1324-1331.
23. Ichimura, T., et al., *Visualizing cell state transition using Raman spectroscopy*. *PLoS One*. **9**(1): p. e84478.
24. Ichimura, T., et al., *Visualizing the appearance and disappearance of the attractor of differentiation using Raman spectral imaging*. *Scientific Reports*. **5**: p. 11358.

25. Poplineau, M., et al., *Raman microspectroscopy detects epigenetic modifications in living Jurkat leukemic cells*. Epigenomics. **3**(6): p. 785-794.
26. Okada, M., et al., *Label-free Raman observation of cytochrome c dynamics during apoptosis*. Proc Natl Acad Sci U S A. **109**(1): p. 28-32.
27. Fliedner, T.M., et al., *Structure and function of bone marrow hemopoiesis: mechanisms of response to ionizing radiation exposure*. Cancer Biotherapy and Radiopharmaceuticals, 2002. **17**(4): p. 405-426.
28. Diebold, J., et al., *Hémopathies malignes: définition et différentes variétés selon la classification de l'OMS 2001*. Revue francophone des laboratoires, 2008. **2008**(398): p. 65-71.
29. Troussard, X. and E. Cornet. *Apport des nouvelles petites molécules dans le traitement de la leucémie lymphoïde chronique et place du biologiste dans la surveillance de ces traitements*. in *Annales de Biologie Clinique*. 2016.
30. Hughes, M.E. and J. Gill, *Primer on Chronic Lymphocytic Leukemia: Part II*. The Journal for Nurse Practitioners, 2018.
31. Mertelsmann, R., et al., *Précis d'hématologie et d'oncologie*. 2011: Springer Science & Business Media.
32. Strati, P. and T.D. Shanafelt, *Monoclonal B-cell lymphocytosis and early-stage chronic lymphocytic leukemia: diagnosis, natural history, and risk stratification*. Blood, 2015. **126**(4): p. 454-462.
33. Guillaume, N., et al. *Pertinence des analyses cytologiques et immunophénotypiques dans le diagnostic de la leucémie lymphoïde chronique B*. in *Annales de biologie clinique*. 2002.
34. Hallek, M., et al., *Guidelines for the diagnosis and treatment of chronic lymphocytic leukemia: a report from the International Workshop on Chronic Lymphocytic Leukemia (IWCLL) updating the National Cancer Institute-Working Group (NCI-WG) 1996 guidelines*. Blood, 2008.
35. Kaur, P., *Chronic Lymphocytic Leukemia/Small Lymphocytic Lymphoma Introduction-Definition, Diagnosis, Cell of Origin*, in *Chronic Lymphocytic Leukemia*. 2018, Springer. p. 1-33.
36. Binet, J.L., et al., *A new prognostic classification of chronic lymphocytic leukemia derived from a multivariate survival analysis*. Cancer, 1981. **48**(1): p. 198-206.
37. Aurrant, T., et al., *Recommandations 2012 de la SFH pour le diagnostic, le traitement et le suivi de la leucémie lymphoïde chronique*. Hématologie, 2013. **19**(1): p. 4-9.
38. Leblanc, R.-M., *Quelle place pour les marqueurs pronostiques de la leucémie lymphoïde chronique en 2010?* Option/Bio, 2010. **21**(432): p. 16-17.
39. Nguyen-Khac, F., et al. *Place de la cytogénétique dans la prise en charge de la leucémie lymphoïde chronique: actualisation du Groupe francophone de cytogénétique hématologique (GFCH)*. in *Annales de Biologie Clinique*. 2016.
40. Eclache, V. and F. Baran-Marszak, *Intérêt de la cytogénétique et de la biologie moléculaire pour le diagnostic des syndromes lymphoprolifératifs B*. Revue francophone des laboratoires, 2013. **2013**(452): p. 49-59.
41. Muller, C., et al., *Cellular pharmacokinetics of doxorubicin in patients with chronic lymphocytic leukemia: comparison of bolus administration and continuous infusion*. Cancer chemotherapy and pharmacology, 1993. **32**(5): p. 379-384.
42. Rozman, C. and E. Montserrat, *Chronic lymphocytic leukemia*. New England Journal of Medicine, 1995. **333**(16): p. 1052-1057.
43. Jain, P. and S. O'Brien, *Chronic lymphocytic leukemia*, in *Targeted therapy in translational cancer research*. 2015, John Wiley & Sons, Hoboken, NJ. p. 130-144.
44. Lepretre, S. and E. Van Den Neste, *Place de la chimiothérapie dans la leucémie lymphoïde chronique*. Hématologie, 2006. **12**(6): p. 28-34.
45. Robak, T., *Therapy of chronic lymphocytic leukemia with purine analogs and monoclonal antibodies*. Transfusion and apheresis science, 2005. **32**(1): p. 33-44.
46. Byrd, J.C., et al., *Targeting BTK with ibrutinib in relapsed chronic lymphocytic leukemia*. New England Journal of Medicine, 2013. **369**(1): p. 32-42.

47. Nair, K.S. and B. Cheson, *The role of idelalisib in the treatment of relapsed and refractory chronic lymphocytic leukemia*. Therapeutic advances in hematology, 2016. **7**(2): p. 69-84.
48. Roberts, A.W., et al., *Substantial susceptibility of chronic lymphocytic leukemia to BCL2 inhibition: results of a phase I study of navitoclax in patients with relapsed or refractory disease*. Journal of Clinical Oncology, 2011. **30**(5): p. 488-496.
49. Seymour, J.F., et al., *ABT-199 (GDC-0199) in relapsed/refractory (R/R) chronic lymphocytic leukemia (CLL) and small lymphocytic lymphoma (SLL): High complete-response rate and durable disease control*. 2014, American Society of Clinical Oncology.
50. Billard, C., *Apoptosis inducers in chronic lymphocytic leukemia*. Oncotarget, 2014. **5**(2): p. 309.
51. Keating, M.J., et al., *Therapeutic role of alemtuzumab (Campath-1H) in patients who have failed fludarabine: results of a large international study*. Blood, 2002. **99**(10): p. 3554-3561.
52. Hillmen, P., et al., *Alemtuzumab compared with chlorambucil as first-line therapy for chronic lymphocytic leukemia*. Journal of Clinical Oncology, 2007. **25**(35): p. 5616-5623.
53. McLaughlin, P., et al., *Rituximab chimeric anti-CD20 monoclonal antibody therapy for relapsed indolent lymphoma: half of patients respond to a four-dose treatment program*. Journal of Clinical Oncology, 1998. **16**(8): p. 2825-2833.
54. Hainsworth, J.D., et al., *Single-agent rituximab as first-line and maintenance treatment for patients with chronic lymphocytic leukemia or small lymphocytic lymphoma: a phase II trial of the Minnie Pearl Cancer Research Network*. Journal of Clinical Oncology, 2003. **21**(9): p. 1746-1751.
55. Dalle, S. and C. Dumontet, *Rituximab: mode d'action et de résistance*. Bulletin du cancer, 2007. **94**(2): p. 198-202.
56. Dreger, P., et al., *Indications for allogeneic stem cell transplantation in chronic lymphocytic leukemia: the EBMT transplant consensus*. Leukemia, 2007. **21**(1): p. 12.
57. Monnereau, A., et al., *Estimation nationale de l'incidence des cancers en France entre 1980 et 2012. Partie 2- Hémopathies malignes*. Saint-Maurice (Fra): Institut de veille sanitaire; 2013. 88 p, 2013.
58. Larkin, P., *Infrared and Raman spectroscopy: principles and spectral interpretation*. 2017: Elsevier.
59. Schrader, B., *Infrared and Raman spectroscopy: methods and applications*. 2008: John Wiley & Sons.
60. McCreery, R.L., *Raman spectroscopy for chemical analysis*. Vol. 225. 2005: John Wiley & Sons.
61. Hollas, J.M., *Modern spectroscopy*. 2004: John Wiley & Sons.
62. Barbillat, J., et al., *Spectrométrie raman*. Techniques de l'ingénieur. Analyse et caractérisation, 1999. **9**(P2865): p. 1-31.
63. Derruau, S., et al., *Shedding light on confounding factors likely to affect salivary infrared biosignatures*. Analytical and bioanalytical chemistry, 2019: p. 1-8.
64. Perez-Guaita, D., et al., *High resolution FTIR imaging provides automated discrimination and detection of single malaria parasite infected erythrocytes on glass*. Faraday discussions. **187**: p. 341-352.
65. Wood, B.R., et al., *Malaria Diagnosis Using ATR-FTIR Spectroscopy*. FTIR Spectroscopy in Microbiological and Medical Diagnostics, 2015.
66. Wood, B.R., et al., *Method and system for rapid malaria detection*. 2018, Google Patents.
67. Bilal, M., et al., *Optical diagnosis of malaria infection in human plasma using Raman spectroscopy*. Journal of biomedical optics, 2015. **20**(1): p. 017002.
68. Frame, L., et al., *Development of a label-free Raman imaging technique for differentiation of malaria parasite infected from non-infected tissue*. Analyst, 2015. **143**(1): p. 157-163.
69. Ichimura, T., et al., *Non-label immune cell state prediction using Raman spectroscopy*. Scientific reports. **6**.
70. Chen, M., et al., *The use of Wavelength Modulated Raman Spectroscopy in label-free identification of T lymphocyte subsets, Natural Killer cells and Dendritic cells*. PloS one. **10**(5): p. e0125158.

71. Weselucha-Birczynska, A., et al., *Raman micro-spectroscopy tracing human lymphocyte activation*. Analyst. **138**(23): p. 7157-63.
72. Hobro, A.J., et al., *Raman spectroscopy as a tool for label-free lymphocyte cell line discrimination*. Analyst. **141**(12): p. 3756-3764.
73. Chan, J.W., et al., *Micro-Raman spectroscopy detects individual neoplastic and normal hematopoietic cells*. Biophys J, 2006. **90**(2): p. 648-56.
74. Vanna, R., et al., *Label-free imaging and identification of typical cells of acute myeloid leukaemia and myelodysplastic syndrome by Raman microspectroscopy*. Analyst. **140**(4): p. 1054-1064.
75. Managá, S., et al., *A reliable Raman-spectroscopy-based approach for diagnosis, classification and follow-up of B-cell acute lymphoblastic leukemia*. Scientific reports. **6**.
76. Kerr, L.T. and B.M. Hennelly, *A multivariate statistical investigation of background subtraction algorithms for Raman spectra of cytology samples recorded on glass slides*. Chemometrics and Intelligent Laboratory Systems. **158**: p. 61-68.
77. Liland, K.H., A. Kohler, and N.K. Afseth, *Model based pre-processing in Raman spectroscopy of biological samples*. Journal of Raman Spectroscopy.
78. Kerr, L.T. and B.M. Hennelly, *A multivariate statistical investigation of background subtraction algorithms for Raman spectra of cytology samples recorded on glass slides*. Chemometrics and Intelligent Laboratory Systems, 2016. **158**: p. 61-68.
79. Fazio, E., et al., *A micro-Raman spectroscopic investigation of leukemic U-937 cells in aged cultures*. Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy, 2016. **159**: p. 21-29.
80. Managá, S., et al., *Identification and classification of acute lymphoblastic leukemia cells from peripheral blood by using Raman spectroscopy*.
81. Read, D.S. and A.S. Whiteley, *Chemical fixation methods for Raman spectroscopy-based analysis of bacteria*. Journal of microbiological methods, 2015. **109**: p. 79-83.
82. Mlynáriková, K., et al., *Influence of culture media on microbial fingerprints using Raman spectroscopy*. Sensors, 2015. **15**(11): p. 29635-29647.
83. Happillon, T., et al., *Diagnosis approach of chronic lymphocytic leukemia on unstained blood smears using Raman microspectroscopy and supervised classification*. Analyst, 2015. **140**(13): p. 4465-4472.
84. Bocklitz, T., et al., *How to pre-process Raman spectra for reliable and stable models?* Analytica chimica acta, 2011. **704**(1-2): p. 47-56.
85. Wehrens, R., *Chemometrics with R: multivariate data analysis in the natural sciences and life sciences*. 2011: Springer Science & Business Media.
86. Savitzky, A. and M.J.E. Golay, *Smoothing and differentiation of data by simplified least squares procedures*. Analytical chemistry, 1964. **36**(8): p. 1627-1639.
87. Shaver, J.M., *Chemometrics for Raman spectroscopy*. Practical Spectroscopy Series, 2001. **28**: p. 275-306.
88. McCreery, R.L., *Signal-to-noise in Raman spectroscopy*. Raman Spectroscopy for Chemical Analysis, 2005.
89. Lieber, C.A. and A. Mahadevan-Jansen, *Automated method for subtraction of fluorescence from biological Raman spectra*. Applied spectroscopy, 2003. **57**(11): p. 1363-1367.
90. Eilers, P.H.C., *Parametric time warping*. Analytical chemistry, 2004. **76**(2): p. 404-411.
91. Liland, K.H., T. Almås, and B.r.-H. Mevik, *Optimal choice of baseline correction for multivariate calibration of spectra*. Applied spectroscopy, 2010. **64**(9): p. 1007-1016.
92. Xi, Y., et al., *A Novel Pre-Processing Algorithm Based on the Wavelet Transform for Raman Spectrum*. Applied spectroscopy, 2018: p. 0003702818789695.
93. Martens, H. and E. Stark, *Extended multiplicative signal correction and spectral interference subtraction: new preprocessing methods for near infrared spectroscopy*. Journal of pharmaceutical and biomedical analysis, 1991. **9**(8): p. 625-635.

94. Afseth, N.K. and A. Kohler, *Extended multiplicative signal correction in vibrational spectroscopy, a tutorial*. Chemometrics and Intelligent Laboratory Systems, 2012. **117**: p. 92-99.
95. Ly, E., et al., *Combination of FTIR spectral imaging and chemometrics for tumour detection from paraffin-embedded biopsies*. Analyst, 2008. **133**(2): p. 197-205.
96. Sockalingum, G.D., et al. *Label free technologies 3: infrared imaging applied to paraffinized tissue microarrays for colon cancer diagnosis*. in *Diagnostic Pathology*. 2013. Springer.
97. de Lima, F.A., et al., *Digital de-waxing on FTIR images*. Analyst, 2017. **142**(8): p. 1358-1370.
98. Gaydou, V., et al., *Vibrational Analysis of Lung Tumor Cell Lines: Implementation of an Invasiveness Scale Based on the Cell Infrared Signatures*. Analytical chemistry, 2016. **88**(17): p. 8459-8467.
99. Kerr, L.T., et al., *Methodologies for bladder cancer detection with Raman based urine cytology*. Analytical Methods, 2016. **8**(25): p. 4991-5000.
100. Duby, C. and S.p. Robin, *Analyse en composantes principales*. Institut National Agronomique, Paris-Grignon, 2006. **80**.
101. Pearson, K., *LIII. On lines and planes of closest fit to systems of points in space*. The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science, 1901. **2**(11): p. 559-572.
102. Hyvärinen, A., J. Karhunen, and E. Oja, *Independent component analysis*. Vol. 46. 2004: John Wiley & Sons.
103. Lee, D.D. and H.S. Seung, *Learning the parts of objects by non-negative matrix factorization*. Nature, 1999. **401**(6755): p. 788.
104. Mallat, S.G., *A theory for multiresolution signal decomposition: the wavelet representation*. IEEE transactions on pattern analysis and machine intelligence, 1989. **11**(7): p. 674-693.
105. Zhang, C., et al., *Label-free Raman spectroscopy provides early determination and precise localization of breast cancer-colonized bone alterations*. Chemical Science, 2018.
106. Ramos, I.R., et al., *Raman spectroscopy for cytopathology of exfoliated cervical cells*. Faraday discussions, 2016. **187**: p. 187-198.
107. Vuiblet, V., et al., *Renal graft fibrosis and inflammation quantification by an automated fourier transform infrared imaging technique*. Journal of the American Society of Nephrology: p. ASN. 2015050601.
108. Meade, A.D., et al., *Vibrational spectroscopy in sensing radiobiological effects: analyses of targeted and non-targeted effects in human keratinocytes*. Faraday discussions, 2015. **187**: p. 213-234.
109. Wood, B.R., et al., *Diagnosing malaria infected cells at the single cell level using focal plane array Fourier transform infrared imaging spectroscopy*. Analyst, 2014. **139**(19): p. 4769-4774.
110. Bibby, J.M., J.T. Kent, and K.V. Mardia, *Multivariate analysis*. 1979, Academic Press, London.
111. Mastrangelo, C., *Statistical Monitoring of Complex Multivariate Processes with Applications in Industrial Process Control*. Journal of Quality Technology. **45**(1): p. 118-120.
112. Hollander, M., D.A. Wolfe, and E. Chicken, *Nonparametric statistical methods*. Vol. 751. 2013: John Wiley & Sons.
113. Gu, Q., Z. Li, and J. Han, *Generalized fisher score for feature selection*. arXiv preprint arXiv:1202.3725, 2012.
114. Liu, H. and H. Motoda, *Feature extraction, construction and selection: A data mining perspective*. Vol. 453. 1998: Springer Science & Business Media.
115. Leardi, R., M.B. Seasholtz, and R.J. Pell, *Variable selection for multivariate calibration using a genetic algorithm: prediction of additive concentrations in polymer films from Fourier transform-infrared spectral data*. Analytica chimica acta, 2002. **461**(2): p. 189-200.
116. Jarvis, R.M. and R. Goodacre, *Genetic algorithm optimization for pre-processing and variable selection of spectroscopic data*. Bioinformatics, 2004. **21**(7): p. 860-868.
117. Breiman, L., *Random forests*. Machine learning, 2001. **45**(1): p. 5-32.
118. Celebi, M.E. and K. Aydin, *Unsupervised Learning Algorithms*. 2016: Springer.

119. Ward Jr, J.H., *Hierarchical grouping to optimize an objective function*. Journal of the American statistical association, 1963. **58**(301): p. 236-244.
120. Roux, M., *Algorithmes de classification*. 1985: Masson.
121. MacQueen, J. *Some methods for classification and analysis of multivariate observations*. in *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*. 1967. Oakland, CA, USA.
122. Bezdek, J.C., *Objective Function Clustering*, in *Pattern recognition with fuzzy objective function algorithms*. 1981, Springer. p. 43-93.
123. Nguyen, T.N.Q., et al., *Development of a hierarchical double application of crisp cluster validity indices: a proof-of-concept study for automated FTIR spectral histology*. Analyst, 2015. **140**(7): p. 2439-2448.
124. Wang, X.Y., J. Garibaldi, and T. Ozen. *Application of the fuzzy C-means clustering method on the analysis of non pre-processed FTIR data for cancer diagnosis*. in *Internat. Conf. on Australian and New Zealand Intelligent Information Systems (ANZIIS)*. 2003.
125. Hughes, C., et al., *FTIR microspectroscopy of selected rare diverse sub-variants of carcinoma of the urinary bladder*. Journal of biophotonics, 2013. **6**(1): p. 73-87.
126. Barker, M. and W. Rayens, *Partial least squares for discrimination*. Journal of Chemometrics, 2003. **17**(3): p. 166-173.
127. Maguire, A., et al., *Competitive evaluation of data mining algorithms for use in classification of leukocyte subtypes with Raman microspectroscopy*. Analyst, 2015. **140**(7): p. 2473-2481.
128. Neugebauer, U., et al., *Towards detection and identification of circulating tumour cells using Raman spectroscopy*. Analyst. **135**(12): p. 3178-3182.
129. Cortes, C. and V. Vapnik, *Support-vector networks*. Machine learning, 1995. **20**(3): p. 273-297.
130. Lerner, J.M., *Imaging spectrometer fundamentals for researchers in the biosciences - a tutorial*. Cytometry Part A, 2006. **69**(8): p. 712-734.
131. Dietrich, B. and K.F. Dombrowski, *Experimental challenges of stress measurements with resonant micro-Raman spectroscopy*. Journal of Raman Spectroscopy, 1999. **30**(10): p. 893-897.
132. Jewett, J.W. and R.A. Serway, *Physics for scientists and engineers with modern physics, figure 38.13*. 2008: Cengage Learning EMEA.
133. Bocklitz, T., et al., *How to pre-process Raman spectra for reliable and stable models?* Analytica Chimica Acta, 2011. **704**(1): p. 47-56.
134. Afseth, N.K., V.H. Segtnan, and J.P. Wold, *Raman spectra of biological samples: A study of preprocessing methods*. Applied spectroscopy, 2006. **60**(12): p. 1358-1367.
135. Lasch, P., *Spectral pre-processing for biomedical vibrational spectroscopy and microspectroscopic imaging*. Chemometrics and Intelligent Laboratory Systems, 2012. **117**: p. 100-114.
136. Guo, S., et al., *Common mistakes in cross-validating classification models*. Anal. Methods, 2017.
137. Ming, L.C., et al., *Real time near-infrared Raman spectroscopy for the diagnosis of nasopharyngeal cancer*. Oncotarget, 2017. **8**(30): p. 49443.
138. Wold, S., H. Martens, and H. Wold, *The multivariate calibration problem in chemistry solved by the PLS method*, in *Matrix pencils*. 1983, Springer. p. 286-293.
139. Bohm, K., E. Smidt, and J. Tintner, *Application of multivariate data analyses in waste management*, in *Multivariate analysis in management, engineering and the sciences*. 2013, InTech.
140. Vapnik, V., *The nature of statistical learning theory*. 2013: Springer science & business media.
141. Cover, T.M., *Geometrical and statistical properties of systems of linear inequalities with applications in pattern recognition*. IEEE transactions on electronic computers, 1965(3): p. 326-334.
142. Breiman, L., *Classification and regression trees*. 1984: Routledge.
143. Breiman, L., *Bagging predictors*. Machine learning, 1996. **24**(2): p. 123-140.

144. Filzmoser, P., B. Liebmann, and K. Varmuza, *Repeated double cross validation*. Journal of Chemometrics: A Journal of the Chemometrics Society, 2009. **23**(4): p. 160-171.
145. Varmuza, K. and P. Filzmoser, *Repeated double cross validation (rdCV) - a strategy for optimizing empirical multivariate models, and for comparing their prediction performances*. Current Applications of Chemometrics. Ed. Khanmohammadi, M., Nova Science Publishers, Hauppauge, NY, USA, 2014: p. 15-32.
146. Filzmoser, P., B. Liebmann, and K. Varmuza, *Repeated double cross validation*. Journal of Chemometrics, 2009. **23**(4): p. 160-171.
147. Chen, M., et al., *The use of wavelength modulated Raman spectroscopy in label-free identification of T lymphocyte subsets, natural killer cells and dendritic cells*. PLoS One, 2015. **10**(5): p. e0125158.
148. McReynolds, N., et al., *Multimodal discrimination of immune cells using a combination of Raman spectroscopy and digital holographic microscopy*. Scientific Reports, 2017. **7**: p. 43631.
149. Hobro, A.J., et al., *Raman spectroscopy as a tool for label-free lymphocyte cell line discrimination*. Analyst, 2016. **141**(12): p. 3756-3764.
150. Pully, V.V., A.T.M. Lenferink, and C. Otto, *Time-lapse Raman imaging of single live lymphocytes*. Journal of Raman Spectroscopy, 2011. **42**(2): p. 167-173.
151. Manago, S., et al., *A reliable Raman-spectroscopy-based approach for diagnosis, classification and follow-up of B-cell acute lymphoblastic leukemia*. Scientific Reports, 2016. **6**: p. 24821.
152. Chan, J.W., et al., *Micro-Raman spectroscopy detects individual neoplastic and normal hematopoietic cells*. Biophysical journal, 2006. **90**(2): p. 648-656.
153. Brown, K.L., et al., *Raman spectroscopic differentiation of activated versus non-activated T lymphocytes: An in vitro study of an acute allograft rejection model*. Journal of immunological methods, 2009. **340**(1): p. 48-54.
154. Chan, J.W., et al., *Nondestructive identification of individual leukemia cells by laser trapping Raman spectroscopy*. Analytical chemistry, 2008. **80**(6): p. 2180-2187.
155. Happillon, T., et al., *Diagnosis approach of chronic lymphocytic leukemia on unstained blood smears using Raman microspectroscopy and supervised classification*. Analyst. **140**(13): p. 4465-72.
156. Ramoji, A., et al., *Toward a spectroscopic hemogram: Raman spectroscopic differentiation of the two most abundant leukocytes from peripheral blood*. Analytical chemistry, 2012. **84**(12): p. 5335-5342.
157. Bankapur, A., et al., *Raman tweezers spectroscopy of live, single red and white blood cells*. PLoS One, 2010. **5**(4): p. e10427.
158. Zoladek, A.B., et al., *Label-free molecular imaging of immunological synapses between dendritic and T cells by Raman micro-spectroscopy*. Analyst, 2010. **135**(12): p. 3205-3212.
159. Ashok, P.C., et al. *Label-free haemogram using wavelength modulated Raman spectroscopy for identifying immune-cell subset*. in *Biomedical Vibrational Spectroscopy Vi: Advances in Research and Industry*. 2014. International Society for Optics and Photonics.
160. Notingher, I., et al., *Spectroscopic study of human lung epithelial cells (A549) in culture: living cells versus dead cells*. Biopolymers, 2003. **72**(4): p. 230-240.
161. Okada, M., et al., *Label-free Raman observation of cytochrome c dynamics during apoptosis*. Proceedings of the National Academy of Sciences, 2012. **109**(1): p. 28-32.
162. Moscato, P. and C. Cotta, *Memetic algorithms*. Handbook of Applied Optimization, 2002. **157**: p. 168.
163. Farah, I., et al., *Development of a memetic clustering algorithm for optimal spectral histology: application to FTIR images of normal human colon*. Analyst, 2016. **141**(11): p. 3296-3304.
164. West, T., S. Prasad, and L.M. Bruce. *Multiclassifiers and decision fusion in the wavelet domain for exploitation of hyperspectral data*. in *Geoscience and Remote Sensing Symposium, 2007. IGARSS 2007. IEEE International*. 2007. IEEE.

165. Luan, L., et al., *Application of multiple classifier fusion in the discriminant analysis of near infrared spectroscopy for agricultural products*. *Journal of Near Infrared Spectroscopy*, 2016. **24**(4): p. 363-372.
166. Trevisan, J.I., et al., *Extracting biological information with computational analysis of Fourier-transform infrared (FTIR) biospectroscopy datasets: current practices to future perspectives*. *Analyst*, 2012. **137**(14): p. 3202-3215.
167. Gajjar, K., et al., *Fourier-transform infrared spectroscopy coupled with a classification machine for the analysis of blood plasma or serum: a novel diagnostic approach for ovarian cancer*. *Analyst*, 2013. **138**(14): p. 3917-3926.
168. Acquarelli, J., et al., *Convolutional neural networks for vibrational spectroscopic data analysis*. *Analytica chimica acta*, 2017. **954**: p. 22-31.
169. Bjerrum, E.J., M. Glahder, and T. Skov, *Data Augmentation of Spectral Data for Convolutional Neural Network (CNN) Based Deep Chemometrics*. arXiv preprint arXiv:1710.01927, 2017.
170. Liu, J., et al., *Deep convolutional neural networks for Raman spectrum recognition: a unified solution*. *Analyst*, 2017. **142**(21): p. 4067-4074.
171. Malek, S., F. Melgani, and Y. Bazi, *One-dimensional convolutional neural networks for spectroscopic signal regression*. *Journal of Chemometrics*, 2018. **32**(5): p. e2977.
172. Krauss, S.D., et al., *Hierarchical deep convolutional neural networks combine spectral and spatial information for highly accurate Raman microscopy based cytopathology*. *Journal of biophotonics*, 2018: p. e201800022.

# **ANNEXES**

# **ANNEXE 1**

Séchage des lames 30' du fixatif à l'air libre

## Procédure d'immunomarquage avec fixation 50/50 méthanol/acétone

### Réactifs :

- MACH 4 Universal AP-Polymer Kit (Biocare Medical – réf. M4U536H)
- Warp Red Chromogen Kit (Biocare Medical – réf. WR806H)

### Préparation des réactifs:

1. Solution de lavage : Dans du PBS (pH 7.2), diluer de la BSA afin d'obtenir une concentration protéique à g/L. Un volume total de 250mL de solution de lavage est suffisant.
2. Tampon de dilution des Anticorps : Mélanger 1mL de PBS à 1mL de SVF
3. Diluer chaque anticorps nécessaire avec ce tampon de dilution. La dilution est propre à l'anticorps utilisé.
  - a. Dilution 1/20
    - i. CD19 (Becton Dickinson Clone 4G7 – réf. 347540) *B méiçoses*
    - ii. CD3 (Dako Polyclonal – réf. A0452) *Lympho T*
    - iii. CD20 (Dako Polyclonal Clone L26 – réf. M0755) *Lympho B & matures*
    - iv. CD2 (Becton Dickinson Clone S5.2 – réf. 347590) – *Lympho T et NK.*
    - v. CD41a (Becton Dickinson Clone HIP8 – réf. 555465)
  - b. Dilution 1/50
    - i. CD45 (Becton Dickinson Clone 2D1 – réf. 347460)
4. Solution de Warp Red (ATTENTION : Attendre l'étape 7 du marquage avant la préparation de la solution), pour 8 à 10 lames :
  - a. Sortir le kit Warp Red du Frigo (Frigo HEFR06).
  - b. Dans un tube gradué, ajouter une goutte de Warp Red (Biocare Medical – Réf. WR806CHC) à 2,5mL de Warp Red Buffer (Biocare Medical – Réf. WR806BFL).
  - c. Conserver la solution à l'abri de la lumière à t° ambiante.
  - d. La solution doit reposer pendant 10 minutes et doit être utilisées dans les 30 minutes après sa préparation.

2 lames / 1 ml  
à  
préparer

### Fixation des lames :

1. Plonger les lames pendant 10 minutes dans un mélange extemporané d'acétone/méthanol (50-50) à température ambiante.
2. Après séchage, délimiter la zone de marquage au DakoPen (Dako – réf. S2002) comme ci-dessous :



### Marquage

- L'ensemble des incubations se réalise dans une chambre humide et à température ambiante.
  - Les lavages se font dans un coplin, rempli de solution de lavage. Après 5 secondes d'agitation douce, laisser reposer deux minutes. Eliminer la solution de lavage.
1. Sortir le kit MACH 4 AP Polymere Kit du Frigo (Frigo HEFR06).
  2. Déposer la solution d'anticorps dilués sur le spot délimité par le Dakopen
  3. Incuber 20 minutes puis laver une fois.
  4. Couvrir le spot avec le liquide jaune du Kit MACH4 (réf. UP536H – MACH4 Universal AP-Probe)
  5. Incuber 15 minutes puis laver une fois
  6. Couvrir le spot avec du liquide rouge du Kit MACH4 (réf. MRAP536H. – MACH4 MR AP-Polymer)
  7. Incuber 15 minutes et laver une fois

8. Couvrir le spot avec la solution de Warp Red fraîchement préparée.
9. Incuber 7 minutes précisément.
10. Après incubation, plonger immédiatement les lames dans l'eau distillée pendant 1 à 2 minutes.
11. Eliminer l'eau et couvrir avec de l'Hématoxyline Solution Gill 3 (*Sigma Aldrich – réf. GHS3-50ML*).
12. Incuber 1 minute et rincer à l'eau courante pendant 1 minute
13. Sécher à l'air libre ou à froid avec un sèche-cheveux et monter entre lame et lamelle à l'Aquatex

## **ANNEXE 2**

# Publications et communications

## Publications scientifiques

### Brevet:

V VUIBLET, P RIEU, O PIOT, M FERE, C GOBINET. **Automated quantification of fibrosis by infrared spectroscopy**. European patent N° 9S176BTFR6 /1457323. Applicant: Université de Reims Champagne Ardenne. 01/08/2014

### Publication des 4 dernières années :

- V. Vuiblet, M. Fere, C. Gobinet, P. Birembaut, O. Piot, P. Rieu, **Renal graft fibrosis and inflammation quantification by an automated fourier transform infrared imaging technique**, Journal of the American Society of Nephrology, (2015) ASN. 2015050601.
- V. Vuiblet, M. Fere, E. Bankole, A. Wynckel, C. Gobinet, P. Birembaut, O. Piot, P. Rieu, **Raman-based detection of hydroxyethyl starch in kidney allograft biopsies as a potential marker of allograft quality in kidney transplant recipients**, Scientific reports, 6 (2016) 33045.
- G.K. Chandra, C. Eklouh-Molinier, M. Fere, J.-F.o. Angiboust, C. Gobinet, L. Van-Gulick, P. Jeannesson, O. Piot, **Probing in vitro ribose induced DNA-glycation using Raman microspectroscopy**, Analytical chemistry, 87 (2015) 2655-2664.
- B. Dekky, A. Wahart, H. Sartelet, M. Féré, J.-F. Angiboust, S. Dedieu, O. Piot, J. Devy, H. Emonard, **Cellular cholesterol distribution influences proteolytic release of the LRP-1 ectodomain**, Frontiers in pharmacology, 7 (2016) 25.
- V. Vuiblet, T.T. Nguyen, A. Wynckel, M. Fere, L. Van-Gulick, V. Untereiner, P. Birembaut, P. Rieu, O. Piot, **Contribution of Raman spectroscopy in nephrology: a candidate technique to detect hydroxyethyl starch of third generation in osmotic renal lesions**, Analyst, 140 (2015) 7382-7390.

## **Publications en cours de finalisation et soumission**

- **M. Féré**, O. Piot, L.H Liu, A. Beljebbar, V. Untereiner, D. Gheldof, M. Chollat, J. Klossa, B. Chatelain, C. Gobinet, **Focus on pre-processing step to ensure the clinical transferability of Raman data acquired on lymphocytes in different experimental and instrumental conditions**
- **M. Féré**, C. Gobinet, L.H Liu, A. Beljebbar, D. Gheldof, M. Chollat, J. Klossa, B. Chatelain, O. Piot, **Implementation of a classification strategy of Raman data collected in different clinical conditions: application to the diagnosis of chronic lymphocytic leukemia**

## **Communications orales**

- **M. Féré**, L.H Liu, C. Gobinet, A. Beljebbar, V. Untereiner, M. Manfait, D. Gheldof, E. Cornet, X. Troussard, A. Civet, A. Templier, M.Chollat, J. Klossa, S. Roux, P. Rideau, B. Chatelain, O. Piot, **Diagnosis of the chronic lymphocytic leukemia (CLL) using a Raman-based scanner optimized for blood smear analysis (M3S Project)**, 13th European Congress on Digital Pathology, Berlin, 25-28 mai 2016.

## **Communications par poster**

- **M. Féré**, A. Beljebbar, C. Gobinet, M. Manfait, O. Piot, V. Untereiner, V. Béasse, N. Rau, M. Bowe, L. Payne, A. Cauchon, P. Rideau, B. Chatelain, E. Cornet, X. Troussard, A. Civet, A. Templier, J. Nolte, M. Chollat, J. Klossa, S. Roux, **Apport de la spectroscopie Raman en cytopathologie dans le cadre du projet européen M3S (Multi Modal Microscopy Scanner)**, 21èmes Journées du Groupe Français de Spectroscopie Vibrationnelle (GFSV 2015).
- **M. Féré**, V. Untereiner, L.H Liu, C. Gobinet, A. Beljebbar, M. Manfait, J. Angulo, P. Rideau, D. Gheldof, E. Cornet, X. Troussard, A. Civet, A. Templier, M.Chollat, J. Klossa, S. Roux, B. Chatelain, O. Piot, **Application de la microspectroscopie Raman pour le diagnostic automatique de la leucémie lymphoïde chronique (LLC)**, École thématique CNRS en spectroscopie vibrationnelle, Ile de Porquerolles (Var), France, 9-14 octobre 2016.
- O. Piot, **M. Féré**, L.H. Liu, C. Gobinet, A. Beljebbar, V. Untereiner, M. Manfait, J. Angulo, P. Rideau, D. Gheldof, E. Cornet, X. Troussard, A. Civet, A. Templier, M. Chollat, J. Klossa, S. Roux, B. Chatelain, **Application of Raman microscopy for the automatic diagnosis of the chronic lymphocytic leukemia (CLL) in routine clinics**, 4th Conference on Advanced Applied Raman Spectroscopy, Berlin, Allemagne, 19-20 mai 2016.

- **M. Féré**, L. H. Liu, C. Gobinet, A. Beljebbar, V. Untereiner, J.-F. Angiboust, M. Manfait, D. Gheldof, H. Jacquemin, S. Walbrecq, E. Cornet, X. Troussard, B. Chatelain, J. Angulo, M. Chollat, J. Klossa, O. Piot, **Application of Raman microscopy for the diagnosis of the chronic lymphocytic leukemia (CLL)**, 13th European Congress on Digital Pathology, Berlin, Allemagne, 25-28 mai 2016.
- **M. Féré**, L.H Liu, C. Gobinet, A. Beljebbar, V. Untereiner, M. Manfait, D. Gheldof, E. Cornet, X. Troussard, A. Civet, A. Templier, M.Chollat, J. Klossa, S. Roux, P. Rideau, B. Chatelain, O. Piot, **Application of Raman microscopy for the diagnosis of the chronic lymphocytic leukemia (CLL)**, Advanced Vibrational Spectroscopy for Biomedical Applications: Faraday Discussion, Cambridge, Angleterre, 21-23 mars 2016.
- L.H. Liu, **M. Féré**, V. Untereiner, A. Beljebbar, D. Gheldof , B. Chatelain, M. Chollat, J. Klossa, O. Piot, C. Gobinet, **Effect of negative cell sorting on Raman spectra of lymphocytes**, International Conference “Shedding New Light on Disease” (SPEC 2016).

---

## **M3S – Développement de la spectroscopie Raman en cytopathologie : Application au diagnostic de la leucémie lymphoïde chronique**

---

---

En cytopathologie, le diagnostic de routine repose sur l'utilisation de coloration ou d'immunomarquage. Le développement des techniques optiques vibrationnelles, telles que la diffusion Raman, offre l'opportunité d'accéder à des informations sur la composition moléculaire globale de l'échantillon, sans en altérer l'intégrité. L'objectif de cette thèse est d'évaluer les capacités diagnostiques de la spectroscopie Raman dans le contexte de la leucémie lymphoïde chronique. Au-delà de la preuve de concept, nous avons cherché à intégrer les caractéristiques d'une application clinique pour construire un classifieur prédictif robuste à partir de données Raman collectées sur frottis sanguins non colorés. Dans une première étape, nous avons montré qu'il était possible d'homogénéiser les données acquises dans différentes conditions expérimentales. Ensuite, nous avons établi une stratégie d'entraînement et de validation permettant de créer des modèles stables et insensibles au sur-apprentissage, et adaptés aux exigences cliniques.

---

spectroscopie Raman ; leucémie lymphoïde chronique ; analyse des données ; diagnostic de la leucémie lymphoïde chronique ; du laboratoire au chevet du patient

---

### **M3S - Development of Raman spectroscopy in cytopathology: Application to the diagnosis of chronic lymphocytic leukaemia**

---

In cytopathology, routine diagnosis is based on the use of staining or immunolabelling. The development of optical vibration techniques, such as Raman scattering, provides the opportunity to access information on the global molecular composition of the sample, without altering its integrity. The objective of this thesis is to evaluate the diagnostic capabilities of Raman spectroscopy in the context of chronic lymphocytic leukaemia. Beyond proof of concept, we sought to integrate the characteristics of a clinical application to build a robust predictive classifier from Raman data collected on uncolored blood smears. In a first step, we showed that it was possible to homogenize the data acquired under different experimental conditions. Secondly, we established a training and validation strategy to create stable models that are insensitive to overlearning and adapted to clinical requirements.

---

Raman spectroscopy; chronic lymphocytic leukaemia; data analysis; diagnosis of chronic lymphocytic leukaemia; from bench to bedside

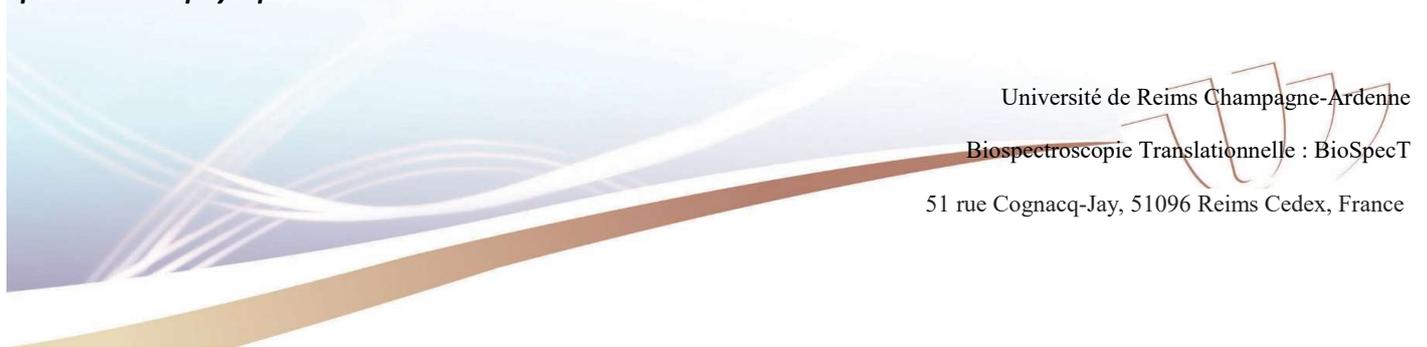
---

**Discipline : SCIENCES DE LA VIE ET DE LA SANTE**

---

**Spécialité : Biophysique**

---



Université de Reims Champagne-Ardenne  
Biospectroscopie Translationnelle : BioSpecT  
51 rue Cognacq-Jay, 51096 Reims Cedex, France