# Une approche pour regrouper des sujets atteints de cancers, sur la base des similarités des répartitions cellulaires au sein de leurs biopsies

Yassine El Ouahidi, Matis Feller, Matthieu Talagas, Bastien Pasdeloup

▶ **To cite this version:**

Yassine El Ouahidi, Matis Feller, Matthieu Talagas, Bastien Pasdeloup. Une approche pour regrouper des sujets atteints de cancers, sur la base des similarités des répartitions cellulaires au sein de leurs biopsies. EGC 2021 : 21ème conférence Extraction et Gestion des Connaissances, Jan 2021, Montpellier, France. pp.333-340. hal-03131445

**HAL Id: hal-03131445**
**https://imt-atlantique.hal.science/hal-03131445**

Submitted on 4 Feb 2021

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

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

# Une approche pour regrouper des sujets atteints de cancers, sur la base des similarités des répartitions cellulaires au sein de leurs biopsies

Yassine El Ouahidi*, Matis Feller*
Matthieu Talagas**,***, Bastien Pasdeloup*,****

*IMT Atlantique, Technopole Brest-Iroise, France
prénom.nom@imt-atlantique.fr
**Univ Brest, LIEN, F-29200 Brest, France
matthieu.talagas@chu-brest.fr
***Département de Dermatologie, CHU de Brest, Brest, France
****Lab-STICC, UMR CNRS 6285, Brest F-29238, France

**Résumé.** Nous présentons une méthodologie interprétable pour regrouper des sujets souffrant de cancer, basée sur des attributs extraits d'images numériques de leurs biopsies. Nous proposons d'analyser des répartitions cellulaires à l'aide d'histogrammes, et de comparer les sujets via ceux-ci. Nous décrivons ici la méthodologie pour définir ces histogrammes et les exploiter à des fins de clustering. Nous illustrons notre approche sur une base de données de tissus colorés à l'hématoxyline et à l'éosine de sujets atteints d'un adénocarcinome pulmonaire de stade I, où nos résultats coincident avec les connaissances existantes en matière d'estimation du pronostic de survie, avec une confiance statistique élevée.

## 1 Introduction

La médecine personnalisée en oncologie vise à améliorer le pronostic de survie d'un sujet, en prenant en compte des attributs spécifiques à son métabolisme. La plupart des travaux du domaine caractérisent les sujets sur la base d'attributs quantitatifs, tels que le nombre de lymphocytes infiltrants (TILs) (Mlecnik et al., 2011), leur densité et leur surface (Reichling et al., 2020), ou l'immunoscore (Galon et al., 2012). Cependant, il a été récemment montré que des organisations plus globales de cellules peuvent avoir un impact sur le pronostic vital (Fridman et al., 2012; Yener, 2016; Saltz et al., 2018).

Les biopsies sont numérisées via des scans à haute résolution de tissus cellulaires, préalablement colorés pour révéler les noyaux cellulaires. La coloration par hématoxyline et éosine (H&E) est la plus couramment utilisée pour des raisons financières et de simplicité d'acquisition. Des techniques plus complexes telles que l'immunohistochimie multiplex (mIHC) permettent de capturer des informations plus riches telles que la distinction des lymphocytes CD8/CD4/CD3, mais à un coût plus élevé. Une fois qu'une image haute résolution a été obtenue, une pratique courante consiste à effectuer sa segmentation et son phénotypage, afin de localiser et caractériser les cellules. Pour les images H&E, des outils utilisant des modèles

d'apprentissage profond (Wang et al., 2019) ont montré des performances impressionnantes pour localiser des cellules et leur associer un phénotype (cancer/stroma/lymphocyte). Des travaux existent pour les données mIHC, impliquant aussi des modèles d'apprentissage profond (Mercadier et al., 2019). Ces outils produisent des informations similaires. Pour chaque cellule, ils fournissent les attributs suivants : coordonnées `x` et `y`, et `phenotype`. Les phénotypes dépendent de la méthode de coloration et des cellules en présence. Des attributs additionnels (forme, niveau de kératine, etc.) peuvent être générés selon les outils considérés.

Dans cet article, nous introduisons une méthodologie pour capturer des répartitions de cellules dans les biopsies. Notre approche peut être utilisée à différentes échelles (cellulaire, groupes de cellules, etc.), fournissant des informations nombreuses et complémentaires sur le tissu. Contrairement aux approches existantes (e.g., Saltz et al. (2018)), nous ne regroupons pas les sujets dans des classes prédéfinies. Au lieu de cela, nous travaillons de manière non supervisée, et de regroupons les sujets en fonction de leurs similarités entre des attributs. Cela permet 1) l'exploration de nouveaux attributs (ou *bio-marqueurs computationnels*) qui peuvent être corrélés avec la survie ; et 2) la recherche de sujets similaires. La Section 2 détaille la méthodologie proposée ; la Section 3 illustre notre approche sur une base de données de biopsies issues de sujets atteints d'adénocarcinome pulmonaire de stade I, et montre que nos résultats coïncident avec les connaissances existantes dans le domaine. La Section 4 conclut.

# 2 Méthodologie proposée

Notre méthode prend en entrée un ensemble $\mathcal{S} = \{s^{(i)}\}_{i\in\{1,\dots,|\mathcal{S}|\}}$ de sujets. Chaque sujet $s \in \mathcal{S}$, dispose d'un ensemble $\mathcal{C}_s = \{c^{(i)}\}_{i\in\{1,\dots,|\mathcal{C}_s|\}}$ de cellules. Chaque cellule $c \in \mathcal{C}_s$ a des attributs, que nous notons $c[a]$, avec $a \in \{\texttt{x}, \texttt{y}, \texttt{phenotype}, \dots\}$. Nous procédons ainsi :

1. On définit un ensemble $\mathcal{F} = \{f^{(i)}\}_{i\in\{1,\dots,|\mathcal{F}|\}}$ de fonctions associant à un sujet $s \in \mathcal{S}$ un histogramme $H_s^{(i)}$ représentant un attribut complexe. Ces fonctions sont évaluées sur les biopsies des sujets afin d'obtenir les histogrammes $\{H_s^{(i)}\}_{s\in\mathcal{S},i\in\{1,\dots,|\mathcal{F}|\}}$ ;
2. Pour chaque attribut $f^{(i)} \in \mathcal{F}$, et pour toutes paires de sujets distincts $s^{(j)}, s^{(k)} \in \mathcal{S}$, on calcule une matrice de distances $\mathbf{D}^{(i)}[j,k] = d_H\left(H^{(i)}_{s^{(j)}}, H^{(i)}_{s^{(k)}}\right)$ associée à $f^{(i)}$ ;
3. Pour chaque matrice $\mathbf{D}^{(i)}$, on applique un algorithme de clustering sur $\mathcal{S}$ afin de regrouper les sujets en fonction de leur similarité vis à vis de $f^{(i)}$ ;
4. Enfin, on effectue une analyse de survie des sous-populations obtenues afin d'évaluer l'importance de $f^{(i)}$ sur le pronostic de survie.

## 2.1 Définition des attributs complexes

Nous présentons trois attributs modélisant des répartitions cellulaires [1]. Un attribut est une fonction $f^{(i)}$ qui, pour un sujet $s \in \mathcal{S}$ avec ses cellules $\mathcal{C}_s$, produit un histogramme $H_s^{(i)}$.

Introduisons dans un premier temps la notion de filtre. Soit $\mathcal{C}_s^- \subseteq \mathcal{C}_s$ un sous-ensemble des cellules du sujet $s \in \mathcal{S}$. On note $\mathcal{C}_s^-[P] \subseteq \mathcal{C}_s^-$ le sous-ensemble de cellules dans $\mathcal{C}_s^-$ pour lequel

1. D'autres attributs, ainsi que leur description sous forme d'algorithme, sont présentés dans la version longue de cet article, accessible à cette adresse : `https://arxiv.org/abs/2007.00135`.

une proposition $P : \mathcal{C}_s^- \to \mathbb{B}$ est vraie. Rappelons que des attributs sont associés aux cellules, notamment leurs phénotypes. Supposons que l'attribut `phenotype` prend des valeurs dans {`"cancer"`, `"stroma"`, `"lymphocyte"`}. Nous définissons le filtre suivant :

$$\begin{array}{rccl} lymphocyte: & \mathcal{C}_s^- & \to & \mathbb{B} \\ & c & \mapsto & c[\texttt{phenotype}] = \texttt{"lymphocyte"} \ . \end{array} \tag{1}$$

La notation $\mathcal{C}_s[lymphocyte]$ nous permet d'obtenir les lymphocytes de $\mathcal{C}_s$. De même, nous pouvons définir les filtres $stroma$ et $cancer$ qui renvoient les cellules correspondantes.

### 2.1.1 $f^{(1)}$ : distances lymphocytes – cellules cancéreuses

Soit un sujet $s \in \mathcal{S}$ aux cellules $\mathcal{C}_s$. Nous retranscrivons ici la proximité entre les lymphocytes et les cellules cancéreuses. Introduisons le filtre qui, pour une cellule donnée $c \in \mathcal{C}_s$, renvoie la cellule dans un ensemble $\mathcal{C}_s^- \subseteq \mathcal{C}_s$ minimisant la distance Euclidienne $d_{\text{xy}}$ à $c$ :

$$\begin{array}{rccl} closest(c): & \mathcal{C}_s^- & \to & \mathbb{B} \\ & c' & \mapsto & c' = \arg\min\limits_{c'' \in \mathcal{C}_s^-} d_{\text{xy}}(c'', c) \ , \end{array} \tag{2}$$

Ainsi, l'histogramme représentant ce premier attribut contient les valeurs :

$$\forall c \in \mathcal{C}_s[lymphocyte] : d_{\text{xy}}\left(c, \mathcal{C}_s[cancer][closest(c)]\right) \in H_s^{(1)} \ . \tag{3}$$

### 2.1.2 $f^{(2)}$ : distances lymphocytes – interface cancer/stroma

Nous retranscrivons ici la proximité entre les lymphocytes et l'interface (ou frontière) entre cellules cancéreuses et stromales. Soient $\mathcal{C}_s^- \subseteq \mathcal{C}_s$ et $\mathcal{C}_s^{-'} \subseteq \mathcal{C}_s$ deux sous-ensembles disjoints des cellules $\mathcal{C}_s$ d'un sujet $s \in \mathcal{S}$. De plus, soit $\mathcal{T}_s$ l'ensemble des triangles obtenus par triangulation de Delaunay des cellules dans $\mathcal{C}_s$, construite à partir des attributs `x` et `y` des cellules. Chaque triangle $t \in \mathcal{T}_s$ est un ensemble de trois cellules. Nous pouvons définir le filtre :

$$\begin{array}{rccl} interface(\mathcal{C}_s^{-'}): & \mathcal{C}_s^- & \to & \mathbb{B} \\ & c & \mapsto & \exists t \in \mathcal{T}_s : c \in t \text{ and } \exists c' \in t : c' \in \mathcal{C}_s^{-'} \ , \end{array} \tag{4}$$

qui renvoie le sous-ensemble des cellules de $\mathcal{C}_s^-$ adjacentes à des cellules de $\mathcal{C}_s^{-'}$, en vérifiant l'existence d'un triangle défini par des cellules des deux sous-ensembles. On note donc les cellules stromales à l'interface cancer/stroma $\mathcal{C}_s[stroma][interface(\mathcal{C}_s[cancer])]$, et les cellules cancéreuses à cette même interface $\mathcal{C}_s[cancer][interface(\mathcal{C}_s[stroma])]$.

Ainsi, l'histogramme représentant ce second attribut contient les valeurs :

$$\forall c \in \mathcal{C}_s[lymphocyte] : \left\{ \begin{array}{ll} d_c & \text{if } d_c \geq d_s \\ -d_c & \text{otherwise} \end{array} \right. \in H_s^{(2)} \ , \tag{5}$$

$$\text{où : } d_c = d_{\text{xy}}(c, \mathcal{C}_s[cancer][interface(\mathcal{C}_s[stroma])][closest(c)]) \tag{6}$$

$$\text{et : } d_s = d_{\text{xy}}(c, \mathcal{C}_s[stroma][interface(\mathcal{C}_s[cancer])][closest(c)]). \tag{7}$$

Nous distinguons ici les lymphocytes au sein de la tumeur ($d_c < d_s$) et ceux en dehors de celle-ci ($d_c \geq d_s$) en associant aux lymphocytes intra-tumoraux une distance négative.

### 2.1.3 $f^{(3)}$ : distances entre aggrégats de lymphocytes

Nous retranscrivons ici la proximité entre aggrégats de lymphocytes, définis comme suit : utilisons à nouveau la triangulation de Delaunay des cellules $\mathcal{C}_s$ d'un sujet $s \in \mathcal{S}$. Soit $\mathcal{E}_s$ l'ensemble des arêtes de la triangulation. On retire de $\mathcal{E}_s$ les arêtes reliant deux cellules dont au moins une n'est pas un lymphocyte. Le graphe composé des sommets $\mathcal{C}_s[lymphocyte]$ et des arêtes $\mathcal{E}_s$ est constitué de composantes connexes $\mathcal{K} = \{k^{(i)}\}_i$. Ainsi, l'histogramme représentant ce troisième attribut contient les valeurs :

$$\forall k, k' \in \mathcal{K}\ ; k \neq k'\ ; \forall c' \in k' : \min_{c \in k} d_{\mathrm{xy}}\left(c, c'\right) \in H_s^{(3)}\ . \tag{8}$$

## 2.2 Fonction de distance entre histogrammes

Nous voulons regrouper les sujets en fonction de la similarité de leurs histogrammes. Cela nécessite une fonction de distance $d_H$ qui soit conforme avec l'information capturée, *i.e.*, pour deux histogrammes $H_s^{(i)}$ et $H_{s'}^{(i)}$ – obtenus en calculant $f^{(i)}$ pour deux sujets $s, s' \in \mathcal{S}$ – nous voulons une distance augmentant à mesure que les formes de $H_s^{(i)}$ et $H_{s'}^{(i)}$ diffèrent. La distance de Wasserstein (Peyré et al., 2019) mesure la quantité de travail nécessaire pour transformer une distribution en une autre. Etant définie entre distributions, nous normalisons nos histogrammes pour que la somme des valeurs soit égale à 1. Une conséquence est que nous ne pouvons plus distinguer certains histogrammes. Cela peut sembler problématique, car des attributs tel que $f^{(1)}$ imposent une relation entre le nombre de lymphocytes dans le tissu et le nombre de bandes dans l'histogramme. Cependant, ces aspects quantitatifs peuvent être directement inclus comme caractéristiques numériques lors de la classification. De plus, cette normalisation ne change pas la forme de l'histogramme, ce qui nous permet de comparer – en reprenant l'exemple de $f^{(1)}$ – les repartitions de lymphocytes autour des cellules cancéreuses. Nous utiliserons donc la distance de Wasserstein $d_{wass}$ entre histogrammes normalisés, *i.e.*,

$$d_H\left(H_s^{(i)}, H_{s'}^{(i)}\right) = d_{wass}\left(\frac{H_s^{(i)}}{\left|H_s^{(i)}\right|_1}, \frac{H_{s'}^{(i)}}{\left|H_{s'}^{(i)}\right|_1}\right)\ , \tag{9}$$

où $|\cdot|_1$ est la norme $\ell_1$ de l'ensemble des valeurs d'un histogramme. La matrice de coûts dans la formulation de $d_{wass}$ (Peyré et al., 2019) est choisie linéaire.

## 2.3 Clustering des sujets

Nous regroupons à présent tous les sujets de $\mathcal{S}$ – de manière non-supervisée – de sortes à créer des ensembles de sujets aux repartitions de cellules similaires. Pour chaque attribut $f^{(i)} \in \mathcal{F}$, et pour chaque paire de sujets $s^{(j)}, s^{(k)} \in \mathcal{S}$, calculons la matrice $|\mathcal{S}| \times |\mathcal{S}|$ suivante :

$$\mathbf{D}^{(i)}[j,k] = d_H\left(H_{s^{(j)}}^{(i)}, H_{s^{(k)}}^{(i)}\right)\ . \tag{10}$$

Cette matrice est ensuite donnée en entrée à un algorithme de clustering. De par sa forte interprétabilité, nous avons choisi un algorithme de clustering hiérarchique (AHC, single-linkage) (Szekely et Rizzo, 2005). Cet algorithme produit un dendrogramme, dans lequel les

feuilles sont les sujets de $\mathcal{S}$, et les nœuds sont les aggrégations des sous-arbres. Ainsi, plus on descend profond dans le dendrogramme, plus les sous-arbres y sont homogènes à l'attribut utilisé pour sa construction. Nous utilisons donc AHC avec $\mathbf{D}^{(i)}$ pour séparer les sujets de $\mathcal{S}$ en deux sous-populations disjointes $\mathcal{S}_1^{(i)}$ et $\mathcal{S}_2^{(i)}$, comme suit :

— $\mathcal{S}_1^{(i)} \subseteq \mathcal{S}$ est obtenue en coupant le dendrogramme à profondeur $\tau_{\text{AHC}}$, et en conservant les sujets dans le plus profond sous-arbre obtenu. Ce cluster est homogène à $f^{(i)}$ ;

— $\mathcal{S}_2^{(i)} \subseteq \mathcal{S}$ est obtenue en regroupant l'ensemble des sujets n'appartenant pas à $\mathcal{S}_1^{(i)}$. Ainsi, ce cluster contient des sujets hétérogènes à $f^{(i)}$.

L'objectif de cette décomposition est de vérifier si des sujets ayant une répartition cellulaire analogue ont un pronostic de survie similaire, contrairement à un autre groupe pouvant exprimer des profils très personnels d'un même attribut. $\tau_{\text{AHC}}$ est obtenu par recherche exhaustive : pour $\tau_{\text{AHC}}$ fixé, nous calculons les courbes de survie de Kaplan-Meier (Kaplan et Meier, 1958) des sous-populations obtenues, et effectuons un log-rank test (Mantel, 1966) entre ces courbes pour quantifier leur séparabilité. La valeur retenue pour $\tau_{\text{AHC}}$ est celle qui minimise la $p$-value de ce test. Ce choix a pour but de maximiser l'homogénéité de $\mathcal{S}_1^{(i)}$ sans a priori sur sa taille.

### 2.4 Combinaison des différents attributs

Nous aimerions combiner plusieurs attributs pour le clustering, afin de regrouper des sujets montrant des similarités selon plusieurs critères. Pour cela, nous construisons une matrice $\mathbf{D}$ dénotant l'appartenance de deux sujets à des clusters différents, pour chaque attribut $f^{(i)} \in \mathcal{F}$. Cette agrégation est préférée à une combinaison linéaire car certaines matrices $\mathbf{D}^{(i)}$ corrèlent positivement ou négativement avec le pronostic, selon l'attribut. Soient $s^{(j)}, s^{(k)} \in \mathcal{S}$ :

$$\mathbf{D}[j,k] = \sum_{i=1}^{|\mathcal{F}|} w^{(i)} \cdot diff\left(s^{(j)}, s^{(k)}, \mathcal{S}_1^{(i)}\right) \text{ , où} \tag{11}$$

$$diff\left(s^{(j)}, s^{(k)}, \mathcal{S}_1^{(i)}\right) = \begin{cases} 0 & \text{si } \left(s^{(j)} \in \mathcal{S}_1^{(i)} \text{ et } s^{(k)} \in \mathcal{S}_1^{(i)}\right) \text{ ou } \left(s^{(j)} \notin \mathcal{S}_1^{(i)} \text{ et } s^{(k)} \notin \mathcal{S}_1^{(i)}\right) \\ 1 & \text{sinon ,} \end{cases} \tag{12}$$

et où $\{w^{(i)}\}_{i\in\{1,\ldots,|\mathcal{F}|\}}$ sont des valeurs réelles pondérant l'importance des attributs, données par un expert, ou choisies de manière plus automatique, comme proposé en Section 3.3.

## 3 Expériences

### 3.1 Base de données considérée

Nous considérons un ensemble de 140 sujets atteints d'adénocarcinome pulmonaire de stade I, issus de TCGA [2]. Pour que sujet, nous disposons de scans de biopsies colorées en H&E, et d'informations cliniques, dont le *time to last follow-up* (TTLFU) et l'état vital à cette date, permettant le calcul des courbes de survie. Nous avons retiré les sujets ayant un TTLFU inférieur à $\tau_{\text{DB}} = 366$ jours (pas assez de recul), ainsi que ceux décédés au delà de $\tau_{\text{DB}}$ jours

2. `http://cancergenome.nih.gov/`

| Attribut | $\lvert\mathcal{S}_1^{(i)}\rvert/\lvert\mathcal{S}\rvert$ | $\lvert\mathcal{S}_2^{(i)}\rvert/\lvert\mathcal{S}\rvert$ | $p$-value |
|---|---|---|---|
| $f^{(1)}$ | 0.55 | 0.45 | 0.0574 |
| $f^{(2)}$ | 0.31 | 0.69 | 0.0002 |
| $f^{(3)}$ | 0.52 | 0.48 | 0.0124 |

TAB. 1: Analyse de survie de la population en fonction de l'attribut utilisé pour le clustering.

(évolution possible de la tumeur). La valeur de $\tau_{\text{DB}}$ a été choisie pour maximiser la taille de $\mathcal{S}$ en encourageant les propriétés décrites. Après ce filtrage, la base de données contient 97 sujets. Pour chacun, nous avons considéré une biopsie, et avons choisi manuellement des régions d'intérêt (ROIs) afin d'inclure à la fois des cellules cancéreuses et stromales. Pour chaque ROI, nous avons segmenté et phénotypé les cellules via l'outil ConvPath (Wang et al., 2019). Enfin, nous avons retiré les ROI ne contenant pas un ratio cancer/stroma dans l'intervalle $[0.3, 0.7]$. Nous avons au final $94$ sujets dont $80$ vivants. Chacun dispose de $1$ à $9$ ROIs, (moyenne $3.86$). Ci-après, les histogrammes sont calculés par ROI, et agrégés avant calcul de distance.

## 3.2 Résultats par attribut (en isolation des autres)

Pour vérifier si les attributs introduits en Section 2.1 corrèlent significativement avec le pronostic vital, nous appliquons la méthode proposée en Section 2 pour chaque attribut individuellement. La Table 1 présente les résultats obtenus. Pour chaque attribut $f^{(i)} \in \mathcal{F}$, nous reportons les tailles des sous-populations obtenues, ainsi que la $p$-value du log-rank test entre leurs courbes de survie. Les résultats montrent que les attributs $f^{(2)}$ et $f^{(3)}$ séparent les courbes de survie significativement ($p$-value $< 0.05$). En particulier, $f^{(2)}$ est très significatif ($p$-value $\ll 0.05$). Toutefois, $f^{(1)}$ ne corrèle pas significativement à la survie.

Une manière d'analyser visuellement les clusters obtenus est d'observer un histogramme *central* pour chacun d'eux [3] : pour un attribut $f^{(i)} \in \mathcal{F}$ et un cluster $\mathcal{S}_j^{(i)}$ ($j \in \{1, 2\}$) donnés, l'histogramme $H_{s_j^*}^{(i)}$ le plus central du cluster $\mathcal{S}_j^{(i)}$, est celui qui minimise :

$$H_{s_j^*}^{(i)} = \arg\min_{s \in \mathcal{S}_j^{(i)}} \sum_{s' \in \mathcal{S}_j^{(i)}} d_H\left(H_s^{(i)}, H_{s'}^{(i)}\right) . \tag{13}$$

## 3.3 Combinaison d'attributs

Considérons les attributs ayant une significativité suffisante pour regrouper les sujets. Les poids $\{w^{(i)}\}_i$ dans l'Équation 11 sont choisis pour donner plus d'importance aux attributs plus significatifs. Soit $p^{(i)}$ la $p$-value du log-rank test pour l'attribut $f^{(i)} \in \mathcal{F}$. On choisit $w^{(i)} = \log\left(\frac{1}{p^{(i)}}\right)$ si $p^{(i)} > 0.05$, ou 0 sinon. En utilisant la matrice $\mathbf{D}$ en Équation 11 avec ces poids dans AHC, nous obtenons deux sous-populations $\mathcal{S}_1$ et $\mathcal{S}_2$. Le log-rank test entre les courbes de survie associées a une $p$-value de $7.36 \cdot 10^{-8}$, montrant une forte corrélation des clusters avec le pronostic vital [4].

3. Les figures présentant les courbes de survie obtenues, ainsi que les histogrammes centraux pour chaque cluster et attribut sont disponibles dans la version longue de cet article.

4. Ces résultats prennent en compte les attributs supplémentaires présentés dans la version longue de cet article.

### 3.4 Discussion

Notons que $f^{(1)}$ n'est pas assez significatif ($p$-value $> 0.05$). Une explication est qu'il ne distingue pas les lymphocytes profondément infiltrés dans la tumeur, de ceux dans l'environnement stromal, à la bordure de la tumeur, sans infiltration.

Cela est pris en compte dans $f^{(2)}$, qui corrèle fortement avec la survie. Les histogrammes centraux des clusters obtenus par $f^{(2)}$ révèlent que les deux sujets possèdent à la fois des lymphocytes infiltrés et non infiltrés dans la tumeur. Une différence notable entre est que le représentant de la sous-population homogène $\mathcal{S}_1^{(2)}$ a une infiltration moins importante (environ $-40\mu m$) que le représentant de la sous-population hétérogène $\mathcal{S}_2^{(2)}$ (environ $-80\mu m$). Afin de vérifier si cette observation se généralise à tous les sujets, nous avons listé la valeur minimale par histogramme, et avons analysé les distributions de ces valeurs par cluster. Ainsi, $\mathcal{S}_1^{(2)}$ a une moyenne de $-43.29\mu m$, (écart type de $20.56\mu m$); et $\mathcal{S}_2^{(2)}$ a une moyenne de $-112.66\mu m$ (écart type de $75.60\mu m$). Le test U de Mann-Whitney indique que ces distributions sont significativement différentes ($U = 201$, $p$-value $= 1.85e-09$).

Les histogrammes centraux des clusters obtenus par $f^{(3)}$ révèlent que le représentant du cluster homogène $\mathcal{S}_1^{(3)}$ a une distance maximale entre aggrégats de lymphocytes bien plus petite (environ $180\mu m$) que le représentant du cluster hétérogène (environ $390\mu m$). Les courbes de survie obtenues révèlent que ce dernier cluster a un pronostic de survie bien moindre que le premier. Une analyse des histogrammes de ces représentants révèle des aggrégats de lymphocytes plus proches les uns des autres chez le représentant du cluster homogène, suggérant une répartition régulière des aggrégats de lymphocytes, afin de faire barrière à la progression de la tumeur, *i.e.*, une réponse immunitaire. Nous avons listé la distance inter-aggrégats maximale pour chaque histogramme, et avons analysé les distributions de ces valeurs par cluster. Ainsi, $\mathcal{S}_1^{(3)}$ a une moyenne de $309.639\mu m$, (écart type de $213.06\mu m$); et $\mathcal{S}_2^{(2)}$ a une moyenne de $431.16\mu m$ (écart type de $274.72\mu m$). Le test U de Mann-Whitney indique que ces distributions sont significativement différentes ($U = 611$, $p$-value $= 0.008$).

Enfin, la combinaison des attributs permet une très bonne séparation des sous-populations, suggèrant que les sous-populations tendent à être cohérentes d'un attribut à l'autre. Cette agrégation offre une approche intéressante pour une estimation multi-critères du pronostic.

## 4 Conclusion

Nous avons proposé une méthodologie pour regrouper des sujets via des histogrammes. Nous avons illustré notre approche sur une population de sujets atteints d'adénocarcinome pulmonaire de stade I, et avons obtenu des clusters cohérents au regard des connaissances existantes en matière d'estimation du pronostic, avec une confiance statistique élevée, tout en permettant l'interprétabilité des résultats. Notre méthode permet l'exploration de nouvelles hypothèses visant à lier l'organisation des cellules et le pronostic, et offre une approche systémique pour comparer des sujets. Elle peut donc aider les oncologues dans leur analyse.

Les suites à ces travaux sont nombreuses. Une direction consiste en l'extension du catalogue d'attributs. Une autre direction est l'analyse de données plus riches, comme les images mIHC, ainsi que l'étude d'autres familles de cancers. Enfin, nous voulons explorer des variations des éléments de notre approche, en changeant par exemple le nombre de clusters produits.

# Références

Fridman, W. H., F. Pagès, C. Sautes-Fridman, et J. Galon (2012). The immune contexture in human tumours : impact on clinical outcome. *Nature Reviews Cancer 12*(4), 298–306.

Galon, J., F. Pagès, F. M. Marincola, M. Thurin, G. Trinchieri, B. A. Fox, T. F. Gajewski, et P. A. Ascierto (2012). The immune score as a new possible approach for the classification of cancer. *Journal of Translational Medicine 10*.

Kaplan, E. L. et P. Meier (1958). Nonparametric estimation from incomplete observations. *Journal of the American statistical association 53*(282), 457–481.

Mantel, N. (1966). Evaluation of survival data and two new rank order statistics arising in its consideration. *Cancer Chemother. Rep. 50*, 163–170.

Mercadier, D. S., B. Besbinar, et P. Frossard (2019). Automatic segmentation of nuclei in histopathology images using encoding-decoding convolutional neural networks. In *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1020–1024. IEEE.

Mlecnik, B., G. Bindea, F. Pagès, et J. Galon (2011). Tumor immunosurveillance in human cancers. *Cancer and Metastasis Reviews 30*(1), 5–12.

Peyré, G., M. Cuturi, et al. (2019). Computational optimal transport. *Foundations and Trends® in Machine Learning 11*(5-6), 355–607.

Reichling, C., J. Taieb, V. Derangere, Q. Klopfenstein, K. Le Malicot, J.-M. Gornet, H. Becheur, F. Fein, O. Cojocarasu, M. C. Kaminsky, et al. (2020). Artificial intelligence-guided tissue analysis combined with immune infiltrate assessment predicts stage iii colon cancer outcomes in petacc08 study. *Gut 69*(4), 681–690.

Saltz, J., R. Gupta, L. Hou, T. Kurc, P. Singh, V. Nguyen, D. Samaras, K. R. Shroyer, T. Zhao, R. Batiste, et al. (2018). Spatial organization and molecular correlation of tumor-infiltrating lymphocytes using deep learning on pathology images. *Cell reports 23*(1), 181–193.

Szekely, G. J. et M. L. Rizzo (2005). Hierarchical clustering via joint between-within distances : Extending ward's minimum variance method. *Journal of classification 22*(2).

Wang, S., T. Wang, L. Yang, D. M. Yang, J. Fujimoto, F. Yi, X. Luo, Y. Yang, B. Yao, S. Lin, et al. (2019). Convpath : A software tool for lung adenocarcinoma digital pathological image analysis aided by a convolutional neural network. *EBioMedicine 50*, 103–110.

Yener, B. (2016). Cell-graphs : image-driven modeling of structure-function relationship. *Communications of the ACM 60*(1), 74–84.

# Summary

We introduce an interpretable methodology to cluster subjects suffering from cancer, based on features extracted from their biopsies. We capture patterns in cell distributions using histograms, and compare subjects based on these. We describe here a workflow to define these histograms and use them for clustering purposes. We illustrate our approach on a database of hematoxylin and eosin-stained tissues of subjects with Stage I lung adenocarcinoma, where our results match existing knowledge with high confidence.