Laboratory Investigation (2011) 91, 799–811

Automation of an algorithm based on fuzzy clustering for analyzing tumoral heterogeneity in human skin carcinoma tissue sections

David Sebiskveradze[1], Valeriu Vrabie[2], Cyril Gobinet[1], Anne Durlach[3], Philippe Bernard[4], Elodie Ly[1], Michel Manfait[1], Pierre Jeannesson[1] and Olivier Piot[1]

This study aims to develop a new FT–IR spectral imaging of tumoral tissue permitting a better characterization of tumor heterogeneity and tumor/surrounding tissue interface. Infrared (IR) data were acquired on 13 biopsies of paraffin-embedded human skin carcinomas. Our approach relies on an innovative fuzzy C-means (FCM)-based clustering algorithm, allowing the automatic and simultaneous estimation of the optimal FCM parameters (number of clusters *K* and fuzziness index *m*). FCM seems more suitable than classical 'hard' clusterings, as it permits the assignment of each IR spectrum to every cluster with a specific membership value. This characteristic allows differentiating the nuances in the assignment of pixels, particularly those corresponding to tumoral tissue and those located at the tumor/peritumoral tissue interface. FCM images permit to highlight a marked heterogeneity within the tumor and characterize the interconnection between tissular structures. For the infiltrative tumors, a progressive gradient in the membership values of the pixels of the invasive front was also revealed.
Laboratory Investigation (2011) **91,** 799–811; doi:10.1038/labinvest.2011.13; published online 28 February 2011

KEYWORDS: clustering methods; FT–IR spectroscopy; fuzziness index; fuzzy C-means; number of clusters; skin cancers

Fourier transform mid-infrared (FT–IR) microspectroscopy is a label-free optical method based on the interaction between an IR radiation and matter. This vibrational spectroscopy permits to probe the biochemical composition of a sample. Coupled with an imaging system, FT–IR microspectroscopy of human tissues can be used as a very sensitive, non-destructive and non-subjective tool for the detection and localization of tumoral nests. Thus, FT–IR microimaging has demonstrated potential to provide clinically relevant diagnostic information in oncology.[1–15]

The biochemical changes related to carcinogenesis between cancerous and surrounding tissue areas are subtle. As a consequence, IR hyperspectral images need to be processed by powerful digital signal processing and pattern recognition methods in order to highlight these changes. To date, unsupervised 'hard' clustering techniques including K-means (KM)[3,16–20] or agglomerative hierarchical (AH)[11,16,21–23] clustering have been usually applied to create color-coded images allowing to localize tumoral tissue surrounded by other tissue structures (normal, inflammatory, fibrotic).

The particularity of 'hard' clustering methods is that each pixel is assigned to only one cluster. Consequently, they neither allow to consider the progressive transition between non-cancerous tissues and cancer lesions nor to reveal every nuance of intratumoral heterogeneity.[2]

To overcome this drawback, fuzzy clustering methods such as fuzzy C-means (FCM) can be used instead of 'hard' clustering algorithms.[24] Indeed, FCM allows each pixel to be assigned to every cluster with an associated membership value varying between 0 (no class membership) and 1 (highest degree of cluster membership). FCM has been successfully used in near-IR spectroscopy to distinguish different types of inks in synthesized samples[25] and in mid-IR spectroscopy to characterize adhesive/dentin interface in caries-affected teeth,[26] and to analyze different types of tumor tissues.[16,27] However, as it is the case for 'hard' KM clustering, the number of clusters *K* must be defined *a priori* by the user. The FCM results are thus dependent from the operator experience. In addition, FCM outcomes are dependent on

[1]Unité MéDIAN, CNRS UMR 6237 MEDyC, UFR de Pharmacie, Université de Reims Champagne-Ardenne, Reims, France; [2]CReSTIC, UFR de Sciences Exactes et Naturelles, Université de Reims Champagne-Ardenne, Reims, France; [3]Department of Pathology Pol Bouin, Maison Blanche University Hospital, Reims, France and [4]Department of Dermatology, Robert Debré University Hospital, Reims, France
Correspondence: Dr V Vrabie, PhD, CReSTIC, UFR de Sciences Exactes et Naturelles, Université de Reims Champagne-Ardenne, BP 1039, 51 687 Reims Cedex 2, France. E-mail: valeriu.vrabie@univ-reims.fr

Received 10 June 2010; revised 14 December 2010; accepted 29 December 2010

another important parameter, called the fuzziness index m in the fuzzy logic literature. When $m = 1$, FCM becomes identical to KM and when m increases, the clustering becomes fuzzier. When m tends to infinity, each pixel tends to have its membership values uniformly distributed to all the clusters.[24] In IR data processing, this can create redundant cluster images, in which only some pixels differ from one cluster to another. However, the fuzziness index is classically fixed to 2 in the literature.[24,25,27] The choice of an efficient trade-off between K and m, necessary to fully exploit the information content of IR hyperspectral images, is still an open problem. Indeed, as recently shown for colorectal adenocarcinoma,[16] when the (K, m) couple is not optimized, FCM clustering proved to be less efficient than AH clustering in terms of tissue histopathological recognition.

In this article, an algorithm dedicated to IR spectral images of tumoral tissues is developed in order to automatically estimate the optimal values of K, the number of non-redundant FCM clusters, and m, the fuzziness index, without any *a priori* knowledge of the data set. This innovative algorithm is based on the redundancy between FCM clusters. Results obtained from human skin cancer-tissue sections indicate that this algorithm is particularly well adapted to localize tumoral areas and to highlight transition areas between tumor and surrounding tissue structures. These transition areas are of crucial importance in the promotion of tumor progression, malignant cell escape and consequently metastasis formation.

MATERIALS AND METHODS

Sample Preparation

The developed algorithm was applied on the IR spectral images acquired on 13 biopsies of formalin-fixed paraffin-embedded human skin carcinomas, squamous cell carcinomas (SCC, $n = 3$), basal cell carcinomas (BCC, $n = 4$) and Bowen's diseases ($n = 6$). The samples were selected by the pathologists from the tumor bank of the Pathology Department of the University Hospital of Reims (France). From samples 10-micron thick slices were cut and mounted, without any particular preparation, especially no chemical dewaxing, on a calcium fluoride (CaF_2) (Crystran, Dorset, UK) window for FT–IR imaging. First adjacent slices (5-μm thick) to those used for FT–IR analysis were stained with hematoxylin and eosin (H&E) for conventional histology, except for the infiltrative SCC #1, for which H&E-stained tissue section is not immediately adjacent to the analyzed section, but localized at *circa* 25 μm. From these slices, the cancer outlines defined by the pathologists were drawn on the photomicrographs.

FT–IR Data Set Acquisition

FT–IR hyperspectral images were recorded with a Spectrum Spotlight 300 FT–IR imaging system coupled to a Spectrum one FT–IR spectrometer (Perkin Elmer Life Sciences, France), with a spatial resolution of 6.25 μm and a spectral resolution of 4 cm^{-1}. The device is equipped with a nitrogen-cooled mercury cadmium telluride 16-pixel-line detector for imaging. Spectral images, also called data sets, were collected using 16 accumulations. Before each acquisition, a reference spectrum of the atmospheric environment and the CaF_2 window was recorded with 240 accumulations. This reference spectrum was subsequently subtracted from each data set automatically by a built-in function from the Perkin Elmer Spotlight software. Each spectral image, covering a substantial part of the biopsy, consisted of about 30 000 spectra. Each image pixel represents an IR spectrum, which is the absorbance of one measurement point ($6.25 \times 6.25\ \mu m^2$) over 451 wavenumbers uniformly distributed between 900 and 1800 cm^{-1}. This spectral range, characterized as the fingerprint region, actually corresponds to the most informative region for biological samples.

Data Set Preprocessing

The samples being analyzed without previous chemical dewaxing, the recorded FT–IR hyperspectral image must be digitally corrected for paraffin spectral contribution. To this end, an automated processing method based on extended multiplicative signal correction (EMSC) was applied on each recorded data set.[3] Briefly, the mean spectrum was computed by averaging all recorded spectra of each data set. Light-scattering effects were modeled with a fourth-order polynomial function. The interference matrix was composed of the average spectrum of paraffin and the first nine principal components extracted from a FT–IR spectral image recorded on a pure paraffin block, in order to take into account the spectral variability of paraffin. After the application of the EMSC-based preprocessing, paraffin contribution is neutralized, which permits to retain in the data sets only the spectral variability of the tissue and to normalize the corrected spectra around the mean spectrum.[3] Two IR spectra before and after EMSC-based preprocessing are shown in Supplementary Figure S1, available in the supplementary information document.

In addition, this preprocessing permits to discard from the analysis outliers and poor tissue signal to noise ratio spectra.[3] The corresponding pixels are white colored on the clustering color-coded images for better visualization.

Clustering Methods

The spectral differences between different skin structures (such as dermis, epidermis and tumor) are weak after the EMSC-based preprocessing step. To highlight the different biological structures of the analyzed sample, clustering methods can be used. The main objective of clustering is to group together similar spectra in order to reveal areas of interest within tissue sections. In IR spectral imaging of cancerous tissues, clustering methods allow to create highly contrasted color-coded images permitting to localize tumoral areas within a complex tissue.[3,16] For cluster assignment, each color-coded map was then provided to the pathologists for a comparison with the corresponding H&E-stained sections.

'Hard' Clustering

KM clustering is the most popular non-hierarchical partition clustering method because of simple algorithm and fast execution speed. The aim of KM is to divide the spectra into a partition composed of K clusters that minimizes an objective function based on a distance measure between each spectrum and the nearest cluster centroid.[28] Its major drawback is that the results depend on a random initialization, and the number of clusters must be fixed in advance by the user. In this study, KM clustering was performed several times ($n>10$) to make sure a stable solution was reached, and to overcome the random initialization dependence. In this study, KM was applied using the Matlab Statistics Toolbox with the classical Euclidean distance. The process continued until no spectrum was reassigned from one iteration to the following, otherwise it was stopped after 10^4 iterations.

AH clustering is a hierarchical partition clustering, in which each object (spectrum in our case) is one cluster at the beginning of the algorithm. At each iteration step, AH regroups the two clusters that are the most similar into a new cluster. The algorithm is stopped when the all spectra are combined into one single cluster.[29] For n spectra, the number of iterations equals to n–1. AH clustering process is independent of initialization. However, like for KM, in AH clustering, the number of clusters K is empirically chosen. Compared with KM, AH clustering is significantly more time and resource consuming. To reduce the computational time of AH clustering on our large data set, we used here an efficient hybrid hierarchical agglomerative clustering (HHAC) technique that combines KM and AH clusterings using Euclidean distance and Ward's algorithm.[30] KM is first applied to reduce the data sets to 1000 cluster centers. AH is then carried out on these 1000 KM centroids.

In addition, a common characteristic feature of 'hard' KM and AH clusterings methods is that each spectrum belongs to an unique cluster. This feature becomes a limitation in case of tissular IR spectra, probing the multichemical composition of the tissue at the microscopic scale.

FCM Clustering

FCM clustering is based on the minimization of the sum of weighted distance measures between each spectrum and each centroid. The weight is controlled by the fuzziness index m. Therefore, contrary to 'hard' clustering, FCM permits to affect each pixel (spectrum) to every cluster with an associated membership value varying between 0 and 1; the sum of the K cluster membership values for one pixel being equal to 1.[24]

In this study, we applied the FCM function from the Matlab Statistics Toolbox with the Euclidean distance. A maximum number of 500 iterations and a setting of 10^{-5} for the minimal amount of improvement of the objective function between two consecutive iterations were used as the stopping criteria.

However, FCM requires to fix the number of clusters K and the fuzziness index m. An inappropriate choice of these parameters could lead to an uninterpretable clustering of the data. The development of an automatic method to optimally estimate these parameters is thus essential.

Development of the Redundancy-Based Algorithm for the Optimal Estimation of FCM Parameters

The redundancy-based algorithm (RBA) is an innovative algorithm proposed to automatically estimate the optimal couple (K_{opt}, m_{opt}) of the FCM. It is based on the FCM clusters redundancy measured in this paper for two clusters i and j by the intercorrelation coefficient $\mathbf{R_{ij}}$:

$$\mathbf{R_{ij}}(K,\ m) = \frac{\mathbf{C}(i,j)}{\sqrt{\mathbf{C}(i,\ i)\mathbf{C}(j,j)}}$$

where $\mathbf{C}(i,\ j)$ is the covariance between the membership values of clusters i and j, and $\mathbf{C}(i,\ i)$ (respectively $\mathbf{C}(j,\ j)$) is the variance of the membership values of cluster i (respectively cluster j).

The RBA is composed of three steps. The first one can be mainly divided into three imbricated loops, and performs an iterative process of cluster-number reduction. For this step, N different values of the fuzziness index m uniformly distributed around the classical value $m=2$ are considered and form the set $\mathbf{m}=\{m_1, \ldots, m_n, \ldots, m_N\}$. In the same manner, L different values of the correlation coefficient threshold s uniformly distributed into the high correlation coefficients range 50 to 95% compose the set $\mathbf{s}=\{s_1, \ldots, s_l, \ldots, s_L\}$. Furthermore, a parameter of this algorithm is the maxi-

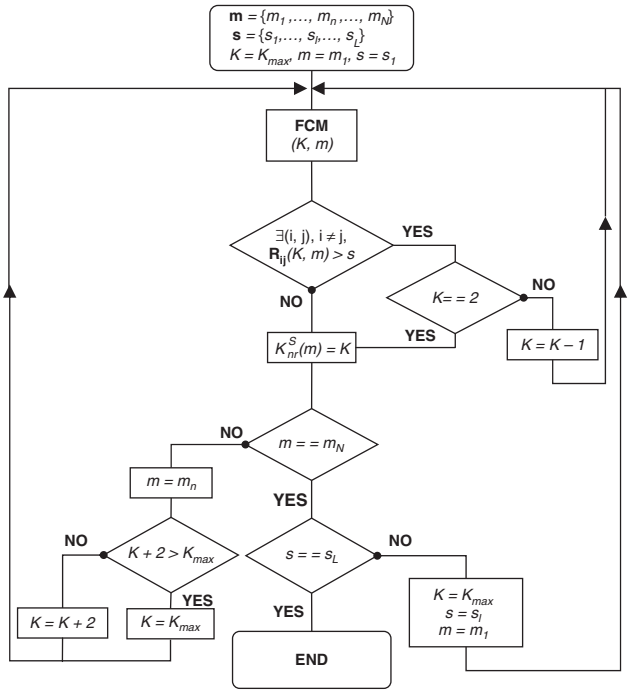


Figure 1 Flowchart of the redundancy-based algorithm (RBA). This flowchart permits to construct the curves of the number of non-redundant clusters $K^s_{nr}(m)$ as a function of m for different values of the threshold s.

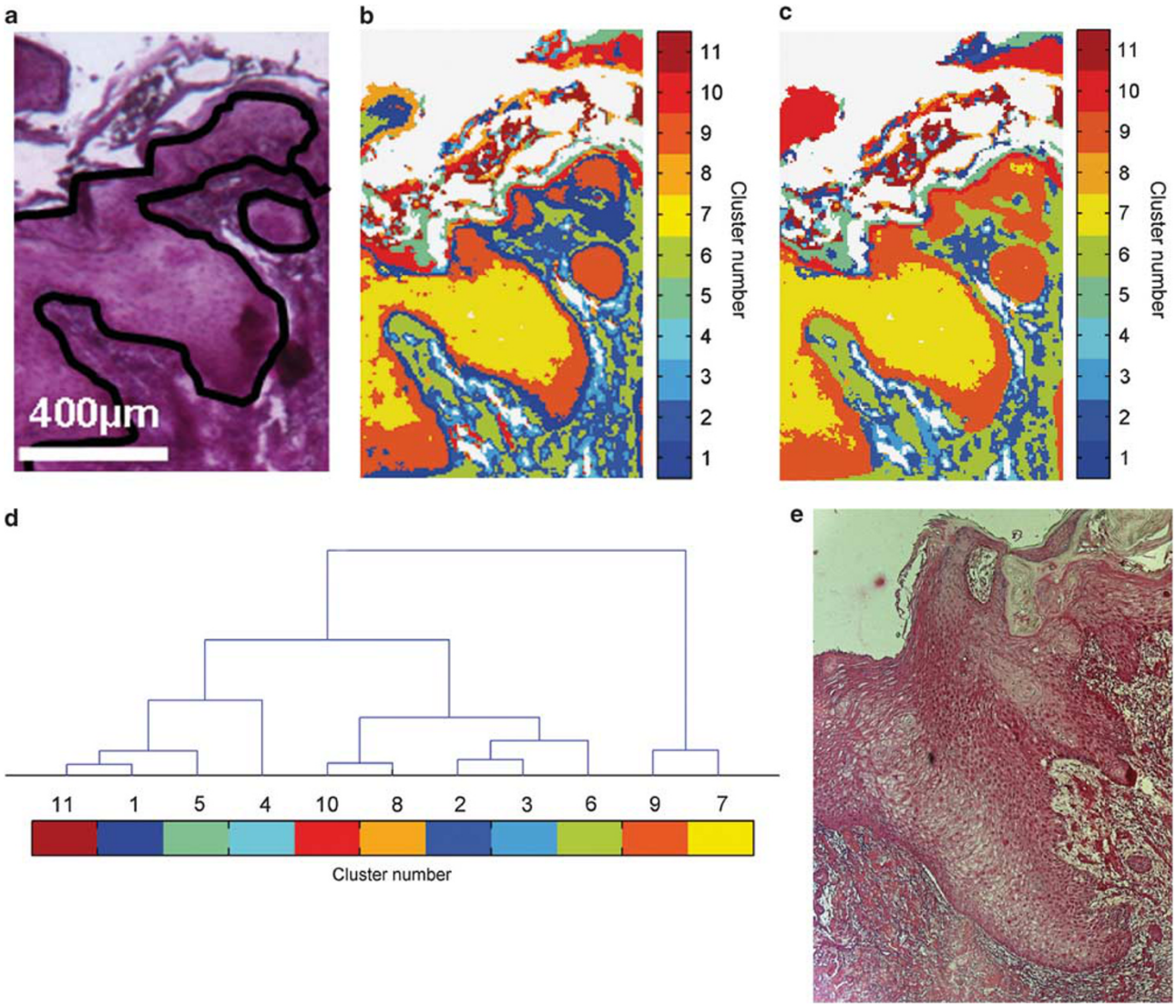


Figure 2 'Hard' clustering color-coded images on the Fourier transform mid-infrared (FT–IR) data set of an infiltrative human skin squamous cell carcinoma (SCC) sample #1. (**a**) Direct hematoxylin and eosin (H&E) staining of the thick (10 μm) section examined by infrared (IR) imaging; tumor is outlined. (**b**) K-means (KM) color-coded image. (**c, d**) Hybrid hierarchical agglomerative clustering (HHAC) color-coded image and its corresponding dendrogram. Each color corresponds to one cluster. (**e**) H&E staining of a thin section (5 μm) localized at *circa* 25 μm from that analyzed by IR imaging.

mum number of considered clusters noted K_{max} and fixed by the user.

The RBA begins by initializing m to m_1, s to s_1 and K to K_{max}. Thereafter the most internal loop can start. FCM are run with the fuzziness index equal to m and a number of clusters equal to K. Once the FCM algorithm has converged, if one of the computed $\mathbf{R_{ij}}$ values is superior to the threshold equal to s, it means that two clusters are redundant. One of them is thus useless. FCM are thus run again with the same m, but with a number of clusters equal to the previous value minus 1. On the contrary, if all the computed $\mathbf{R_{ij}}$ values are inferior to s, it means that no clusters are redundant. All the estimated cluster carrying different information, this first loop can hence be stopped. The current value of the number of clusters K is saved as the number of non-redundant clusters noted $K^s_{nr}(m)$ (note that for the first loop, $s = s_1$ and $m = m_1$, and for the following iterations $s = s_l$ and $m = m_n$).

The middle loop of the algorithm is a repetition of the most internal one for the different values of m among the set $\mathbf{m}$. To gain computation time, the most internal loop should begin for the next value of m with an initial number of clusters K equal to the number of non-redundant clusters estimated for the previous value of m. However, the FCM algorithm being randomly initialized, the estimated number of non-redundant clusters can vary from one clustering to another. To take into account this possible variation, the initial value of K for the next m is set to the number of non-redundant clusters for the previous m plus two. Note

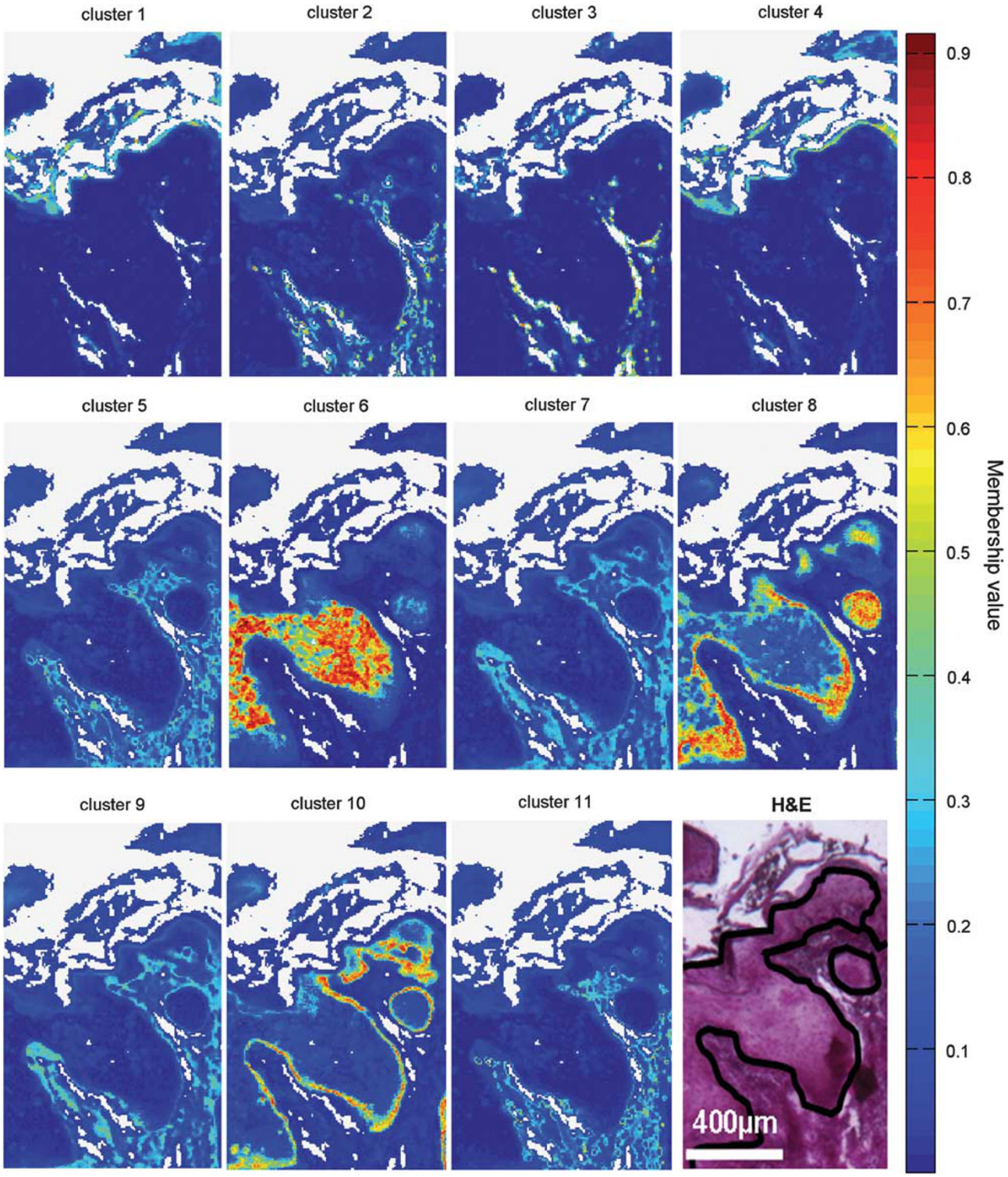


Figure 3 Fuzzy C-means (FCM) images with unoptimized parameters ($K = 11$ and $m = 2$) on the Fourier transform mid-infrared (FT–IR) dataset of the human skin squamous cell carcinoma (SCC) sample #1, and the corresponding hematoxylin and eosin (H&E)-stained section. Clusters 1 and 4 are redundant clusters of the epidermis, whereas cluster 3 is non-redundant. For the dermis, clusters 2, 5 and 11 are redundant, as for clusters 7 and 9. Clusters 6, 8 and 10 are dissociated clusters describing the tumor. The color bar represents the membership value for each pixel. In the corresponding H&E-stained section, tumor is outlined.

that the threshold s keeps its value in this loop. By performing this procedure for the different values of m, a curve $K^s_{nr}(\mathbf{m})$ of the number of non-redundant clusters $K^s_{nr}(m)$ is obtained as a function of m with a threshold equals to s.

The last loop (the most external one) is a repetition of the first two ones for the different values of s among the set **s**. At the beginning of each iteration of this loop, the number of clusters K is initialized to K_{max} and m to m_1. At the end of the algorithm, the L resulting $K^s_{nr}(\mathbf{m})$ curves are obtained for each threshold value s. The complete flowchart of this global procedure is shown in Figure 1.

The second step of the RBA consists in the optimal estimation of FCM parameters from the obtained curves $K^s_{nr}(\mathbf{m})$. As it will be presented in the Results and discussion section, each curve decreases rapidly and becomes stable at a $\hat{K}^s_{opt}$ value that can be different from one threshold to another.

However, in practice, whatever the threshold s, we usually observe that the breakings in these curves appear for close and often for the same number of clusters. A majority voting algorithm can thus be used to identify the final optimal value $\hat{K}_{opt}$ of the number of clusters. The optimal value $\hat{m}_{opt}$ of the fuzziness index is then computed as the average of values $\hat{m}^s_{opt}$ for which the curves $K^s_{nr}(\mathbf{m})$ have presented a break at $\hat{K}^s_{opt}$.

Hereafter in the manuscript, FCM clustering carried out with these RBA-optimized parameters will be defined as FCM–RBA.

RESULTS AND DISCUSSION

The FCM–RBA clustering was assessed on EMSC-preprocessed FT–IR hyperspectral images acquired on thin tissue sections of 13 human skin carcinomas. The results were compared with KM, HHAC and classical FCM outcomes.

To improve the reading of this section, we present these comparative results for one infiltrative SCC #1. For a superficial state of a BCC #1 and a Bowen's disease #1, only FCM–RBA clustering data are given, whereas corresponding KM, HHAC and FCM outcomes are shown in the Supplementary information (Supplementary Figures S2–S5). In addition, the FCM–RBA results of the remaining samples (infiltrative SCC #2 and #3, superficial BCC #2–4 and Bowen's diseases #2–6) are shown in Supplementary Figures S6–S15. The histological characteristics of the studied human skin cancers are indicated in Supplementary Table S1.

'Hard' Clustering Results

The H&E-stained histological image of the studied SCC sample #1, on which the tumor is outlined, is shown in Figure 2a.

To highlight the distinctive histological regions of this paraffin-embedded tissue section, KM clustering was applied with an empirical choice of 11 clusters as already described by our group.[4] To generate comparable results, the dendrogram of the HHAC clustering was cut to 11 clusters. The resulting color-coded images are shown in Figures 2b for KM and 2c for HHAC. Each color is associated to one cluster. The corresponding dendrogram used to construct the HHAC color-coded image is shown in Figure 2d. In addition, Figure 2e corresponds to the H&E staining of a thin section (5 μm), not immediately adjacent to the analyzed section, but localized at *circa* 25 μm.

The comparison of KM and HHAC images with the corresponding H&E-stained section permits an assignment of the clusters. As shown for KM clustering in Figure 2b, the pixels belonging to the tumor are grouped into clusters 1, 7 and 9, revealing an intratumoral heterogeneity. The dermis is represented by clusters 2, 3 and 6, and the ulcerated epidermis by clusters 4, 5, 8, 10 and 11. As shown in Figure 2c, HHAC clustering results are quite similar to those of KM.

These results indicate that 'hard' clustering algorithms are able to retrieve the histological structures, and especially to localize tumoral areas within the tissue section. However, the choice of the number of clusters is a difficult problem that is usually empirically resolved. When less than 11 clusters are chosen, the histological regions identified by clustering

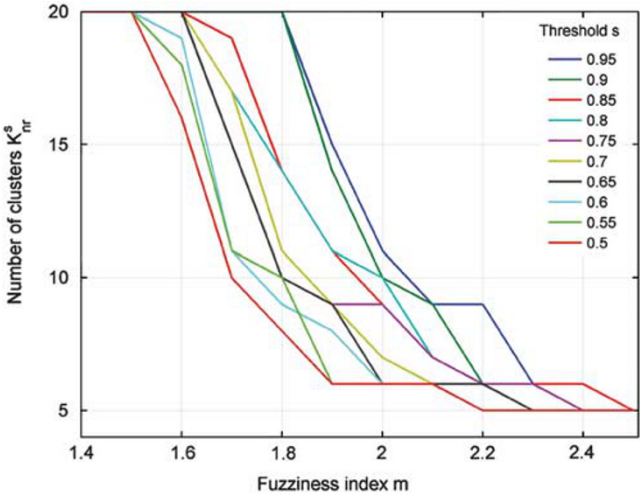


Figure 4 Number of non-redundant clusters $K^s_{nr}(m)$ as a function of the fuzziness index m estimated by the redundancy-based algorithm (RBA) for the squamous cell carcinoma (SCC) sample #1. Each curve corresponds to a given value of the threshold s.

Table 1 Optimal parameters of FCM estimated by RBA in function of the threshold *s* for the human skin SCC sample #1

S	0.95	0.9	0.85	0.8	0.75	0.7	0.65	0.6	0.55	0.5
$\hat{K}^s_{opt}$	9	6	6	6	9	6	6	6	6	6
$\hat{m}^s_{opt}$	2.1	2.2	2.2	2.2	1.9	2.1	2	2	1.9	1.9

Optimal number of clusters $\hat{K}^s_{opt}$ and the corresponding optimal values of the fuzziness index $\hat{m}^s_{opt}$ have been determined for 10 different values of the threshold s from the curves shown in Figure 3.

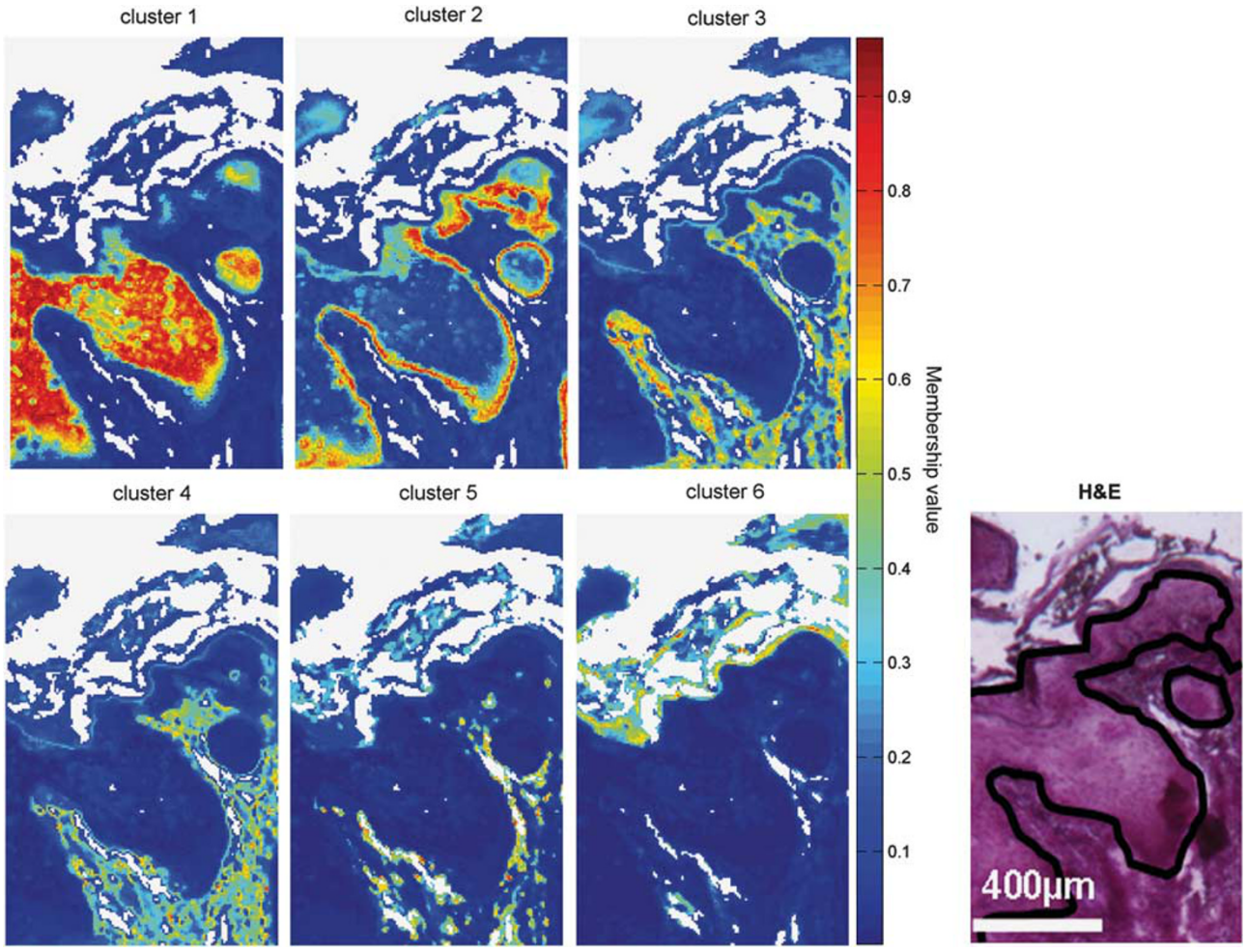


Figure 5 Fuzzy C-means (FCM) images on the Fourier transform mid-infrared (FT–IR) data set of the human skin squamous cell carcinoma (SCC) sample #1 constructed with redundancy-based algorithm (RBA)-optimized parameters $\hat{K}_{opt} = 6$ (number of clusters) and $\hat{m}_{opt} = 2.1$ (fuzziness index) and the corresponding hematoxylin and eosin (H&E)-stained section. Assignment of the clusters: cluster 1 (tumor); 2 (invasive front); 3, 4 and 5 (dermis); 6 (epidermis). The color bar represents the membership value for each pixel. In the corresponding H&E-stained section, SCC is outlined.

algorithms are mixed, and the intratumoral heterogeneity is no more revealed (data not shown). With more than 11 clusters, no further interpretable information is obtained (data not shown). Furthermore, the principal drawback of these 'hard' clustering methods is that the cluster membership grade of each individual spectrum equals to 0 or 1. Nuances of pixel membership are thus not accessible with 'hard' clustering methods. Consequently, these techniques do not allow to consider progressive transitions likely to exist at the invasion front of a tumor or between heterogeneous intratumoral areas.

Classical FCM Clustering

The results obtained by using the FCM algorithm without optimized parameters on the same data set are shown in Figure 3. The fuzziness index m was fixed to the commonly used default value of 2, according to investigations of other groups.[31] A total of 11 clusters were chosen as they allow an unequivocal reproduction of the H&E-based histology as previously described with 'hard' clusterings (Figure 2). Each cluster is presented into a separate image instead of superimposing them into only one color-coded image as reported by others.[16,27,31] Indeed, the superimposition presentation makes it difficult to highlight transitional structures.

A visual comparison of the clusters shown in Figure 3 reveals important redundancies. This was confirmed by the intercorrelation coefficients $\mathbf{R_{ij}}$ between the computed images. Indeed, clusters 7 and 9 are correlated with a $\mathbf{R_{ij}}$ coefficient equal to 98.3%, 5 and 7 with 82.6%, 5 and 11 with 78.6% and finally 1 and 4 with 76.7%. Similarly, redundancies have also been observed between certain FCM cluster pairs on all examined skin-cancer samples. This is supported by the FCM redundant images shown in the supplementary information for two other representative cancer samples (Supplementary Figure S3 for BCC #1 and Supplementary Figure S5 for Bowen's disease #1); and by the

Supplementary Table S2, which shows the maximal values of the intercorrelation coefficients for all the samples.

These results demonstrate that classical FCM creates non-informative redundant images, in which only few pixels differ from one cluster to another when K and m are incorrectly chosen. Therefore, it is essential to choose the optimal values of K and m parameters to obtain a biologically relevant clustering.

Optimization of FCM Parameters Using RBA

Simultaneous determination of optimal K and m parameters was performed using the innovative RBA. In our investigation, varying settings of the number of clusters from 2 to 20, and varying values of m from the set $\mathbf{m} = \{1.4, 1.5, \ldots, 2.5\}$ were tested. The curves $K_{nr}^{s}(\mathbf{m})$, representing the number of non-redundant clusters as a function of m obtained by this method for 10 different values of threshold s from the set $\mathbf{s} = \{0.5, 0.55, \ldots, 0.95\}$ are shown in Figure 4 for the SCC sample #1. Each curve tends to quickly decrease towards a $\hat{K}_{opt}^{s}$ value, from which each curve becomes quite stable. The $\hat{K}_{opt}^{s}$ values and the corresponding $\hat{m}_{opt}^{s}$ values for these thresholds are indicated in Table 1. The optimal number of clusters $\hat{K}_{opt}$ has thus been estimated by using a majority voting algorithm as equal to 6. The resulting optimal value $\hat{m}_{opt}$ is determined as the average of the values of $\hat{m}_{opt}^{s}$ obtained for $\hat{K}_{opt}^{s} = 6$, and is equal to 2.1. This developed RBA was applied on all remaining samples and permitted to estimate an optimal couple of values, K and m, for each analyzed IR image.

It has to be mentioned, that in our case, classical methods used to determine the optimal number of FCM clusters K failed to correlate with standard histopathology. Indeed, the partition coefficient and classification entropy[24] applied with $m = 2$ give an aberrant value of $K = 2$ that does not permit to reveal the different tissue structures. These data reinforce the relevancy of our developed RBA in terms of tissue-structure differentiation.

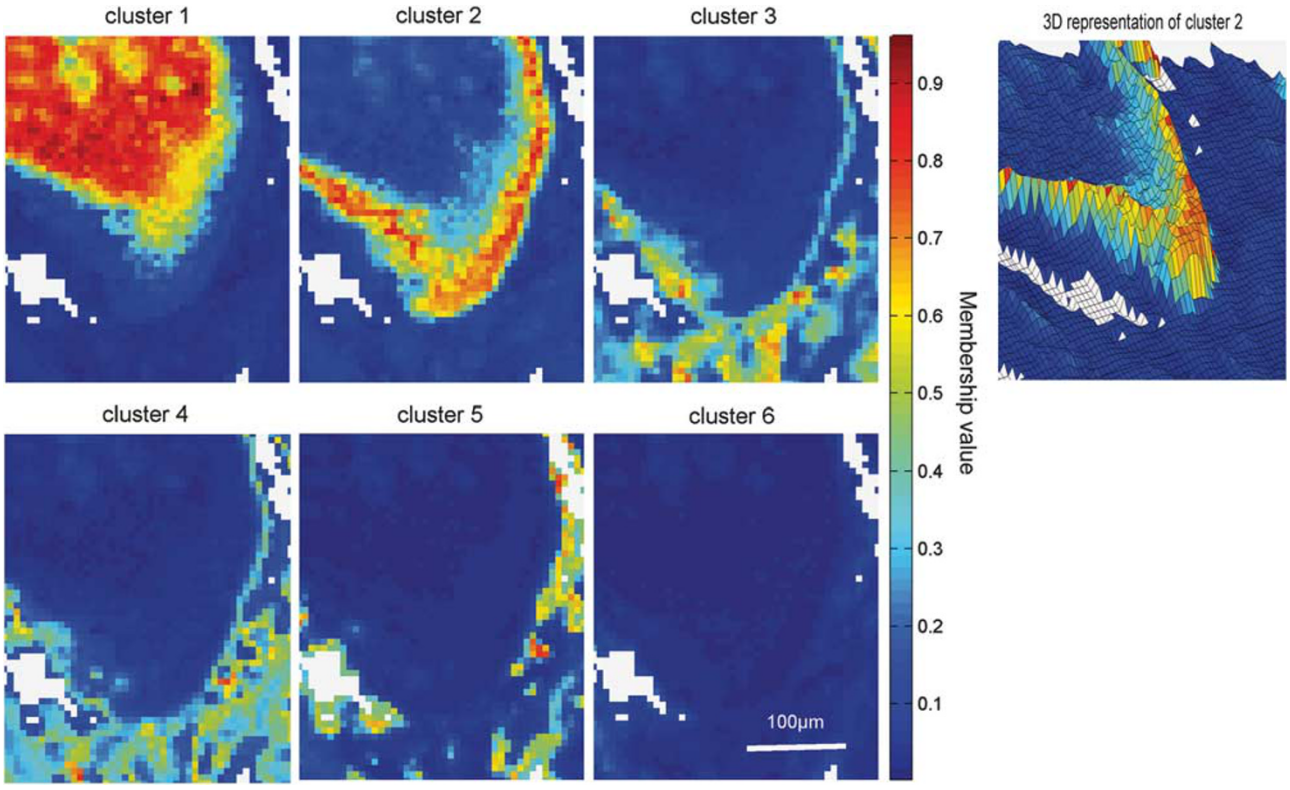
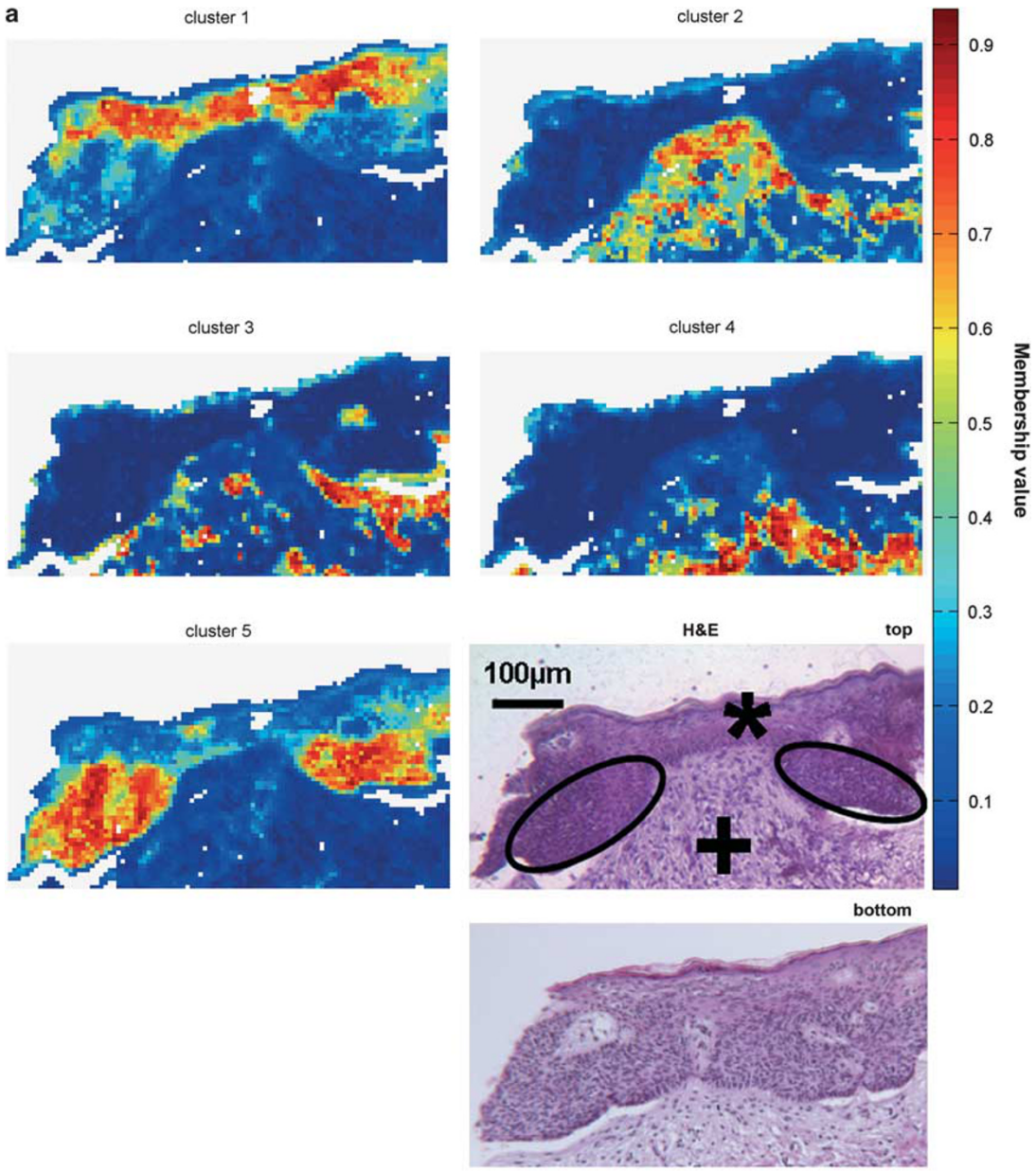
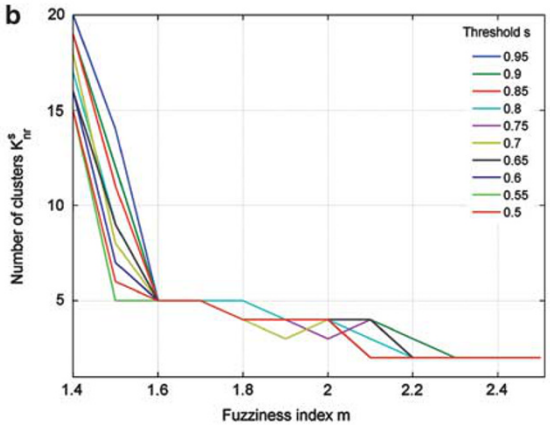


Figure 6 Analysis of the tumor/surrounding dermis interface by zooming the fuzzy C-means (FCM) images shown in Figure 5. Cluster 2, characterizing the invasive front of the tumor is also shown in a 3-D representation. The color bar represents the membership value for each pixel.

Figure 7 Redundancy-based algorithm (RBA) results on the Fourier transform mid-infrared (FT–IR) data set of the human skin superficial basal cell carcinoma (BCC) sample #1. Fuzzy C-means (FCM) images (**a**) were constructed with optimized parameters $\hat{K}_{opt} = 5$ and $\hat{m}_{opt} = 1.6$. These parameters were defined using the RBA-resulting curves (**b**) and Table 2. Assignment of the clusters: cluster 1 (epidermis); 2, 3 and 4 (dermis); 5 (tumoral areas). The color bar represents the membership value for each pixel. In the corresponding hematoxylin and eosin (H&E)-stained section (top), BCC (outlined), epidermis (*) and dermis (+) are indicated. H&E-stained section (bottom) of higher quality localized at *circa* 25 μm from that analyzed by infrared (IR) imaging.


a
cluster 1
cluster 2
cluster 3
cluster 4
cluster 5
H&E
top
bottom
100µm
Membership value
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1


b
Threshold s
0.95
0.9
0.85
0.8
0.75
0.7
0.65
0.6
0.55
0.5
Number of clusters K^s_{nr}
20
15
10
5
1.4
1.6
1.8
2
2.2
2.4
Fuzziness index m

Histopathological Recognition of Skin Carcinomas Using FCM–RBA

The images generated by the FCM–RBA are shown in Figure 5 for the human infiltrative skin SCC #1. After comparison with the histological image, each generated cluster can be assigned to a precise tissue structure: tumoral area (cluster 1), invasive front (cluster 2), dermis (clusters 3, 4 and 5) and epidermis (cluster 6). Moreover, FCM–RBA reveals new information that is not accessible by conventional histology or classical 'hard' clustering methods. Indeed, it highlights the presence of a marked heterogeneity both within the tumor as shown for cluster 1 and within the invasive front as shown for cluster 2. Compared with 'hard' clustering, FCM–RBA allows to visualize within each of these clusters, spectral nuances corresponding to membership grade variations of the pixels. These spectral differences rely on molecular changes within tissue structures that could reflect changes in the structure/function of the tumor cells present in these areas. Interestingly, as shown in Figure 6 using a 3-D representation of the invasive front (cluster 2), FCM–RBA reveals the presence of a progressive gradient in the membership values of the pixels. From tumor towards dermis, the membership value of each pixel gradually increases to reach a maximum, and then decreases sharply at the edge of the dermis. This indicates both a tight connection between the tumor (cluster 1) and its invasive front (cluster 2), and a surprising clear-cut difference between the invasive front (cluster 2) and the surrounding dermis (clusters 3, 4 and 5). For two other infiltrative SCC samples #2 and #3 (Supplementary Figures S6 and S7, in the Supplementary information), the FCM–RBA also reveals a progressive increase of the membership values of the IR pixels within the clusters 2 (SCC #2 and SCC #3) assigned for their respective invasive front. On a pathological point of view, considering the invasive front is of great interest, as it represents a tumor area where the invasive cells can infiltrate the surrounding tissue. This approach shows significant potential for probing tumor progression, from carcinoma to metastases, and consequently may represent an attractive tool for the early determination of tumor aggressiveness.

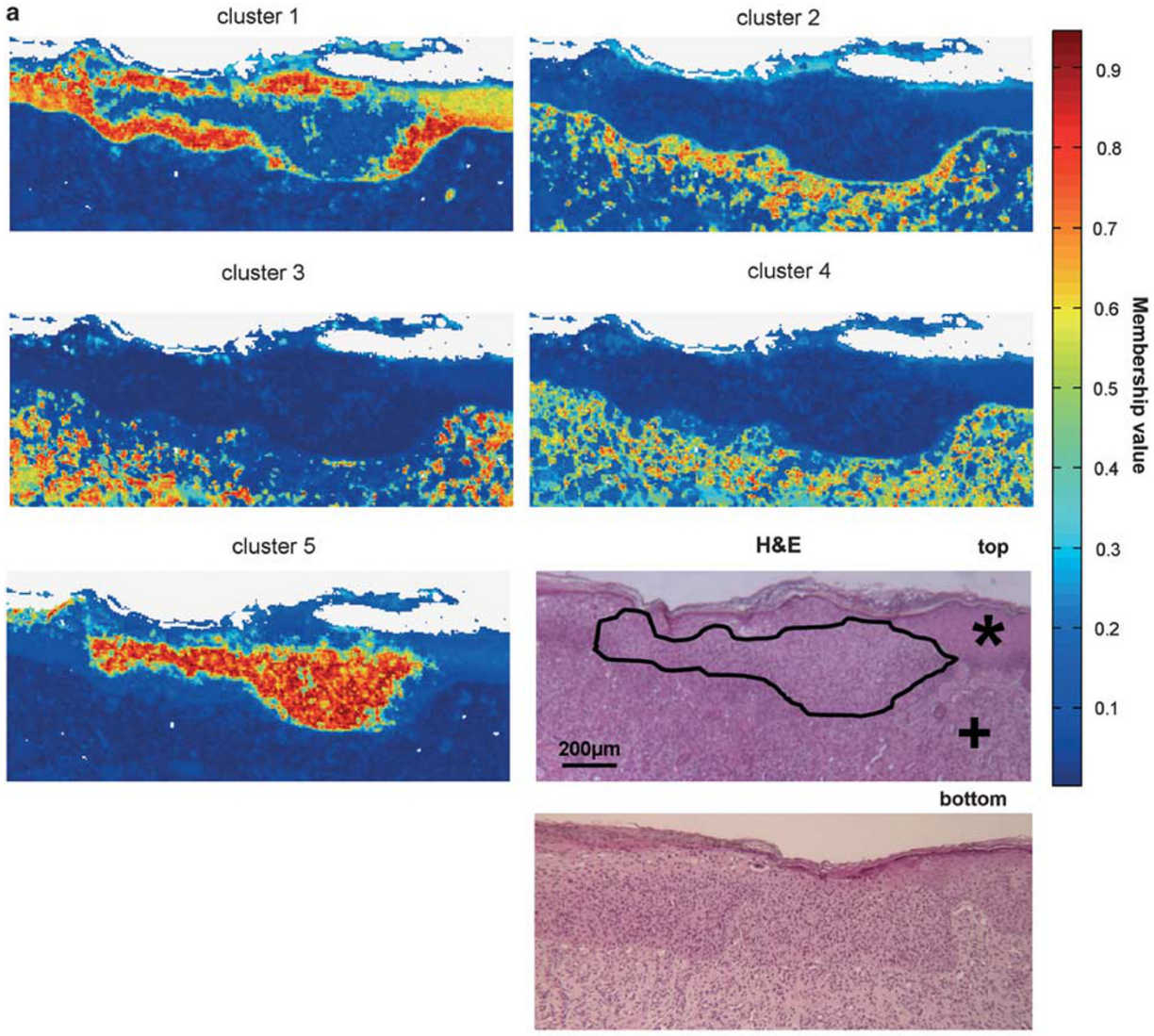
After having analyzed an SCC sample as a model of an infiltrative skin cancer, the FCM–RBA outcomes are presented to describe a superficial state of BCC and a Bowen's disease. The optimization of FCM parameters by RBA are shown for these samples in Figures 7b and 8b and in Table 2 and Table 3, for BCC #1 and Bowen's disease #1, respectively.

As shown in Figure 7a, for the superficial BCC #1, FCM–RBA reveals five clusters that can be easily assigned to following separate tissue structures: epidermis (cluster 1), dermis (clusters 2, 3 and 4) and tumoral areas (cluster 5). Compared with 'hard' clustering (Supplementary Figure S2), fuzzy clustering identifies intratumoral heterogeneities within cluster 5, as already described for cluster 1 of the previous SCC sample. An additional original information is evidenced at the tumor (cluster 5)/normal epidermis (cluster 1) interface. Indeed, a progressive transition from tumor towards epidermis is observed, reflecting an interconnectivity between these two regions. These results, directly based on molecular IR vibrational profiles of intrinsic tissue biomolecules, confirm the morphological interpretation of this connecting area observed by conventional histopathology. In addition, they can be explained by the fact that BCC originates from the transformation of epidermal keratinocytes.[32] Contrary to the infiltrative SCC, the tumor (cluster 5)/dermis (clusters 2, 3 and 4) interface does not present any intermediary structure, but rather the existence of a well-defined edge that confirms the superficial state diagnosed for this BCC sample. Similarly, for the other analyzed superficial BCC samples #2–4, (Supplementary Figures S8–S10, in the Supplementary information), both spectral characteristics of tumor interface, either with epidermis or dermis, are unambiguously identified.

For the Bowen's disease sample #1, (Figure 8a), FCM–RBA reveals five clusters that can be assigned to the following histological structures: epidermis (cluster 1), dermis (clusters 2, 3 and 4) and tumor (cluster 5). Visual comparative analysis of clusters 1 and 5 indicate that the tumor is well localized within the normal epidermis. In addition, FCM–RBA does not reveal the presence of a gradient in the membership values of the pixels at the tumor/neighboring epidermis interface. Contrary to the SCC and BCC studied samples, this absence of interconnectivity was also demonstrated for the other Bowen's samples #2–6 (Supplementary Figures S11–S15, in the Supplementary information). Such spectral features are in accordance with the fact that Bowen's diseases correspond to well-localized *in situ* carcinomas.[33]

In addition, in accordance with the pathologists, we collected data on healthy skin areas from the superficial BCC #4 and the Bowen's disease #6. Our developed FCM–RBA clearly permits to estimate the optimal couple of FCM parameters (K, m) and to retrieve the classical skin histological structures: epidermis that is associated to one cluster (with a

Figure 8 Redundancy-based algorithm (RBA) results on the Fourier transform mid-infrared (FT–IR) data set of the Bowen's disease sample #1. Fuzzy C-means (FCM) images (**a**) were constructed with optimized parameters $\hat{K}_{opt} = 5$ and $\hat{m}_{opt} = 1.8$. These parameters were defined using the RBA-resulting curves (**b**) and Table 3. Assignment of the clusters: cluster 1 (epidermis); 2, 3 and 4 (dermis); 5 (Bowen's disease). The color bar represents the membership value for each pixel. In the corresponding hematoxylin and eosin (H&E)-stained section (top), Bowen's disease (outlined), epidermis (*) and dermis (+) are indicated. H&E-stained section (bottom) of higher quality localized at *circa* 25 μm from that analyzed by infrared (IR) imaging.


a
cluster 1
cluster 2
cluster 3
cluster 4
cluster 5
H&E
top
bottom
200μm
Membership value
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1

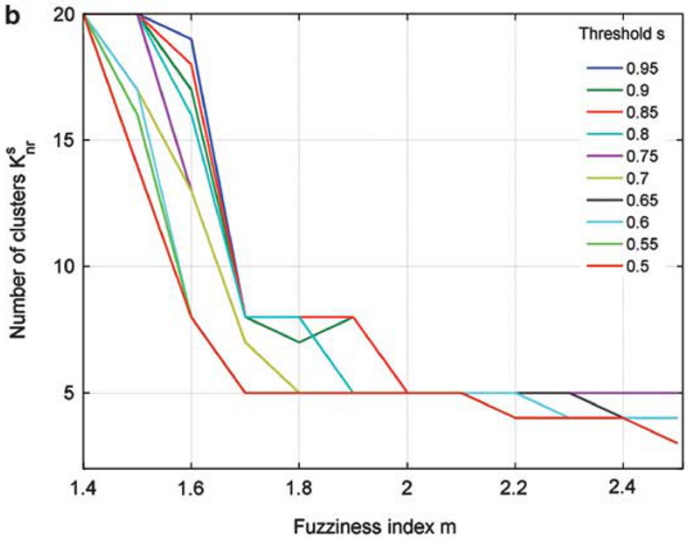

b
Number of clusters K_{nr}^{s}
Fuzziness index m
Threshold s
0.95
0.9
0.85
0.8
0.75
0.7
0.65
0.6
0.55
0.5
20
15
10
5
1.4
1.6
1.8
2
2.2
2.4

Table 2 Optimal parameters of FCM estimated by RBA in function of the threshold *s* for the human skin BCC sample #1

S	0.95	0.9	0.85	0.8	0.75	0.7	0.65	0.6	0.55	0.5
$\hat{K}^s_{opt}$	5	5	5	5	5	5	5	5	5	5
$\hat{m}^s_{opt}$	1.6	1.6	1.6	1.6	1.6	1.6	1.6	1.6	1.5	1.6

Optimal number of clusters $\hat{K}^s_{opt}$ and the corresponding optimal values of the fuzziness index $\hat{m}^s_{opt}$ have been determined for 10 different values of the threshold *s* from the curves shown in Figure 7b.

Table 3 Optimal parameters of FCM estimated by RBA in function of the threshold *s* for the Bowen's disease sample #1

S	0.95	0.9	0.85	0.8	0.75	0.7	0.65	0.6	0.55	0.5
$\hat{K}^s_{opt}$	8	5	8	8	5	5	5	5	5	5
$\hat{m}^s_{opt}$	1.7	2	1.7	1.7	1.8	1.8	1.7	1.7	1.7	1.7

Optimal number of clusters $\hat{K}^s_{opt}$ and the corresponding optimal values of the fuzziness index $\hat{m}^s_{opt}$ have been determined for 10 different values of the threshold *s* from the curves shown in Figure 8b.

membership value >0.9) and dermis associated to four clusters, which is in accordance to its biochemical marked heterogeneity (Supplementary Figures S16 and S17).

Conclusion

IR spectral microimaging associated with clustering techniques shows a great potential for the direct analysis of paraffin-embedded tissue sections of human skin cancers. Our results demonstrate that FCM clustering is more powerful than classical 'hard' clusterings (KM and hierarchical classification) to reveal biologically relevant information, related to the tumor heterogeneity and invasiveness. We have developed an original algorithm dedicated to the simultaneous determination of the optimal FCM parameters (number of clusters *K*, and fuzziness index *m*). This innovative data processing makes FT–IR microimaging a promising tool, integrable to gold standard histology. This could help in the guidance of the therapeutic strategy, especially for predictive extension of infiltrative cancer lesions.

Supplementary Information accompanies the paper on the Laboratory Investigation website (http://www.laboratoryinvestigation.org)

ACKNOWLEDGEMENTS

This study was supported by a grant of Institut National du Cancer (INCa), Canceropôle Grand Est. We would like to thank Ligue contre le Cancer, Comité de l'Aisne, INSERM PNR Imagerie and CNRS Projets Exploratoires Pluridisciplinaires for financial support. D.S. is a recipient of a doctoral fellowship from INCa, and E.L. from CNRS and Région Champagne-Ardenne.

DISCLOSURE/CONFLICT OF INTEREST

The authors declare no conflict of interest.

1. Krishna CM, Sockalingum GD, Bhat RA, *et al.* FTIR and Raman microspectroscopy of normal, benign, and malignant formalin-fixed ovarian tissues. Anal Bioanal Chem 2007;387:1649–1656.
2. Wolthuis R, Travo A, Nicolet C, *et al.* IR spectral imaging for histopathological characterization of xenografted human colon carcinomas. Anal Chem 2008;80:8461–8469.
3. Ly E, Piot O, Wolthuis R, *et al.* Combination of FTIR spectral imaging and chemometrics for tumour detection from paraffin-embedded biopsies. Analyst 2008;133:197–205.
4. Ly E, Piot O, Durlach A, *et al.* Differential diagnosis of cutaneous carcinomas by infrared spectral micro-imaging combined with pattern recognition. Analyst 2009;134:1208–1214.
5. Kendall C, Isabelle M, Bazant-Hegemark F, *et al.* Vibrational spectroscopy: a clinical tool for cancer diagnostics. Analyst 2009;134:1029–1045.
6. Kong R, Reddy RK, Bhargava R. Characterization of tumor progression in engineered tissue using infrared spectroscopic imaging. Analyst 2010;135:1569–1578.
7. Acerbo AS, Miller LM. Assessment of the chemical changes induced in human melanoma cells by boric acid treatment using infrared imaging. Analyst 2009;134:1669–1674.
8. Wang J, Chang KJ, Chen CY, *et al.* Evaluation of the diagnostic performance of infrared imaging of the breast: a preliminary study. Biomed Eng Online [serial on the Internet]. 2010; 9(3): Available from: http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd = Retrieve &db = PubMed&dopt = Citation&list_uids = 20055999.
9. Sobottka SB, Geiger KD, Salzer R, *et al.* Suitability of infrared spectroscopic imaging as an intraoperative tool in cerebral glioma surgery. Anal Bioanal Chem 2009;393:187–195.
10. Krafft C, Sobottka SB, Geiger KD, *et al.* Classification of malignant gliomas by infrared spectroscopic imaging and linear discriminant analysis. Anal Bioanal Chem 2007;387:1669–1677.
11. Bird B, Miljkovic M, Romeo MJ, *et al.* Infrared micro-spectral imaging: distinction of tissue types in axillary lymph node histology. BMC Clin Pathol 2008;8:8.
12. Bogomolny E, Huleihel M, Suproun Y, *et al.* Early spectral changes of cellular malignant transformation using Fourier transform infrared microspectroscopy. J Biomed Opt 2007;12:024003.
13. Fabian H, Thi NA, Eiden M, *et al.* Diagnosing benign and malignant lesions in breast tissue sections by using IR-microspectroscopy. Biochim Biophys Acta 2006;1758:874–882.
14. Diem M, Romeo M, Boydston-White S, *et al.* A decade of vibrational micro-spectroscopy of human cells and tissue (1994–2004). Analyst 2004;129:880–885.
15. Wood BR, Chiriboga L, Yee H, *et al.* Fourier transform infrared (FTIR) spectral mapping of the cervical transformation zone, and dysplastic squamous epithelium. Gynecol Oncol 2004;93:59–68.
16. Lasch P, Haensch W, Naumann D, *et al.* Imaging of colorectal adenocarcinoma using FT-IR microspectroscopy and cluster analysis. Biochim Biophys Acta 2004;1688:176–186.
17. Sebiskveradze D, Gobinet C, Ly E, *et al.* Effects of digital dewaxing methods on K-means-clusterized IR images collected on formalin-fixed paraffin-embedded samples of skin carcinoma. BioInfo BioEng 2008, BIBE 2008, 8th IEEE International Conference on 8–10 Oct. 2008; Athens 2008. p. 1–6.
18. Travo A, Piot O, Wolthuis R, *et al.* IR spectral imaging of secreted mucus: a promising new tool for the histopathological recognition of human colonic adenocarcinomas. Histopathology 2010;56:921–931.
19. Beljebbar A, Dukic S, Amharref N, *et al.* Monitoring of biochemical changes through the c6 gliomas progression and invasion by fourier transform infrared (FTIR) imaging. Anal Chem 2009;81: 9247–9256.
20. Beljebbar A, Amharref N, Leveques A, *et al.* Modeling and quantifying biochemical changes in C6 tumor gliomas by Fourier transform infrared imaging. Anal Chem 2008;80:8406–8415.
21. Ali K, Lu Y, Das U, *et al.* Biomolecular diagnosis of human glioblastoma multiforme using Synchrotron mid-infrared spectromicroscopy. Int J Mol Med 2010;26:11–16.
22. Conti C, Ferraris P, Garavaglia M, *et al.* Microimaging FTIR of head and neck tumors. IV. Microsc Res Tech 2009;72:67–75.
23. Untereiner V, Piot O, Diebold MD, *et al.* Optical diagnosis of peritoneal metastases by infrared microscopic imaging. Anal Bioanal Chem 2009;393:1619–1627.

24. Bezdek JC. Pattern Recognition with Fuzzy Objective Function Algorithms. Plenum Press: New York, 1981.
25. Mansfield JR, Sowa MG, Scarth GB, *et al.* Analysis of Spectroscopic Imaging Data by Fuzzy C-Means Clustering. Anal Chem 1997;69: 3370–3374.
26. Wang Y, Yao X, Parthasarathy R. Characterization of interfacial chemistry of adhesive/dentin bond using FTIR chemical imaging with univariate and multivariate data processing. J Biomed Mater Res A 2009;91:251–262.
27. Richter T, Steiner G, Abu-Id MH, *et al.* Identification of tumor tissue by FTIR spectroscopy in combination with positron emission tomography. Vibrational Spectrosc 2002;28:103–110.
28. Macqueen JB, (ed) Some methods for classification and analysis of multivariate observations. Proceedings of the fifth Berkeley symposium on mathematical statistics and probability. Statistical Laboratory of the University of California, Berkeley, 1967.
29. Jain AK, Dubes RC. Algorithms for clustering data. Upper Saddle River, NJ: Prentice-Hall, Inc, 1988.
30. Vijaya PA, Murty M, Subramanian DK. An Efficient Hybrid Hierarchical Agglomerative Clustering (HHAC) Technique for Partitioning Large Data Sets. Lecture Notes in Computer Science 2005;3776:583–588.
31. Wang X-Y, Garibaldi J, Bird B, *et al.* A novel fuzzy clustering algorithm for the analysis of axillary lymph node tissue sections. App Intell 2007;27:237–248.
32. Crowson AN. Basal cell carcinoma: biology, morphology and clinical implications. Mod Pathol 2006;19:S127–S147.
33. Rinker MH, Fenske NA, Scalf LA, *et al.* Histologic variants of squamous cell carcinoma of the skin. Cancer Control 2001;8:354–363.