Available online at www.sciencedirect.com

SCIENCE DIRECT®

Biochimica et Biophysica Acta 1688 (2004) 176–186

BIOCHIMICA ET BIOPHYSICA ACTA

BBA

www.bba-direct.com

Imaging of colorectal adenocarcinoma using FT-IR microspectroscopy and cluster analysis

Peter Lasch[a,*], Wolfgang Haensch[b], Dieter Naumann[c], Max Diem[a]

[a] *Department of Chemistry and Biochemistry, City University of New York, Hunter College, 695 Park Avenue, New York, NY 10021, USA*

[b] *Department of Pathology, Charité Medical Faculty of the Humboldt University, Robert-Rössle-Klinik at the Max-Delbrück-Centrum for Molecular Medicine, D-13125 Berlin, Germany*

[c] *Robert Koch-Institut, P13 "Biomedical Spectroscopy", Nordufer 20, D-13353 Berlin, Germany*

Received 26 September 2003; received in revised form 26 November 2003; accepted 3 December 2003

Abstract

In this paper, three different clustering algorithms were applied to assemble infrared (IR) spectral maps from IR microspectra of tissues. Using spectra from a colorectal adenocarcinoma section, we show how IR images can be assembled by agglomerative hierarchical (AH) clustering (Ward's technique), fuzzy C-means (FCM) clustering, and k-means (KM) clustering. We discuss practical problems of IR imaging on tissues such as the influence of spectral quality and data pretreatment on image quality. Furthermore, the applicability of cluster algorithms to the spatially resolved microspectroscopic data and the degree of correlation between distinct cluster images and histopathology are compared.

The use of any of the clustering algorithms dramatically increased the information content of the IR images, as compared to univariate methods of IR imaging (functional group mapping). Among the cluster imaging methods, AH clustering (Ward's algorithm) proved to be the best method in terms of tissue structure differentiation.

Keywords: Cluster analysis; IR imaging; Biomedical spectroscopy; FT-IR microspectroscopy; Colorectal adenocarcinoma; Pattern recognition; Tissue classification

1. Introduction

Fourier transform infrared (FT-IR) imaging is becoming an increasingly applied technique in biomedical spectroscopy [1–6]. This technique provides spatially resolved information on the basis of the chemical composition of the different structural compartments. In infrared (IR) microspectroscopy, the combination of IR spectroscopy and microscopy is exceptionally well suited for differentiating distinct tissue structures and for identifying tissue pathology. As the image contrast is based on the vibrational signature of the tissue components, FT-IR imaging does not require the use of dyes, tags or stains. Thus, compared with conventional histological techniques, IR imaging of tissues may simplify sample preparation procedures and minimizes sample modifications. Furthermore, since FT-IR microspectroscopy is a computer-based digital technique, the procedure of tissue evaluation can be automated. Based on validated databases of tissue spectra, this would permit to objectively diagnose pathological states of tissues. The answer to the crucial question, whether the IR technique for non-subjective tissue diagnosis will become a useful tool in the hand of the pathologist, is found in the quality and validation of these spectral databases. To this point, the compilation of spectral databases seems to be the main challenge biomedical IR microspectroscopy will be faced in the future [6].

A number of different techniques for IR spectral imaging of tissues were proposed in the past. Aside from the univariate technique of chemical mapping (also called functional group mapping), several authors employed mul-

Abbreviations: AH, agglomerative hierarchical (clustering); ANN, artificial neural network; FCM, fuzzy C-means (clustering); FT-IR, Fourier transform infrared; IR, infrared; H&E, hematoxylin–eosin; KM, k-means (clustering); MCT, mercury cadmium telluride; S/N, signal-to-noise ratio

* Corresponding author. Current address: Robert Koch-Institut, P13, Nordufer 20, D-13353 Berlin, Germany. Tel.: +49-30-4547-2405; fax: +49-30-4547-2606.

E-mail address: LaschP@rki.de (P. Lasch).

0925-4439/$ - see front matter © 2004 Elsevier B.V. All rights reserved.
doi:10.1016/j.bbadis.2003.12.006

tivariate pattern recognition methods such as principal component analysis [4], agglomerative hierarchical (AH) clustering [7–9], artificial neural network (ANN) analysis [4,6], k-means and fuzzy C-means (KM, FCM) clustering [10,11], and linear discriminant analysis (LDA) [12]. It turned out that most of the multivariate methods could be more or less successfully applied to problems in biomedical spectroscopy. Systematic comparative tests of these techniques are, however, still missing.

In this manuscript we describe the application of three different unsupervised multivariate imaging approaches (AH clustering, KMC, and FCM) to IR microspectral data while the supervised classification techniques (ANNs, LDA) will be compared in a separate study. Spectra for cluster analysis were acquired from a colorectal adenocarcinoma specimen and the results of the distinct cluster imaging techniques are compared in terms of image quality and consistency with standard histopathology. Furthermore, the applicability of cluster imaging to real problems in biomedicine is discussed.

2. Materials and methods

2.1. Sample preparation

Adenocarcinoma specimens were flash frozen in liquid nitrogen for later cryo-sectioning. The sample used in this study originated from a rectal adenocarcinoma. Its histopathological grade of malignancy was established as moderately differentiated (G2). An 8-μm-thick tissue slice was thaw mounted on a CaF_2 window. After data acquisition, this specimen was stained by hematoxylin–eosin (H&E) and reevaluated by a pathologist (W.H.).

2.2. FT-IR data collection

IR spectra were collected as absorbance spectra using an IRScope II IR microscope (Bruker Optics, Billerica, MA) equipped with a small sized MCT (HgCdTe) detector. The microscope was linked to a Vector 22 FT-IR spectrometer (Bruker) and was equipped with a software controlled *x*,*y*-stage. A specially designed microscope box was purged by dry air reducing spectral contributions from atmospheric water vapor and CO_2. FT-IR spectra were recorded using a 45×45-μm square aperture defined by knife-edges. Nominal spectral resolution was 8 cm^{-1}. Eight interferograms were averaged per pixel spectrum and apodized using a Happ–Genzel apodization function for Fourier transformation. Interferograms were zero-filled by a factor of 4. The background spectrum was recorded at a position of the IR-window outside the tissue sample.

91×91 spectra were collected at a step size of 20 μm in *x* and *y* direction for a total sample area of 1800×1800 μm^2 (i.e., spatial oversampling by a factor of 2 was employed). Data acquisition was carried out by means of the OS/2 based OPUS software package supplied by Bruker.

2.3. Data processing

All data processing and image assembly was performed by a program, written by one of us (P.L.), that is now available from CytoSpec (http://www.cytospec.com). This program differs from software products available from instrument manufacturers in that it was designed and written to operate on entire spectral (imaging) data sets, rather than individual spectra.

2.4. Data pre-processing

In order to permit a meaningful comparison of the clustering methods to be analyzed, uniformly pretreated data were used. The 8281 spectra were subjected to a quality test using settings of 50 and 250 for the minimal and maximal integral intensity criterion in the wavelength range of 950–1750 cm^{-1}. Spectra were also evaluated by a signal-to-noise ratio (S/N) test using a criterion of 500 as a threshold (signal: maximum of the amide I band; noise: the standard deviation in the spectral range 1800–1900 cm^{-1}). All spectra passing the tests were subsequently converted to first derivatives using a nine-smoothing point Savitzky–Golay algorithm. Derivative spectra were scaled before the cluster analysis such that the sum squared deviation over the indicated wavelengths (950–1750 cm^{-1}) equals unity (vector normalization):

$$\sum_k (A_k - \bar{A})^2 = 1 \quad (1)$$

where A_k is the absorbance at wavelength k and $\bar{A}$ is the average absorbance value in the corresponding spectral region.

Spectra with a negative quality test (19 out of 8281, ca. 0.23%) were routinely excluded from further data analysis and appeared in the IR images as black areas. Adequate data pretreatment and a high spectral quality turned out to be an essential prerequisite for further multivariate data analysis.

3. Results and discussion

We start the discussion with a short review of the histopathological features of colon tissue. Subsequently, the clustering methods will be introduced, and the images obtained by applying these clustering methods to the data set will be compared to the photomicrographs of the stained tissue.

3.1. Histopathological architecture of colorectal adenocarcinomas

Carcinoma cells of colorectal adenocarcinomas are malignant transformed subsets of the colonocytes. The vast majority of these tumors originate from resorptive epithelial cells, mainly located at the mucosal surface and in the upper

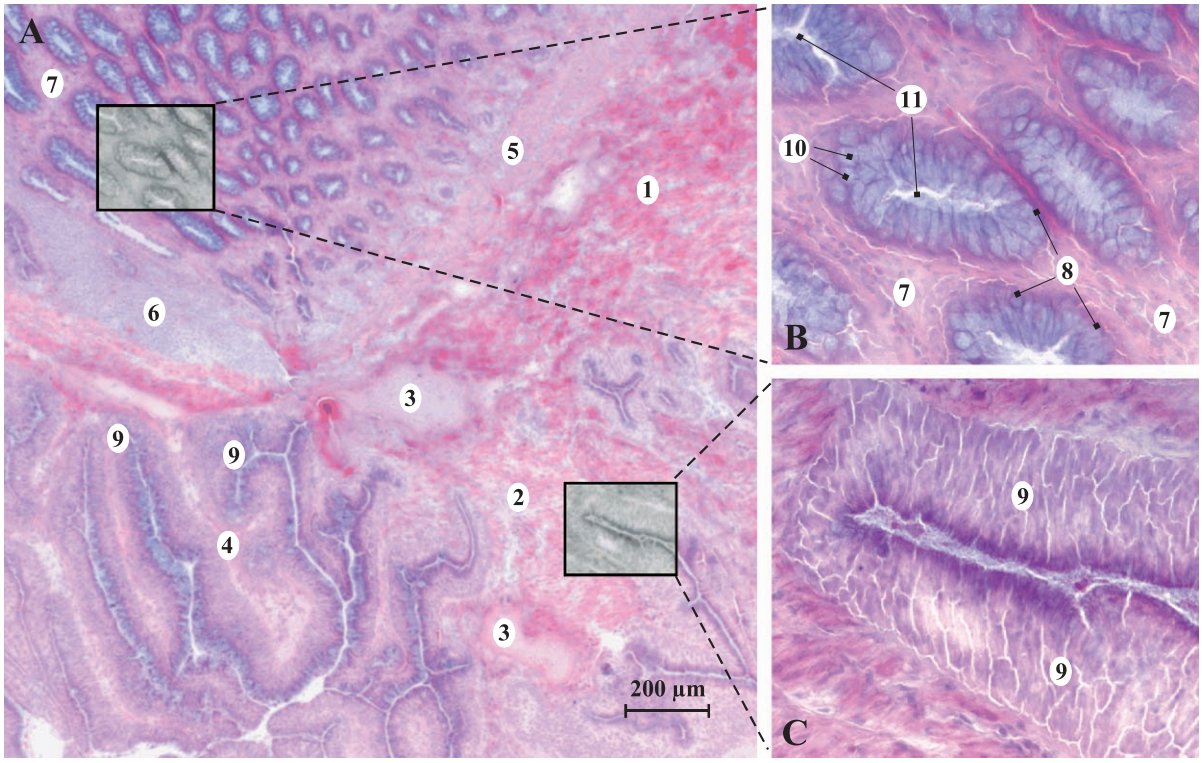
third of tubular glands, also called crypts. In the colorectal mucosa the epithelial cells form a layer of columnar epithelium with well-preserved architecture. In the crypts these cells are arranged together with other epithelial cell types (goblet cells, stem cells, various types of functionally differentiated endocrine cells). The lamina propria mucosae, which is poor in lymphangia, crypts and the muscularis mucosae constitute the mucous membrane of the colorectum. Noncancerous transitional mucosa is depicted in Fig. 1A (upper left), and at higher magnification in Fig. 1B. Fig. 1B shows the crypts cross-sectioned with the epithelial cells (colonocytes), goblet cells, and the lamina propria mucosae (cf. numbering of the inset of Fig. 1).

Colorectal adenocarcinomas originate from the epithelial cells and are able to infiltrate the subjacent layers (submucosa, tunica muscularis) of colon and rectum. This is reflected in the lower left and in the central right of the photomicrograph of Fig. 1A. As illustrated in Fig. 1A and C, the adenocarcinoma exhibits typical morphological signs of malignancy: atypical histo-architecture such as multiple layers of cells exhibiting pleomorphism, and infiltration of the submucosa. Fig. 1A displays furthermore regions with lymphoid Peyer's patches. The example of Fig. 1 clearly demonstrates the complexity of the tissue histo-architecture that faces IR spectral imaging. Tissue specimens may contain normal tissue components, tissues exhibiting non-neoplastic alterations (e.g. inflammation), and also neoplastic tissue components.

3.2. Basic principles of IR cluster imaging

In order to partition IR tissue spectra into classes (or clusters), and to assemble and compare "cluster images", we applied three different clustering techniques to the data set of IR microspectra. These spectra were acquired from the tissue specimen shown in Fig. 1. The data pretreatment was identical for all three instances.

Clusters should contain IR spectra from histological regions that exhibit similar spectral features. Spectra in different clusters ideally exhibit different spectral signatures, i.e. the inter-cluster variance should be maximal and the intra-cluster variance minimal. Image assembly on the basis



1- submucosa
2- submucosa and desmoplastic connective tissue, partly infiltrated by the carcinoma
3- blood vessel
4- fibrovascular stock and other connective tissues
5- lamina muscularis mucosae
6- Peyer's patch
7- lamina propria mucosae
8- colonocytes
9- adenocarcinoma (cancerous epithelium)
10- goblet cells
11- central lumen of the crypts

Fig. 1. (A) Photomicrograph of an H&E-stained section of a colorectal adenocarcinoma at a mid-power field magnification. The area shown in panel A (1800 × 1800 µm) was mapped before staining in the IR microspectroscopic measurements using a step size of 20 µm, and an aperture of 45 µm for a total of 8281 individual IR spectra. Panels B and C display regions with benign (B) and malign (C) anatomical features at a higher magnification (see text for further discussion).

of cluster analysis follows the simple idea of assigning a distinct color to all spectra in one cluster. Since each spectrum of a mapping experiment has a unique spatial *x*,*y* position in the map, false-color images can be generated by plotting specifically colored pixels as a function of the spatial coordinates.

3.3. Results from KM clustering

KM clustering is a nonhierarchical clustering method, which obtains a "hard" (crisp) class membership for each spectrum, i.e. the class membership of an individual spectrum can take only the values 0 or 1. It uses an iterative algorithm to update randomly selected initial cluster centers, and to obtain the class membership for each spectrum, assuming well-defined boundaries between the clusters [13,14]. In the present study we used a setting of 100 for the maximal number of iterations and up to 11 for the number of cluster.

The iterative algorithm of KM clustering (k-means algorithm of MacQueen) can be described as follows: IR spectra are illustrated as points in a *p*-dimensional space (*p* is the number of features or data points of the spectra). In this space a number of *k* points is initially chosen, where each point denotes the origin of a future cluster. Then, distance values between the points and all objects (spectra) are calculated. Objects are assigned to a cluster on the basis of a minimal distance value. Next, centroids of the clusters are calculated and distance values between the centroids and each of the objects are recalculated. If the closest centroid is not associated with the cluster to which the object currently belongs, the object will switch its cluster membership to the cluster with the closest centroid. The centroid's positions are recalculated every time a component has changed its cluster membership. This process continues until none of the objects has been reassigned [13].

Fig. 2 depicts the images assembled by KM clustering. For comparison with histopathology, this figure displays also a photomicrograph of the H&E-stained tissue (panel A of Fig. 2). Panels 2B–2F display the KM clustering images, which were assembled by using varying settings for the number of clusters (2, 4, 6, 8, and 11, respectively). At a first glance, the principal correspondence between histopathology and KM cluster imaging is obvious.

In panel D, for example, an IR image is displayed which was assembled using 6 clusters. All of the spectral clusters can be clearly assigned to histological structures: dark blue pixels encode the submucosa and blue pixels determine the central parts of the crypts or of the adenocarcinoma, respectively. Brown-colored areas of Fig. 2D are typical

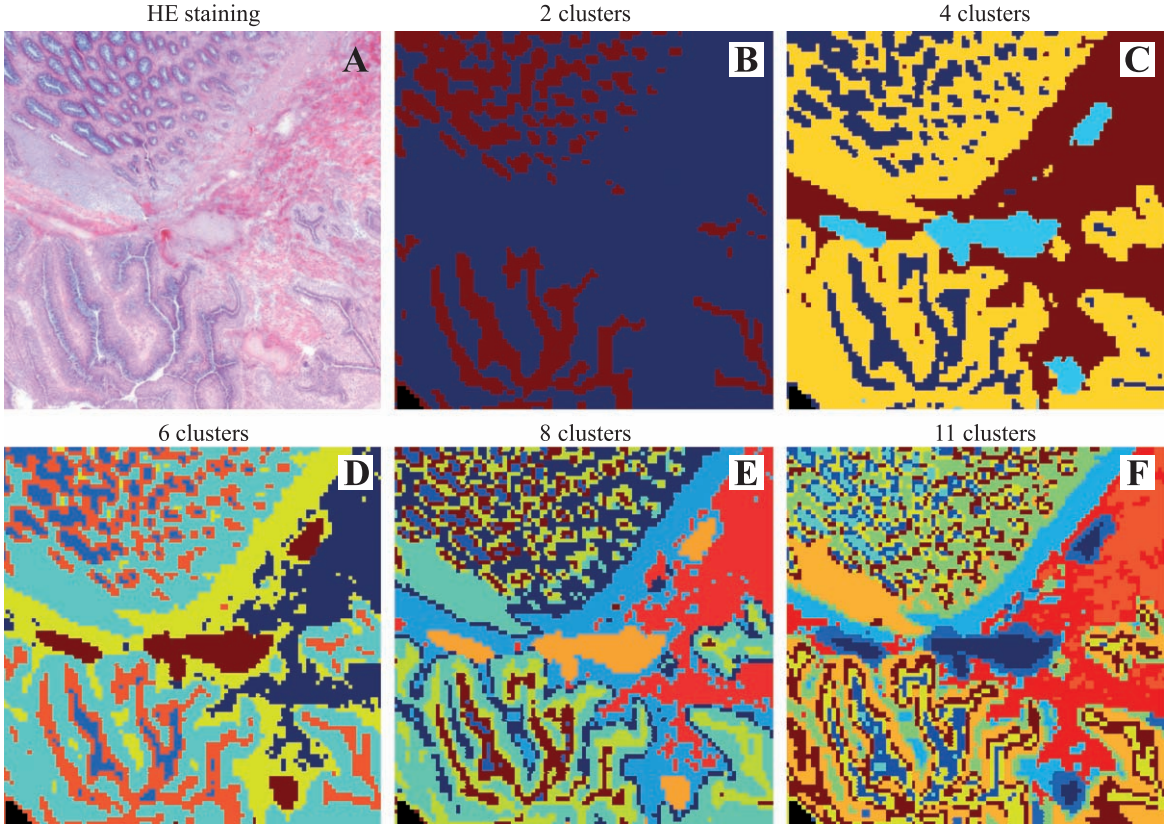


Fig. 2. IR imaging of colorectal adenocarcinoma specimen by KM clustering. Panel A: H&E-stained specimen at a mid-power field magnification (see also Fig. 1A). Panels B–F: IR images, reassembled by KM clustering. Different clusters are encoded by different colors. Note that in panels B–F only the initial parameter of number of cluster was varied (from 2 to 11, see inset).

for the lumen of blood vessels (cf. Fig. 1A). Furthermore, KM cluster imaging permits identification of tissue structures such as propria and fibrovascular stock (light green), and structures formed by smooth muscles such as lamina muscularis mucosae, the wall of blood vessels, and, possibly, remnants of the tunica muscularis (yellow). Finally, the outer cell layers of the crypts and of the adenocarcinoma appear as red regions in the pseudo-color map. It should be noted that we could not discriminate additional histological structures by KM clustering. Even if the number of clusters was increased to 11 (see panel F) or higher (not shown), further differentiation of histological structures was not successful. Particularly, colonocytes from normal and transitional mucosa are indistinguishable from malign cells of the cancerous epithelium (see, e.g. Fig. 2F).

3.4. Results from FCM clustering

FCM clustering is also a nonhierarchical clustering method. This clustering technique partitions objects into groups whose members show a certain degree of similarity. Unlike KM clustering, the output of FCM clustering is a membership function, which defines the degree of membership of a given spectrum to the clusters. The values of the membership function can vary between 1 (highest degree of cluster membership) and 0 (no class membership), where the sum of the C cluster membership values for one object equals 1.

Thus, this method departs from the classical two-valued (0 or 1) logic, and uses "soft" linguistic system variables and a continuous range of truth values in the interval [0,1]. FCM cluster imaging uses a fuzzy iterative algorithm to calculate the class membership grade for each spectrum. The iterations in FCM clustering are based on minimizing an objective function, which represents the distance from any given data point (spectrum) to the actual cluster center weighted by that data points membership grade [15]. In the present study we used a setting of 0.0001 for the minimal amount of improvements (the stopping criterion of the iteration) and up to 11 for the number of classes.

The advantage of the FCM over KM clustering is that both outliers and data, which display properties of more than one class, can be characterized by assigning nonzero class membership values to several clusters. In the IR tissue maps assembled by FCM clustering, the membership values can be encoded by color intensities. A high-class membership value defines high color intensity and vice versa. FCM cluster images can be then reassembled by plotting the color

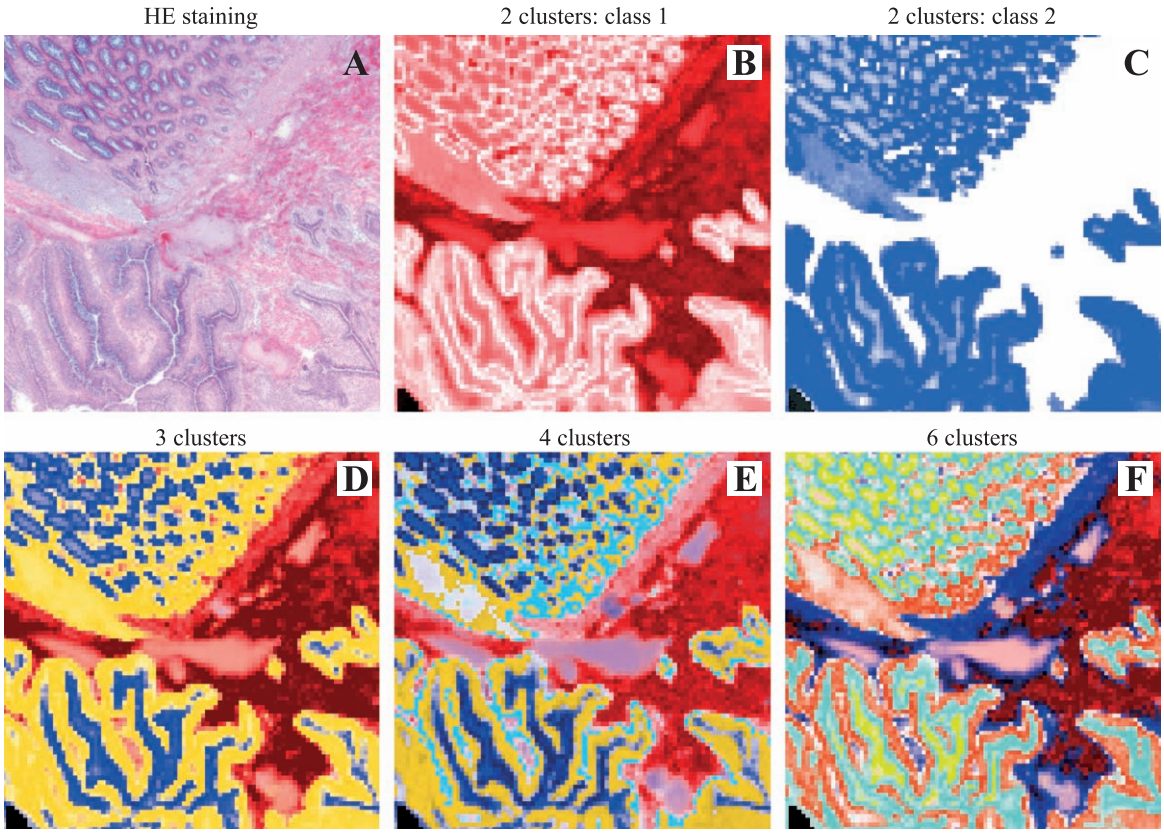


Fig. 3. IR imaging of colorectal adenocarcinoma specimen by FCM clustering. Panel A: H&E-stained specimen at a mid-power field magnification (see also Fig. 1A). Panels B–C: IR images of a two-class classification trial, showing that the sum of class memberships of an individual spectrum equals 1. Panels D–F: The class membership grades of individual IR spectra were converted to color scale levels and plotted for each of the three (D), four (E), or six (F) classes as a function of the spatial coordinates.

scales as the function of the spatial coordinates. In Fig. 3 the results of FCM cluster imaging are illustrated. While Fig. 3A again displays the H&E tissue specimen of Fig. 1, the red (panel B) and blue (panel C) colored images show the scaled spatial distribution of specific IR spectroscopic patterns. In this two-class classification trial, the complementary nature of the red and blue clusters is clearly discernible. The interpretation of these IR spectral maps is, however, not straightforward. Presumably, the most intense red regions in Fig. 3B may show submucosa and smooth muscle tissues, which are known to contain relatively high amounts of collagen (submucosa>smooth muscle). Since collagen exhibits highly specific IR bands [16], collagen-rich tissue structures can be spectroscopically easily differentiated from other tissues of the colorectum. The most intense blue regions of panel C show normal or abnormal epithelial cells.

Panels D–F of Fig. 3 display maps obtained for 3, 4 and 6 clusters, respectively. These images have been generated by superimposing 3, 4, or 6 single color FCM images. We found that each individual color in these examples can be assigned to specific tissue structures. On the other hand, if the number of clusters was further increased (>6), the interpretation of the FCM maps became more and more ambiguous in terms of the known histopathology. To give an example, epithelial cells originating from normal or transitional mucosa and malign cells of the cancerous epithelium could not be discriminated by FCM clustering. Even if the number of clusters was increased to 11, the method of FCM failed to differentiate spectra from these structures (data not shown).

3.5. Results from AH clustering

We also tested the method of AH clustering, a technique which has been successfully applied to many other problems in biomedical IR spectroscopy [8,17,18]. Like KM clustering, AH clustering is also a "hard" clustering method, i.e. spectra may belong to a given cluster or not. The algorithm of this technique can be illustrated in the following way: First, a distance matrix between all spectra is calculated. This matrix contains the complete set of interspectral distances (measures of dissimilarity). The distance matrix is symmetric along its diagonal and has the dimension $n \times n$, where n is the number of objects (spectra). In the AH clustering algorithm, two objects that are closest to one another (most similar) are then merged into a new object (cluster). In this way, the dimension of the distance matrix is reduced to $(n-1) \times (n-1)$. Subsequently, the distances of the new formed object to all remaining objects are reevaluated, and the next two most similar objects are merged. This process is repeated until all objects are combined into one single cluster. The merging process can be visualized in a tree-like diagram, which is called a "dendrogram". The final partition of objects into classes is defined by "cutting" the dendrogram. While the clustering process is completely unsupervised, this step is subjective and defines the number of classes, which will appear in the cluster image. In practice, the number of classes is rarely clearly defined and may require additional information.

A number of different methods are known for calculating the initial interspectral distances, and for obtaining the new distances after merging two objects. A detailed description of these methods can be found elsewhere [19,20]. The best imaging results, as judged by good correlation between histology and spectroscopy, were achieved if a combination of D-values for obtaining the distances measures and Ward's algorithm for clustering was applied.

$$D-\text{values}: \; d_{y1y2} = (1 - r_{y1y2}) \times 1000 \tag{2}$$

$r_{y1,y2}$ is known as the Pearson's correlation coefficient:

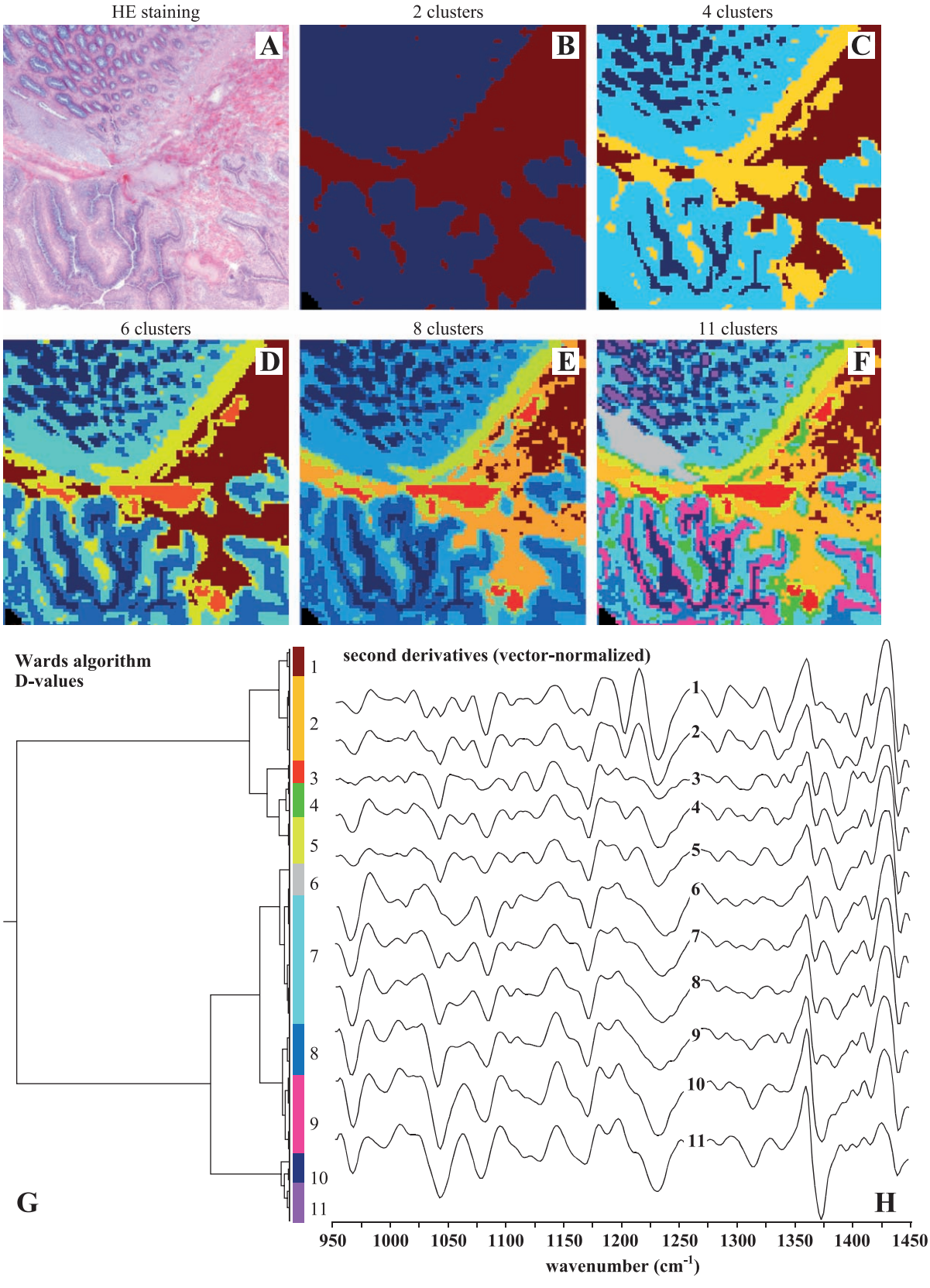
$$r_{y1y2} = \frac{\left(\sum_{i=1}^{p} y_{1i} \cdot y_{2i}\right) - p \cdot y_1 \cdot y_2}{\sqrt{\left(\sum_{i=1}^{p} y_{1i}^2 - p \cdot y_1^2\right) \cdot \left(\sum_{i=1}^{p} y_{2i}^2 - p \cdot y_2^2\right)}}, \tag{3}$$

where p is the total number of absorbance values in the spectra, and y_{1i} and y_{2i} are the i-th absorbance values of spectrum 1 or 2, respectively. Ward's algorithm [21] was used since it tends to produce dense clusters.

3.6. Correlation of spectral and histopathological images

The results of the AH cluster imaging approach are displayed in Fig. 4. Again, panel A shows the photomicrograph of the H&E-stained specimen, and panels B–F display the respective AH cluster images. Like in KM or FCM clustering (cf. Figs. 2 and 3), the cluster images were assembled by varying only one single parameter, the number of clusters. As for KM and FCM cluster analysis, the results of IR microspectroscopy and light microscopy are highly correlated. For example, Fig. 4D (6 clusters) clearly shows that each spectral cluster can be assigned to a unique pathohistological structure: red regions encode blood vessels, yellow-colored areas can be assigned to smooth muscle tissue (muscularis mucosae, wall of blood vessels, and partly fibrovascular stock). Furthermore, turquoise regions encode the structures of propria and fibrovascular stock, while crypts and adenocarcinoma are colored blue (dark and light blue). The submucosa appears brown in the present example. Generally, the images depicted in Fig. 4C–E (assembled by defining 4, 6, or 8 clusters) are very similar to the respective IR images reassembled by KM and FCM cluster imaging.

If the number of clusters is increased to 11 (see panel F of Fig. 4), the AH cluster imaging methodology uncovers the typical morphology of the mucosa and also of the carcinoma. As shown by the dendrogram in Fig. 4G, hierarchical clustering reveals two major clusters, with one cluster


HE staining
2 clusters
4 clusters
A
B
C
6 clusters
8 clusters
11 clusters
D
E
F
Wards algorithm
D-values
1
2
3
4
5
6
7
8
9
10
11
G
second derivatives (vector-normalized)
1
2
3
4
5
6
7
8
9
10
11
H
950
1000
1050
1100
1150
1200
1250
1300
1350
1400
1450
wavenumber (cm-1)

composed of mesenchymal tissues (submucosa, blood vessels) and the lamina muscularis mucosae (see clusters 1–5) and the other cluster of mucosal tissues (except lamina muscularis mucosae, see clusters 6–11). Within the cluster image, crypts of the transitional mucosa can be easily identified by their dark blue epithelia surrounding magenta colored central regions (see upper left of Fig. 4F). Also, the crypt-like architecture of the malignant glands can be recognized as pink rings of epithelial cells encircling dark blue central regions. Pale-blue and turquoise areas reflect properties of the lamina propria mucosae (upper left), but can be also found around the tumor to the lower left. From the point of view of histology, both pale blue and turquoise areas differ fundamentally. It should be noted that at present we do not have a simple answer explaining this finding. Yellow regions in Fig. 4F can be clearly assigned to smooth muscles structures (muscularis mucosae and wall of blood vessels). At the mesenchymal site (submucosa), which is indicated as structure 1 in Fig. 1A, we distinguish between ocher and brown zones. The color shift from brown to ocher correlates well with histopathology: while brown areas depict the original submucosal connective tissue, this tissue is replaced at the tumor site by ocher colored dense desmoplastic connective tissue.

Interestingly, the central parts of the malignant glands of Fig. 4F are encoded by the same color as the colonocytes of the mucous membrane. Thus, glandular structures of the adenocarcinoma are composed of an outer layer of carcinoma cells (which undoubtedly differ spectroscopically from colonocytes) and carcinoma cells of the central parts exhibiting similar spectral features to colonocytes. In the following, we will discuss the spectral changes on which these classification results are based.

3.7. Spectral changes

The software implementation of hierarchical cluster analysis includes an "average spectra" option, i.e. average spectra of spectral clusters can be easily obtained and stored for further analysis. In panel H of Fig. 4, we show average spectra (vector normalized second derivative spectra) as obtained from the 11 class classification trial of Fig. 4F. It should be noted that each spectrum of Fig. 4H was obtained by averaging hundreds of individual spectra. Thus, the spectra obtained have an exceptionally high signal-to-noise ratio (>8.000 in the amide I region). Even tiny spectral features are not fortuitous but represent real IR spectral properties of a certain tissue structure.

The most striking spectral differences can be found between spectra of class 1 (submucosa, brown) and 11 (center of crypts, magenta). In the discussion above, we pointed out that the submucosa is a collagen-rich tissue, exhibiting typical collagen bands at 1202, 1282, 1336 (amide III) [16], and 1453 cm^{-1}. Interestingly, the spectra obtained from the central parts of the crypts exhibit highly characteristic IR bands as well. These bands are different from those of collagen and can be found at 1044, 1079, 1121, 1314, and 1373 cm^{-1} (see traces 1 and 4 of Fig. 5A). Most of these bands can be assigned to mucine, a glycoprotein rich in cysteine. Mucine is known to be present

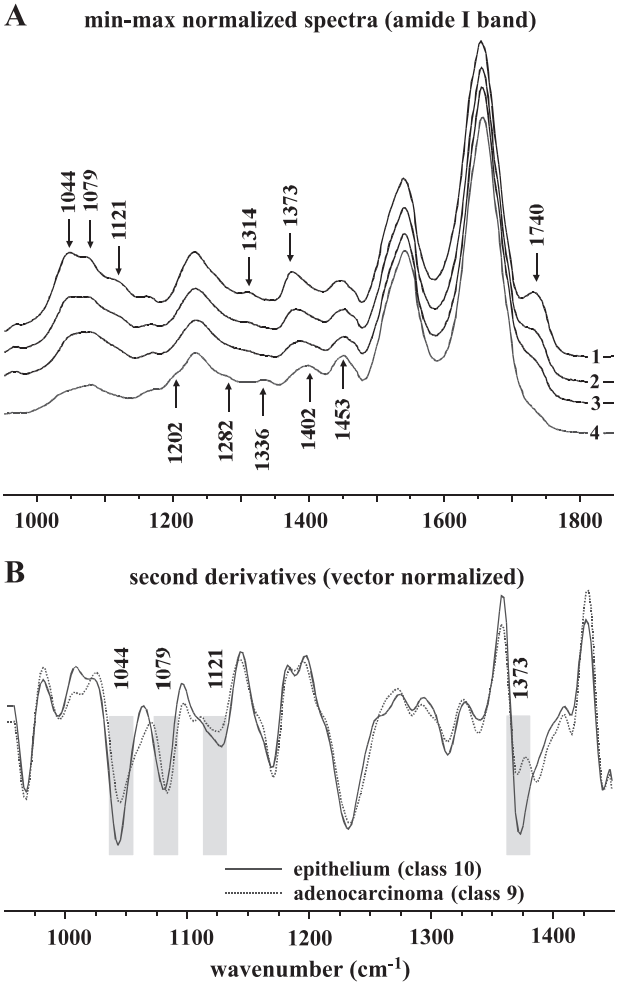


Fig. 5. IR microspectra obtained from a colorectal adenocarcinoma. Panel A: Representative average spectra of the spectral clusters as obtained by AH clustering (11 clusters, cf. Fig. 4F–H). 1—center of the crypts (cluster 11 in Fig. 4F–H), 2—epithelial cells (crypts, cluster 10), 3—malign epithelium (cluster 9), 4—submucosa (cluster 1). Spectra were min–max normalized in the amide I region. Panel B: Vector normalized second derivatives of curves 2 (epithelial cells) and 3 (malign epithelium) of Fig. 5A (for details see text).

Fig. 4. IR imaging by AH clustering (*D*-values, Wards algorithm). In this example of cluster imaging, spectra of a specific cluster are encoded by a specific color. To reassemble the IR images, the assigned color for each spectral cluster is displayed at the coordinates at which each pixel spectrum was collected. IR images are reassembled by color encoding 2, 4, 6, 8, or 11 classes (see panels B–F). In the 11-class classification trial, the malign epithelial cells of the adenocarcinoma (red areas of panel F) can be clearly separated from the respective epithelial cells of non-neoplastic crypts (blue and dark blue regions of panel F, see also Fig. 1). Panel G displays the dendrogram with the respective color assignments of the 11-class classification trial. The class-mean spectra (second derivatives) are given in panel H (numbering corresponds to the class numbering of the dendrogram).

either as a precursor in secretory granules of mature goblet cells or, after its secretion, in the lumen of the crypts (cf. Fig. 1B). Due to the heterogeneity of the distinct types of mucine, the band assignment differs slightly in literature. To give an example, the position of the most prominent mucine peak was found between 1035 and 1050 cm^{-1} [6,22,23]. In the light of the spectral differences between traces 1 (mucine rich regions of the crypts) and 4 (submucosa), the differences between spectra originating from the epithelial cell layer of the crypts (trace 2 of Fig. 5A) and from epithelial cells of the carcinoma (trace 3) are rather small. However, for a better visualization these subtle differences were "amplified" by calculating second derivatives (cf. traces 2 and 3 in Fig. 5B).

A closer inspection of the derivative spectra shown in Fig. 5B revealed reproducible spectral differences at 1044, 1079, 1121 and 1373 cm^{-1}. From the discussion above, it is apparent that most of these changes are characteristic for mucine (see also traces 9–11 of Fig. 4H). In the following, we will illustrate that this surprising observation coincides with basic principles known for a long time in histopathology and also with our findings in earlier studies.

In these studies, we showed that cancerous and noncancerous cultured immortal cells exhibited similar IR spectra, whereas terminally differentiated and metabolically inactive cells display a significantly different spectral pattern [24,25]. The most impressive spectral changes were found for the PO_2^- bands at 1088 (sym) and 1234 cm^{-1} (antisym). It was postulated that the intensity of the PO_2^- bands could be correlated with the grade of divisional activity. Since in cancer cells the division rate is usually higher than in precursor cells, the average spectrum from cancer tissue should differ from that of "normal" (benign) tissues with a lower mitotic index.

The experimental findings of the present study, however, seem to be contradictory to these conclusions. Obviously, instead of signs for a higher divisional activity of the carcinoma, we found only a mucin-like difference pattern. To answer this question, we need to discuss histology and pathohistology of the intact mucosa and of colorectal adenocarcinomas.

The mucosal epithelium of the gastrointestinal tract is a thin layer of epithelial cells (colonocytes), which undergoes a continual turnover and renewal. It is continuously replaced by outward displacement from the crypts with typical cell lifetime of about 6 days. In the descending colon, stem cells of resorptive epithelial cells are located at the crypt's base and cellular migration occurs in an outward direction. Thus, epithelial cells of the mucosa are cells with a very high proliferation rate. The highest mitotic activity can be found at the base of the crypts.

In colorectal adenocarcinomas, which originate from resorptive epithelial cells, the cell dynamic described above is significantly altered. Depending on the histological grading, the proliferative capability, age and renewing rate of carcinoma cells are increased. Histologically, the variability of cell size and shape is higher and the epithelium becomes multilayered. From a spectroscopic point of view, it is also important to note that these changes are accompanied by an overall loss of other cell types, such as secretive goblet cells, as the adenocarcinoma represents a selective clonal proliferation of the malign cell type.

The latter facts could explain the unexpected spectral differences observed between the epithelial layer of the crypts and epithelial cells of the adenocarcinoma. First and foremost, while the aperture size of 45 μm allowed the collection of "pure" spectra from the adenocarcinoma, spectra from the epithelial cell layer should exhibit signs from adjacent histological structures, since the thickness of the epithelial cell layer is about 15 μm, i.e. much smaller than the aperture of 45 μm used in the experiments. As a consequence, spectral "impurities" from adjacent tissue structures, e.g. goblet cells filled with mucine, are highly possible. On the contrary, in adenocarcinomas the layers of epithelial cells are usually thicker (cf. Fig. 1B and C) and contain fewer amounts of goblet cells. Thus, in order to resolve "pure" spectra from the epithelium of normal crypts a spatial resolution of less than 15 μm is required. Furthermore, it is possible that mucine-like spectral differences strongly superimpose existing variances in the PO_2^- region. These divisional activity-related spectral alterations are expected to be rather small as the colonocytes itself are cells with a very high mitotic rate. Additionally, it is likely that in moderately differentiated carcinomas (G2 grading) the physiologically high levels of cellular proliferation and growth are only slightly increased.

3.8. Computational considerations

As discussed above, we find that the best correlation between histopathology and spectral images was observed if the data were processed by hierarchical cluster analysis. This method is an unsupervised computational method in the sense that neither reference data nor any starting conditions like presumed cluster centers are required. The endpoint of the hierarchical cluster analysis is somewhat arbitrary, in that, subsequent to the computations, the number of clusters is selected that gives the best discrimination. In the CytoSpec implementation of hierarchical cluster analysis, this is achieved by terminating the calculations at the level of two clusters. The membership list of the clustering process (up to a preset value of 50 clusters) is maintained. Thus, it is possible to check the images created prior to the last merging processes and to determine empirically at which level the discrimination between the various tissue types and disease state is lost in the final clustering process. This step, in principle, could be rendered completely unsupervised as well if a criterion, such as the density or homogeneity of the cluster, or the merging distance between the clusters is used as an indicator for the endpoint of the calculations.

Table 1
Comparison of the total CPU times taken for KMC, FCM, and AH cluster imaging

Cluster analysis	Calculation time	CPU time dependence
k-means clustering	7 min	$t_{KM} \sim const_1 \times k \times n$
fuzzy C-means clustering	30 min	$t_{FCM} \sim const_2 \times k \times n$
hierarchical clustering	4.5 h	$t_{AH} \sim const_3 \times n^2$

Calculations were carried out on a 966-MHz Intel Pentium IV workstation equipped with 512-MB RAM, running under the Windows 2000 operating system. 8281 vector-normalized first derivative spectra in the spectral region of 950–1750 cm^{-1} were clustered, and images of 11 clusters were reassembled. The CPU times specified are measured from the initialization of the distance matrix calculation to the display of the final cluster images. The third column indicates a first-order approximation of the processing time dependence on the number of objects (n) and the number of classes (k) (see Discussion for details).

The best correlation between histopathology and spectral images was found if the data were processed by hierarchical clustering, using a combination of the so-called *D*-values as the distance measure and Ward's algorithm for clustering. In our opinion, the high grade of correspondence seems to be the major advantage of hierarchical clustering over the methods of KMC and also over FCM clustering. Furthermore, the results of hierarchical clustering are always independent from starting conditions as no random initialization of k (KM clustering) or C (FCM clustering) starting points is required.

The major drawback of AH clustering is the high computational requirements, which becomes more and more important when large data sets have to be analyzed. From the algorithm of hierarchical clustering, it is apparent that the CPU time scales with the square of the number of objects (spectra, n). At large n (n>1000), this dependence is mainly due to the task of searching for the global minimum in the distance matrix. Furthermore, the allocation of computer memory (RAM) by the distance matrix is sizable. To give an example, the size of the distance matrix also scales with n^2 and was in the current study 274.3 MB ($n \times (n-1)/2$ distance values, stored as 8 byte double precision values; $n = 8281$). Since 32-bit operating systems such as Windows or Linux can normally handle 2 GB of address space for applications, the theoretical limit for AH cluster imaging on PCs can be easily calculated (23 170 spectra, ca. 150×150 spectra).

For KM and FCM clustering, the computational effort scales in a first-order approximation linearly with n. Thus, compared to KM and FCM clustering, the AH algorithm is significantly more time-consuming at large n. For the present study, the calculation time for AH clustering was 4.5 h compared to 30 min for FCM, and only 7 min for KM clustering (see Table 1 for details). In a practical environment, this is unacceptable. Although, further developments of computer hardware will considerably increase CPU speed, AH cluster analysis will be also in the future not the appropriate technique for routine analysis of IR imaging data (e.g. recorded by large focal plane array detectors). This task can be taken much more efficiently by supervised classifiers, such as ANNs, trained and validated on the basis of large databases of IR tissue microspectra. For validation and compilation of these databases, however, cluster analysis is an indispensable explorative data analysis technique.

4. Conclusions

Spatially resolved IR microspectroscopy in combination with digital imaging techniques is a powerful new technique which can be used to assemble false color images from histological specimens. In the present study, we applied three different clustering techniques to microspectroscopic data from colorectal adenocarcinoma sections: KM, FCM, and hierarchical clustering. The use of any of the clustering algorithms dramatically increased the information content of the IR images, as compared to the functional group mapping technique. In terms of tissue structure differentiation, AH clustering (Pearson's product momentum correlation coefficient, Ward's algorithm) proved to be the best, but also the most CPU intensive image methodology.

Acknowledgements

We are grateful to Anthony Pacifico (Hunter College) and Luis Chiriboga (Pathology Department, Bellevue Hospital, New York University) for fruitful discussions and support. This work was partially supported by grants from the National Institutes of Health (CA 81675 and GM 60654, to MD) and RR-03037 (supporting the infrastructure of the Chemistry Department at Hunter College).

References

[1] J. Guilment, S. Markel, W. Windig, Infrared chemical micro-imaging assisted by interactive self-modeling multivariate analysis, Appl. Spectrosc. 48 (1994) 320–326.
[2] L.P. Choo, D.L. Wetzel, W.C. Halliday, M. Jackson, S.M. LeVine, H.H. Mantsch, In situ characterization of β-amyloid in Alzheimer's diseased tissue by synchrotron FTIR microspectroscopy, Biophys. J. 71 (1996) 1672–1679.
[3] L.H. Kidder, V.F. Kalasinsky, J.L. Luke, I.W. Levin, E.N. Lewis, Visualization of silicon gel in human breast tissue using new infrared imaging spectroscopy, Nat. Med. 3 (1997) 235–237.
[4] P. Lasch, D. Naumann, FT-IR microspectroscopic imaging of human carcinoma thin sections based on pattern recognition techniques, Cell. Mol. Biol. 44/1 (1998) 189–202.
[5] L. Chiriboga, P. Xie, H. Yee, D. Zarou, W. Zakim, M. Diem, Infrared spectroscopy of human tissue: IV. Detection of dysplastic and neoplastic changes of human cervical tissue via microscopy, Cell. Mol. Biol. 44/1 (1998) 219–229.
[6] P. Lasch, W. Haensch, E.N. Lewis, L.H. Kidder, D. Naumann, Characterization of colorectal adenocarcinoma sections by spatially resolved FT-IR microspectroscopy, Appl. Spectrosc. 48 (2002) 1–10.
[7] C.P. Schultz, H.H. Mantsch, Biochemical imaging and 2D classification of keratin pearl structures in oral squamous cell carcinom, Cell. Mol. Biol. 44/1 (1998) 203–210.
[8] M. Jackson, B. Ramjiawan, M. Hewko, H.H. Mantsch, Infrared mi-

croscopic functional group mapping and spectral clustering analysis of hypercholesteramic rabbit liver, Cell. Mol. Biol. 44/1 (1998) 89–98.
[9] M. Diem, L. Chiriboga, H. Yee, Infrared spectroscopy of human cells and tissue: VIII. Strategies for analysis of infrared tissue mapping data, and applications to liver tissues, Biopolymers 57/5 (2000) 282–290.
[10] L. Zhang, G.W. Small, A.S. Haka, L.H. Kidder, E.N. Lewis, Classification of Fourier transform infrared microscopic imaging data of human breast cells by cluster analysis and artificial neural networks, Appl. Spectrosc. 57/1 (2003) 14–22.
[11] J.R. Mansfield, M.G. Sowa, G.B. Scarth, R.L. Somorjai, H.H. Mantsch, Analysis of spectroscopic imaging data by fuzzy C-means clustering, Anal. Chem. 69/16 (1997) 3370–3374.
[12] J.R. Mansfield, L.M. McIntosh, A.N. Crowson, H.H. Mantsch, M. Jackson, A LDA-guided search engine for the non-subjective analysis of infrared microscopic maps, Appl. Spectrosc. 53 (1999) 1323–1330.
[13] J.B. McQueen, Some methods of classification and analysis of multivariate observations, in: L.M. LeCam, J. Neymann (Eds.), Proceedings of Fifth Berkeley Symposium on Mathematical Statistics and Probability, University of California, Berkeley, 1967, pp. 281–297.
[14] D.L. Massart, L. Kaufmann, The Interpretation of Analytical Chemical Data by the Use of Cluster Analysis, Wiley, New York, 1983.
[15] J.C. Bezdek, Pattern Recognition With Fuzzy Objective Function Algorithms, Plenum, New York, 1981.
[16] M. Jackson, H.H. Mantsch, FTIR spectroscopy in the clinical science, in: H. Clark (Ed.), Biomedical Applications of Spectroscopy, Wiley, New York, 1996, pp. 185–215.
[17] D. Helm, H. Labischinski, G. Schallehn, D. Naumann, Classification and identification of bacteria by Fourier-transform infrared spectroscopy, J. Gen. Microbiol. 137 (1991) 69–79.
[18] M. Diem, S. Boydston-White, L. Chririboga, Infrared spectroscopy of cells and tissues: shining light onto a novel subject, Appl. Spectrosc. 53 (1999) 148A–161A.
[19] OPUS/IDENT Reference Manual, Bruker Optics, Autorenkollektiv Ettlingen, Germany, 1996.
[20] http://www.statsoftinc.com/textbook/stcluan.html.
[21] J.H. Ward, Hierarchical grouping to optimize an objective function, J. Am. Stat. Assoc. 58 (1963) 236–244.
[22] L. Chiriboga, P. Xie, W. Zhang, M. Diem, Infrared spectroscopy of human tissue: III. Spectral differences between squamous and columnar tissue and cells from the human cervix, Biospectroscopy 3 (1997) 253–257.
[23] B.R. Wood, M.A. Quinn, B. Tait, M. Ashdown, T. Hislop, M. Romeo, D. McNaughton, FTIR microspectroscopic study of cell types and potential confounding variables in screening for cervical malignancies, Biospectroscopy 4 (1998) 75–91.
[24] P. Lasch, A. Pacifico, M. Diem, Spatially resolved IR microspectroscopy of single cells, Biopolym. Biospectrosc. 67 (2002) 335–338.
[25] P. Lasch, A. Pacifico, M. Boese, L. Miller, L. Chiriboga, M. Diem, Infrared microspectroscopy and infrared spectral maps of single human cells, Biophys. J. (2004) (in press).