Analyst

PAPER

View Article Online
View Journal

Cite this: DOI: 10.1039/c5an02227d

Development of a memetic clustering algorithm for optimal spectral histology: application to FTIR images of normal human colon†

Ihsen Farah,[a,b] Thi Nguyet Que Nguyen,[a,b] Audrey Groh,[c] Dominique Guenot,[c] Pierre Jeannesson[a,b] and Cyril Gobinet*[a,b]

Received 27th October 2015,
Accepted 7th April 2016

DOI: 10.1039/c5an02227d

www.rsc.org/analyst

The coupling between Fourier-transform infrared (FTIR) imaging and unsupervised classification is effective in revealing the different structures of human tissues based on their specific biomolecular IR signatures; thus the spectral histology of the studied samples is achieved. However, the most widely applied clustering methods in spectral histology are local search algorithms, which converge to a local optimum, depending on initialization. Multiple runs of the techniques estimate multiple different solutions. Here, we propose a memetic algorithm, based on a genetic algorithm and a *k*-means clustering refinement, to perform optimal clustering. In addition, this approach was applied to the acquired FTIR images of normal human colon tissues originating from five patients. The results show the efficiency of the proposed memetic algorithm to achieve the optimal spectral histology of these samples, contrary to *k*-means.

1 Introduction

Fourier transform infrared (FTIR) imaging is a non-destructive, non-invasive and label-free biophotonic technique based on the two-dimensional scan of the light absorbed by a sample. It has been successfully applied to elucidate the histopathological structures of biological tissues.[1,2]

Acquired images are mathematically defined as data cubes composed of two spatial and one spectral dimensions. Each pixel of a FTIR image represents an IR spectrum informative on the molecular composition of the tissue at this acquisition point.

The analysis of a FTIR image is complex for two reasons. Firstly, a large number (several thousands) of spectra composed of numerous wavenumbers (several hundreds) can be acquired. Secondly, the studied phenomenon can generate weak and subtle spectral responses. The interpretation of such a large and highly multi-dimensional data cube is possible by the development and application of advanced numerical data analysis tools. In particular, the application of partitional clustering methods, such as *k*-means (KM)[3,4] and Fuzzy *c*-means (FCM),[4,5] performs an IR spectral histology of the studied tissue, highly correlated to the conventional histology. However, these clustering methods are local search techniques, ensuring only the convergence of the algorithm to a local optimum, which is dependent on the initialization. Thus, applied several times to the same data cube, a partitional clustering algorithm can estimate highly variable solutions.

Metaheuristics are numerical methods designed to find the global optimal solution of any optimization problem. Genetic algorithms (GA),[6] ant colony optimization,[7] and particle swarm optimization[8] are popular examples of population-based metaheuristics. Numerous applications have been reported, such as traveling salesman problem,[9] vehicle routing problem,[10] knapsack problem,[11] bin-packing[12] *etc*. In addition, the hybridization between population-based metaheuristics and local refinement procedures has led to the development of the more efficient memetic algorithms.[13]

In this study, we propose a memetic clustering method combining GA and KM to solve the problem of partitional clustering of the acquired IR spectral images of normal human colon tissue. We show that our method outperforms KM by estimating the optimal histological partition.

2 Materials and methods

2.1 Sample preparation

Five formalin-fixed paraffin-embedded tissue blocks of normal zones were prepared from surgically excised colons from five

[a]Université de Reims Champagne-Ardenne, Equipe MéDIAN-Biophotonique et Technologies pour la Santé, UFR de Pharmacie, 51 rue Cognacq-Jay, 51096 Reims Cedex, France
[b]CNRS UMR 7369, Matrice Extracellulaire et Dynamique Cellulaire (MEDyC), Reims, France. E-mail: cyril.gobinet@univ-reims.fr
[c]Université de Strasbourg (Unistra), EA 3430 Progression tumorale et microenvironnement, Approches translationnelles et Epidémiologie, Fédération de Médecine Translationnelle de Strasbourg (FMTS), Bâtiment U1113, 3 Avenue Molière, 67200 Strasbourg, France

†Electronic supplementary information (ESI) available. See DOI: 10.1039/c5an02227d

patients with colon cancer. For each tissue block, two consecutive 6 μm thick slices were cut with a microtome. For FTIR image acquisition, the first tissue section was mounted onto a calcium fluoride (CaF_2) window (Crystran, Dorset, UK) which is transparent for IR light. For conventional histological analysis, the second tissue section was mounted on a glass window and stained by Harris' Hematoxylin and Eosin (HE). In this study, the stained tissue section is used as a reference for comparison with its corresponding FTIR images.

2.2 FTIR image acquisition

FTIR images were acquired using a Spectrum Spotlight 300 FTIR imaging system coupled with a Spectrum one FTIR spectrometer (Perkin Elmer, Courtabœuf, France), equipped with a nitrogen-cooled mercury cadmium telluride 16-pixel-line detector.

FTIR images were collected with a 6.25 μm spatial resolution, a 4 cm^{-1} spectral resolution, and a 16 scan-averaged accumulation in a mid-IR range of 750 to 4000 cm^{-1}. A 240 scan-averaged reference spectrum was recorded from a blank area of the CaF_2 window in order to substract the background spectrum from the recorded FTIR images, using Spectrum Image software (Perkin Elmer).

On each tissue section, two FTIR images were collected: (i) one on the tissue area for spectral histology, and (ii) one on a pure paraffin zone for numerical dewaxing.

2.3 Spectral data preprocessing

Before applying clustering methods, preprocessing steps must be applied in order to correct the spectra from parasitic signals.

Firstly, an atmospheric correction was performed on each FTIR image by Spectrum Image software to remove water vapor and CO_2 contributions.

Secondly, the spectral range was limited to the 900–1800 cm^{-1} fingerprint region of biological samples.[14]

Thirdly, the spectra were numerically corrected for paraffin signal and baseline, and normalized using the Extended Multiplicative Signal Correction (EMSC) method.[15,16] As previously described in ref. 17, the EMSC parameters were chosen as follows. The reference spectrum is the mean spectrum of the IR spectral image acquired on the tissue area of the considered sample. The interference matrix is composed of the mean spectrum and the first ten principal components computed from the pure paraffin spectra acquired in a pure paraffin zone of the sample. A fourth-order polynomial function was used to model the physical light scattering effects. The efficacy of EMSC is shown in Fig. S1 of the ESI.†

After the preprocessing stage, two different clustering algorithms were applied independently on each IR image in order to highlight the tissue structures of the studied samples: (i) the classical KM clustering which is a local search method, and (ii) a memetic clustering algorithm developed to globally optimize the clustering problem.

2.4 Partitional clustering

2.4.1 *k*-means clustering. KM[18] is the most popular unsupervised classification method. Its aim is to partition into k clusters a set $X = \{x_i | 1 \le i \le n\}$ of n patterns where $x_i = \{x_{il} | 1 \le l \le d\}$ is the i^{th} pattern composed of d features. The clusters are estimated by minimizing the total within-cluster variation defined as:

$$f(X, W, C) = \sum_{j=1}^{k} \sum_{i=1}^{n} w_{ij} \, \| x_i - c_j \|^2 \, . \quad (1)$$

$C = \{c_j | 1 \le j \le k\}$ is the set of barycenters (also called centroids) where $c_j = \{c_{jl} | 1 \le l \le d\}$ is the barycenter of the j^{th} cluster. $\| x_i - c_j \| = \sqrt{\sum_{l=1}^{d} (x_{il} - c_{jl})^2}$ is the Euclidean distance between the i^{th} pattern and the j^{th} barycenter. $W = \{w_{ij} | 1 \le i \le n, 1 \le j \le k\}$ is the membership matrix where:

$$w_{ij} = \begin{cases} 1 & \text{if the } i^{th} \text{ pattern belongs to the } j^{th} \text{ cluster} \\ 0 & \text{otherwise} \end{cases} . \quad (2)$$

W must respect the following constraints:

$$\sum_{i=1}^{n} w_{ij} \ge 1, \quad 1 \le j \le k, \quad (3)$$

$$\sum_{j=1}^{k} w_{ij} = 1, \quad 1 \le i \le n, \quad (4)$$

$$\sum_{j=1}^{k} \sum_{i=1}^{n} w_{ij} = n. \quad (5)$$

Constraint (3) means that each cluster must have at least one pattern. Constraints (4) and (5) imply that each pattern must be assigned to a unique cluster. A KM partition is defined as: $\psi = \{\psi_i | 1 \le i \le n\}$ with $\psi_i = j$ such that $w_{ij} = 1$ and $j \in \{1,2,\ldots,k\}$.

This KM optimization problem can be addressed using the following algorithm:

(i) Randomly choose k patterns as barycenters.

(ii) Assign each pattern to the nearest barycenter in terms of Euclidean distance.

(iii) Update barycenters using $c_j = \left(\sum_{i=1}^{n} w_{ij} x_i\right) / \left(\sum_{i=1}^{n} w_{ij}\right), \quad 1 \le j \le k.$

(iv) Repeat steps (ii) and (iii) until no reassignment of patterns occurs.

2.4.2 Memetic clustering. A memetic algorithm (MA)[13] is a global optimization method based on the hybridization between a population-based approach and a local refinement technique. In this paper, a MA coupling a GA and a KM local search, named memetic clustering (MC), is proposed for the clustering of infrared spectral images. This MC method globally minimizes the KM objective function $f(X,W,C)$ defined in eqn (1).

Problem encoding. A clustering partition is encoded as a chromosome ψ_p, such that $W^{(p)}$ and $C^{(p)}$ correspond to its

membership matrix and centroid matrix, respectively. A chromosome is composed of n genes where each gene ψ_{pi}, $1 \le i \le n$, takes a value in $\{1,2,\ldots,k\}$ according to the string-of-group-number encoding.[19] Then, the i^{th} gene represents the cluster number to which the i^{th} pattern belongs.

Initial population. The initial population is composed of P chromosomes $\Psi = \{\psi_p | 1 \le p \le P\}$. For each chromosome ψ_p, k patterns are randomly selected as barycenters C_p. The i^{th} pattern x_i (*i.e.* the i^{th} gene) is assigned to the cluster with the nearest barycenter $c_j^{(p)}$, *i.e.* the barycenter which minimizes the squared Euclidean distance $\|x_i - c_j^{(p)}\|^2$, $1 \le i \le n$, $1 \le j \le k$. Thus, each chromosome corresponds to an initial partition of the n patterns into k clusters.

Selection operator. GA is based on the generation of children chromosomes from the crossover of parent chromosomes selected from the current population. In this study, a dynamic parent selection was adopted using the roulette wheel strategy[6] coupled with the exponential scaling.[20]

With exponential scaling, the fitness of chromosome ψ_p is measured by $g(\psi_p) = 1/f(X,W^{(p)}C^{(p)})^{E(t)}$, where:

$$E(t) = \tan\left[\left(\frac{t}{T+1}\right)\frac{\pi}{2}\right]\rho. \quad (6)$$

t is the current iteration, T is the total number of iterations. $\rho \in]0;1[$ is a constant controlling the weight of the chromosomes during the selection process. Then, in the roulette wheel strategy, $P/2$ chromosomes are randomly selected with a probability proportional to their fitness values. During the first iterations of MC, the chromosomes have the same probability to be selected (because $E(t) \approx 0$, thus $g(\psi_p) \approx 1$, $\forall p$). At the end of the algorithm, chromosomes with high fitness have a high probability to be chosen (because $E(t) \gg 0$). The higher ρ, the earlier the chromosomes with the highest fitness will be fostered.

In addition, the elitism scheme is used in this work, *i.e.* the best chromosome is automatically selected.

Crossover operator. The goal of the crossover operator is to exchange information between two selected parent chromosomes ψ_{p_1} and ψ_{p_2}, to generate two children chromosomes ψ_{c_1} and ψ_{c_2}. Let b_1 and b_2 be two integers randomly selected in $\{1,2,\ldots,n-1\}$. The first child is composed of genes 1 to b_1 from the first parent, genes $b_1 + 1$ to b_2 from the second parent and genes $b_2 + 1$ to n from the first parent. Thus, $\psi_{c_1} = \{\psi_{p_1 1},\ldots,\psi_{p_1 b_1},\psi_{p_2(b_1+1)},\ldots,\psi_{p_2 b_2},\psi_{p_1(b_2+1)},\ldots,\psi_{p_1 n}\}$. The second child is composed of genes 1 to b_1 from the second parent, genes $b_1 + 1$ to b_2 from the first parent and genes $b_2 + 1$ to n from the second parent. Thus, $\psi_{c_2} = \{\psi_{p_2 1},\ldots,\psi_{p_2 b_1},\psi_{p_1(b_1+1)},\ldots,\psi_{p_1 b_2},\psi_{p_2(b_2+1)},\ldots,\psi_{p_2 n}\}$. This crossover operator can generate chromosomes containing empty clusters. In this case, a correction operator is applied on these chromosomes in order to randomly create new non-empty clusters.

The new population is composed of the $P/2$ selected parents and the $P/2$ generated children.

Mutation operator. The mutation operator acts as a local genetic perturbation on each chromosome to prevent premature algorithm convergence. In our algorithm, we used the mutation operator defined in ref. 21. Each gene has a probability $p_m = 0.05$ to mutate, *i.e.* to change its value. For the p^{th} chromosome, the value of the i^{th} gene subjected to a mutation is selected in the range $\{1,2,\ldots,k\}$ by the roulette wheel procedure described above, using the following fitness function:

$$h_j = d_{\max} - \|x_i - c_j^{(p)}\|^2 \quad (7)$$

where $d_{\max} = \max_{1 \le j \le k}\{\|x_i - c_j^{(p)}\|^2\}$, $c_j^{(p)}$ is the j^{th} updated barycenter of the p^{th} chromosome. The gene value has a high probability to be equal to the cluster number with the nearest barycenter.

Local search operator. The local search operator consists to refine chromosomes computed by the application of the selection, crossover and mutation operators. Its role is to explore the search space in the chromosome neighbourhood in order to accelerate the convergence toward a global optimum. In this work, a limited number N of KM steps is applied on each chromosome as the local search operator.

Finally, MC repeats the selection, crossover, mutation and local search operators until the number of iterations M is reached. The output of the algorithm is the best chromosome of the last population. The different steps of our MC are summarised in the flowchart presented in Fig. 1.

2.5 Choice of the number of clusters

On the acquired IR spectral images of normal human colon tissue sections, KM clustering has been performed for a number of clusters k ranging from 2 to 20. Then, the pathologist compared the estimated partitions of the IR spectral images with the histological structures and substructures that can be identified on the reference HE-stained tissue sections. Consequently, the pathologist has empirically chosen $k = 15$ as the most relevant number of clusters.

This number of clusters was also used for the MC algorithm.

2.6 Clustering pseudo-color-coded images and cluster assignment

After the preprocessing stage, KM and MC clustering were applied separately on each IR image. For each estimated partition, a unique color is attributed to the pixels belonging to the same cluster. Then, the corresponding reconstructed color-coded image is visually analysed by an expert pathologist who annotates each cluster to its corresponding histological class by comparison to a reference HE-stained image.

2.7 Quality measure of a partition

The quality of a partition estimated by KM or MC is measured by the KM objective function f defined in eqn (1). For a given number of clusters k, the smaller the objective function, the better the partition.

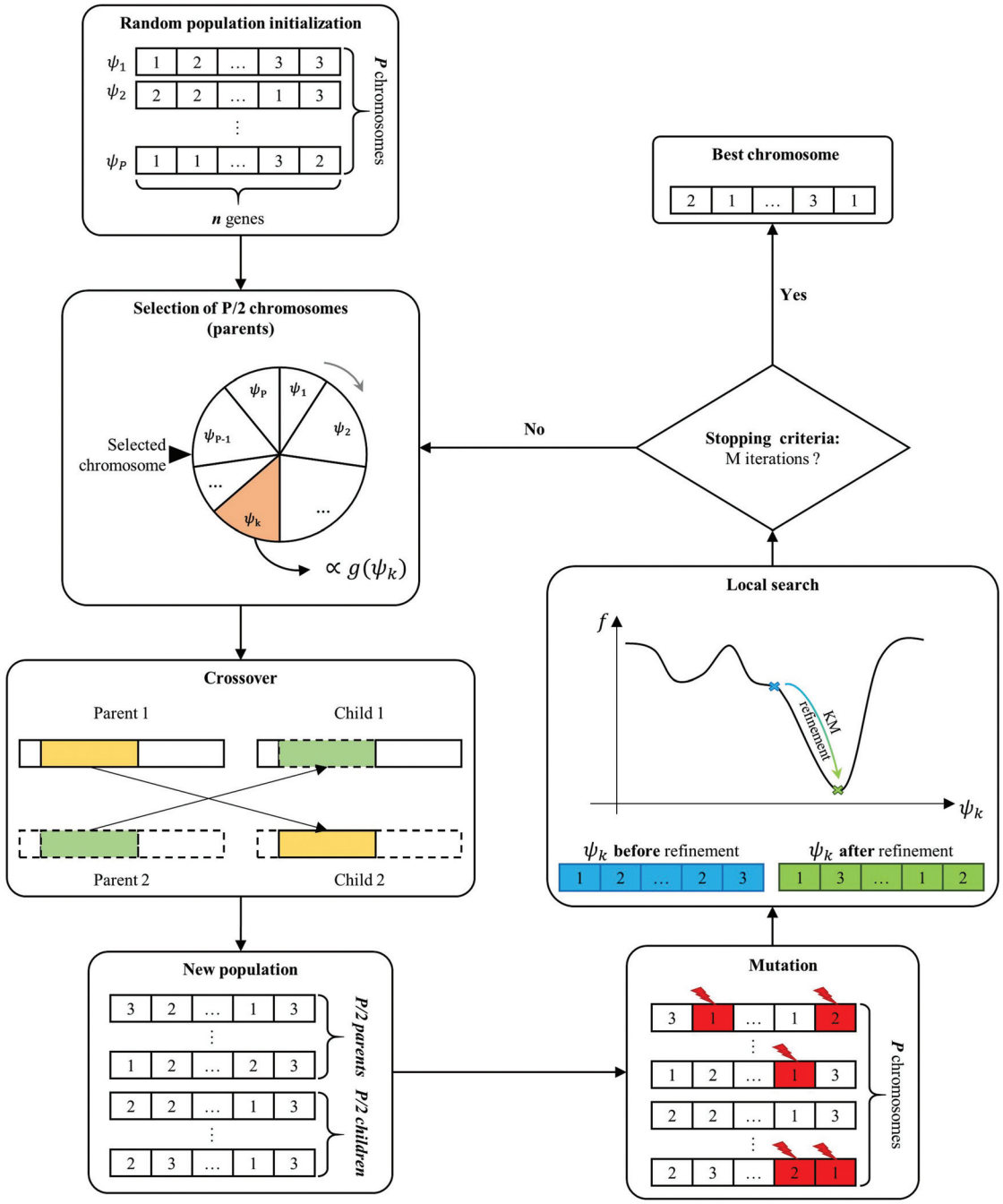

Fig. 1 Flowchart of the developed memetic clustering (MC) algorithm.

3 Results and discussion

3.1 Conventional histology of normal human colon

Here, conventional histology using HE staining is considered as the reference method for the morphological recognition of the tissue structures. Fig. 2 shows the HE-stained colon tissue section of patient #1. This image illustrates the five main histological tissue structures of a normal human colon. The outer layer (mucosa) is composed of tubular glands (crypts of Lieberkühn (structure 1)) and of connective tissue (lamina propria (structure 2)). A thin layer of muscle (muscularis mucosae (structure 3)) links the mucosa to the submucosa (structure 4) which is rich in adipose tissue. Lymphoid aggregates (structure 5) can locally appear in the lamina propria and the submucosa or extend from the lamina propria to the submucosa.[22]

The HE-stained colon tissue sections of the four remaining patients are shown in Fig. S2–S5 of the ESI.†

In this study, the KM and MC pseudo-color-coded images are compared to these HE sections for the

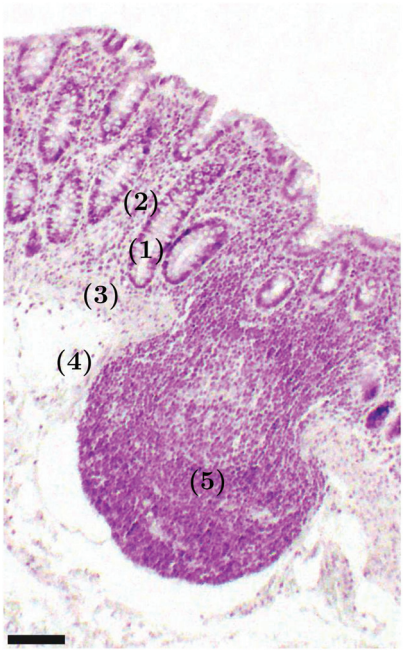

Fig. 2 HE-stained image of normal human colon tissue of patient #1. Its main histological tissue structures are annotated by numbers: (1) the crypts, (2) the lamina propria, (3) the muscularis mucosae, (4) the submucosa and (5) the lymphoid aggregate. Scale bar indicates 100 μm.

assignment of the estimated clusters to the main tissue structures.

3.2 Limitations of KM clustering

In our study, KM is repeated 100 times with $k = 15$ clusters on each tissue section in order to evaluate the variability of KM results due to its random initialization. As mentioned in section 2.7, each partition is evaluated by its estimated objective function value. The second column of Table 1 presents the mean and the standard deviation of these 100 quality measures for the five patients. These results show a high variability of the KM objective function value. Consequently, a high variability of the estimated clusters is visible on KM partitions, as can be seen in Fig. 3 for the tissue section of patient #1. For example, the lamina propria is represented by 2, 3, and 1 clusters on the best (Fig. 3(a)), the most frequent (Fig. 3(b)), and the worst (Fig. 3(c)) partitions, respectively. In addition, only 1% of KM results corresponds to the best partition, justifying the routine application of KM replicates for spectral histology.[23,24] Since KM is a local search method, it converges to a local minimum.[25] Thus, there is no certainty that the best KM partition is the optimal one.

The best, the most frequent, and the worst partitions estimated by KM on the four remaining patients are available in Fig. S2–S5 of the ESI.†

To overcome these KM limitations, MC is proposed to partition data in an optimal way.

3.3 Setting of MC algorithm parameters

A critical phase of MC is the right choice of its parameters, presented in section 2.4.2, in order to ensure the convergence of the algorithm to the optimal solution. In this paper, a grid search using $k = 15$ clusters was used to choose the parameters varying as follows: $P \in \{10, 20, 30, 40, 50, 60, 70\}$, $M \in \{20, 40, 60, 80, 100\}$, $\rho \in \{0.1, 0.3, 0.5, 0.7, 0.9\}$ and $N \in \{1, 5, 10, 15, 20\}$. For each setting of the parameters, the variability of MC is evaluated over 10 replicates, using the acquired IR spectral image of the patient #1 tissue section.

For this spectral dataset, we found that the smallest population size P required to ensure the convergence of MC is for $\rho = 0.1$ (data not shown). Hence, ρ was fixed at 0.1 in the rest of this article.

The remaining parameters were selected by analyzing the mean and standard deviation of the quality measure of the 10 estimated partitions. For each value of N, these results can be represented by a 3D-map. Two kinds of map are observed as shown in Fig. 4. Whatever the values of P and M, the first kind is characterized by a variable mean and a high standard deviation typical of the non-convergence of MC, as shown in Fig. 4(a) for $N = 1$. In contrast, the second kind of map is composed of a flat region defined by a constant mean and a tiny standard deviation, as shown in Fig. 4(b) for $N = 15$. This region is characteristic of the convergence of MC to the global optimal solution. However, on this convergence region, the smaller the P and M values, the shorter the computational time. Thus, this region is summarized by its minimum values of P and M and the results are presented in Table 2. Contrary to the case of $N \in \{1, 5\}$, MC converges to the optimal partition for $N \in \{10, 15, 20\}$ in the tested parameter ranges. Thus, the setting of parameters was driven by the computational time shown in the last column of Table 2. Subsequently, the

Table 1 Mean $\bar{f}$ and standard deviation σ of the partition quality measures computed over 100 replicates for KM, and 10 for MC, GKA and GABC. For each patient, bold values represent the smallest $\bar{f}$ among the four tested clustering methods

	$\bar{f} \pm \sigma$			
Patient	KM	MC	GKA	GABC
#1	21.0075 ± 0.1970	**20.6118** ± 0.0008	20.6317 ± 0.0164	22.4859 ± 0.1769
#2	4.1730 ± 0.0628	**3.9734** ± 0.0004	3.9918 ± 0.0208	4.4089 ± 0.0627
#3	11.4789 ± 0.2418	**10.3126** ± 0.0027	10.3202 ± 0.0001	11.2471 ± 0.1049
#4	15.5702 ± 0.1684	**15.2106** ± 0.0050	15.2366 ± 0.0230	16.8276 ± 0.1886
#5	5.4735 ± 0.0838	**5.3722** ± 0.0004	5.3723 ± 0.0004	6.0431 ± 0.0679

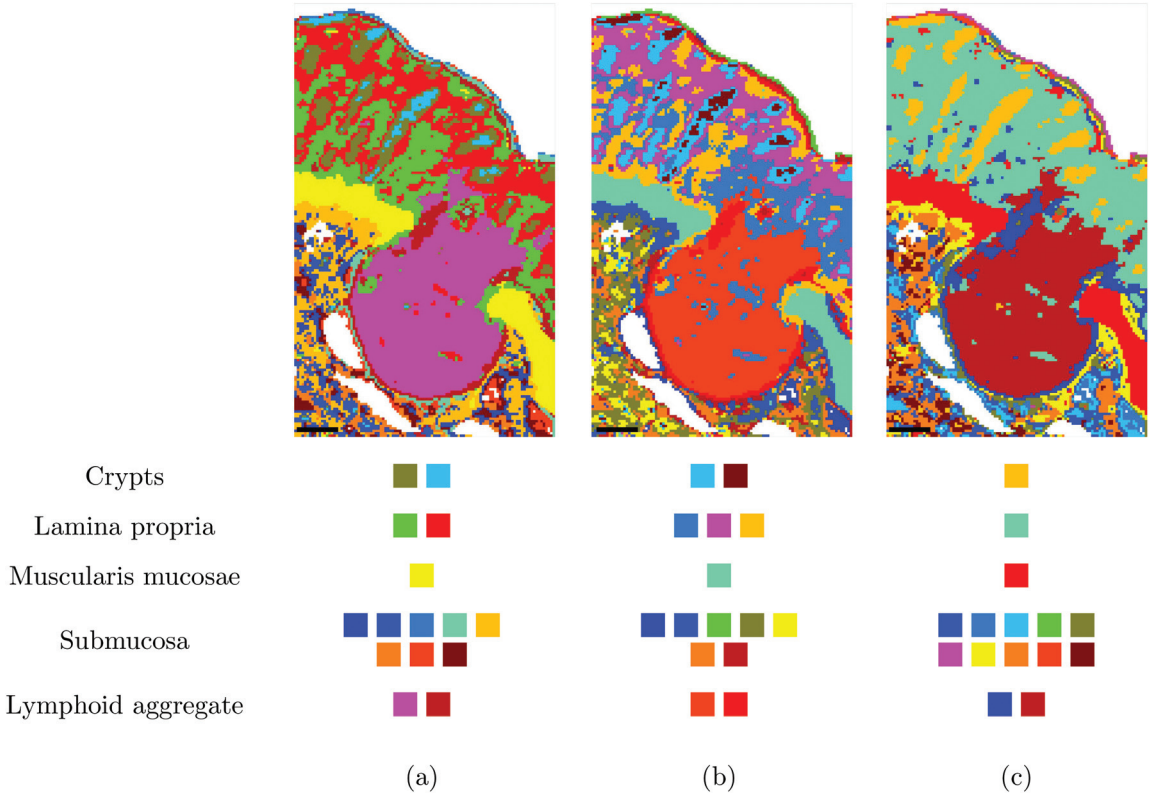

Fig. 3 Examples among the 100 KM partitions estimated for patient #1: (a) the best (f = 20.638), (b) the most frequent (f = 20.931), and (c) the worst (f = 21.771) partitions. Scale bars indicate 100 μm. The cluster assignments are detailed below each pseudo-color-coded image.

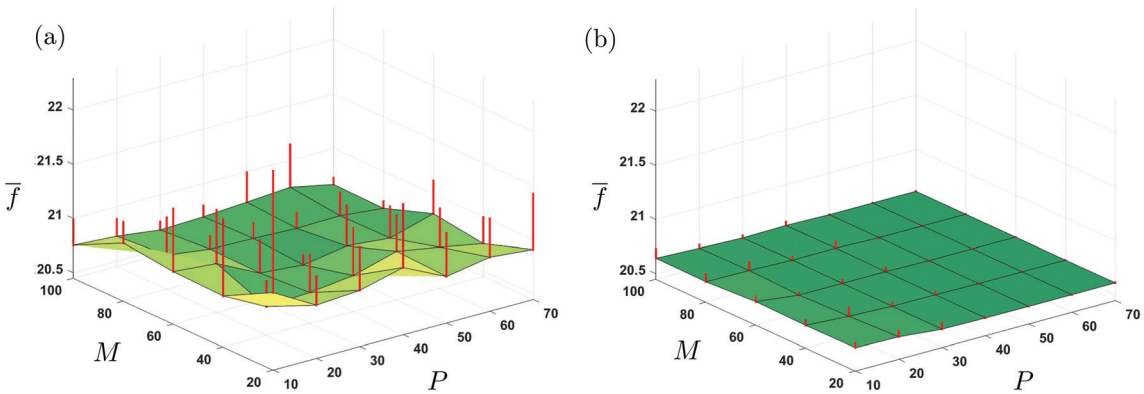

Fig. 4 Examples of grid search maps for (a) N = 1 and (b) N = 15, using ρ = 0.1. Each map represents the quality measure mean $\bar{f}$, over 10 MC replicates, as a function of the population size P and the number of iterations M. Error bars represent the standard deviation σ.

parameters of MC were fixed to ρ = 0.1, N = 15, P = 50, and M = 20 since this setting minimizes the computational time. The efficiency of this setting has been confirmed by the optimal convergence of MC on the spectral images of the remaining patients.

3.4 Efficiency of MC algorithm

In our study, MC is repeated 10 times with k = 15 clusters on each tissue section. MC variability is evaluated by the mean and the standard deviation of the 10 quality measures. These results are summarized in the third column of Table 1. Compared to KM clustering, MC estimates a better solution, since its quality measure is smaller for all the patients. Furthermore, MC is reproducible since its standard deviation is close to zero. These results are consistent with the expectations since MC is a global optimization method contrary to KM which is a local search algorithm.

An example of the optimal partition estimated by MC for patient #1 is given in Fig. 5. The main difference with the best KM partition (Fig. 3(a)) is visible at the level of lamina propria

Table 2 MC convergence regions characterised by the number of KM iterations N, the minimum population size P, the minimum number of iterations M, the mean $\bar{f}$ and the standard deviation σ of the quality measures, and the mean computational time $\bar{t}$ in seconds, over 10 replicates. For a given N, "/" means that MC does not converge whatever the values of P and M

N	P	M	$\bar{f}$	σ	$\bar{t}$
1	/	/	/	/	/
5	/	/	/	/	/
10	70	80	20.6115	0.0005	3906.3951
15	50	20	20.6118	0.0008	793.6144
20	40	40	20.6114	0.0004	1285.7603

represented by two clusters by KM and one by MC. Other small differences can be seen for the other main histological structures.

The optimal partitions estimated by MC on the four remaining patients are available in Fig. S2–S5 of the ESI.† In some cases, KM clustering can converge to the optimal partition as shown in Fig. S5 of the ESI† for patient #5. This event is rare as it depends on the data structure, the chosen number of clusters, and the KM initialization. On the contrary, MC always converges to the optimal solution.

Other clustering algorithms based on metaheuristics have been proposed such as Genetic k-means Algorithm (GKA)[21] and Genetic Algorithm-Based Clustering (GABC).[26] These two algorithms were applied 10 times on each FTIR image. The mean and standard deviation of their corresponding quality measures are given in the columns 4 and 5 of Table 1. These results show that GKA and GABC are less efficient and less reproducible than the MC algorithm, since their mean and standard deviation of quality measures are higher than those of MC. However, since MC and GKA estimated close values of the objective function f, their pseudo-color coded partitions are very similar. On the contrary, the higher value of f for GABC induces a completely different partition of the data. These results are illustrated for patient #1 in Fig. S6 of the ESI.†

An important characteristic of metaheuristics is the computational time. Table 3 presents the mean and standard deviation of the computational time of MC, GKA, and GABC over 10 replicates. These data show that MC is four times faster than GKA and GABC.

Several studies have shown the efficacy of metaheuristics applied to IR spectral data. For example, genetic algorithms have been developed for the supervised classification of the acquired FTIR spectra of different species of bacteria,[27] for the optimal selection of discriminant subsets of wavelengths,[24,28–30] and for the selection of the best sequence of preprocessing steps applied to spectral data.[28] To the best of our knowledge, this is the first study presenting a metaheuristics-based algorithm specifically developed for the clustering of IR images.

The proposed MC has been proven effective to perform the spectral histology of human normal colon tissues from 5 patients. To confirm the validity of our approach, the comparative study of KM, MC, GKA and GABC has been extended to supplementary normal colon tissues from 10 other patients. These results are available in Table S1 of the ESI.† Moreover, MC is based on a general framework suitable to all kinds of data with different cluster shapes. Indeed, a simple and effective approach to adapt our MC algorithm to seek clusters of different shapes is to change the distance metric of the objective function $f(X,W,C)$ (eqn (1)) used in both the memetic algorithm and the local search operator. For example, Minkowski[31] and Mahalanobis[32] metrics can be used for estimating elliptic clusters. Other cluster shapes can be determined using kernel-based metrics.[32,33] MC can thus be applied to the acquired infrared images of human samples from other organs with different physiopathological states, without degrading its performance.

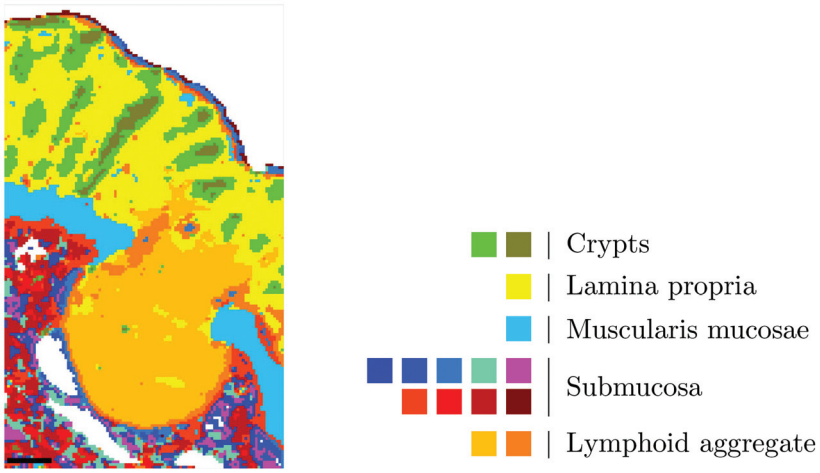

Fig. 5 Optimal partition estimated by MC for patient #1 (f = 20.611). Scale bar indicates 100 μm. The cluster assignments are detailed beside the pseudo-color-coded image.

Table 3 Mean $\bar{t}$ and standard deviation σ_t of the computational time (in seconds) over 10 replicates for MC, GKA and GABC. For each patient, bold values represent the smallest $\bar{t}$ among the three tested clustering methods

	$\bar{t} \pm \sigma_t$		
Patient	MC	GKA	GABC
#1	**793.6144** ± 19.7452	3713.1001 ± 35.3542	3349.3905 ± 6.1103
#2	**553.4608** ± 37.0554	2136.6738 ± 21.0275	2722.8798 ± 26.6593
#3	**632.9907** ± 46.0887	2561.5067 ± 29.0262	2090.8268 ± 27.4400
#4	**728.1194** ± 6.8380	3599.2612 ± 69.6974	2871.8835 ± 21.8029
#5	**412.0091** ± 42.4198	1775.1262 ± 29.5005	2221.7898 ± 18.7104

4 Conclusion

In this study, an optimal memetic clustering (MC) combining a genetic algorithm and a refinement by KM was developed. Applied on the acquired FTIR images of normal human colon tissue samples originating from five patients, this method outperformed standard KM and two popular genetic algorithm-based clustering techniques namely GKA and GABC. Compared to these three methods, our algorithm reproducibly converges to the optimal solution. In addition, MC is four times faster than GKA and GABC. Owing to its general framework, our algorithm may be applied for the spectral histology of any kind of tissue and for the clustering of any kind of data.

Acknowledgements

The authors thank Cancéropôle Grand-Est, Ligue contre le Cancer, the URCA technological platform of cellular and tissular imaging PICT-IBiSA, Région Champagne-Ardenne, Région Alsace, and Ministère de l'Enseignement Supérieur et de la Recherche for financial support, and Shawn Hussain for linguistic assistance.

References

1 M. Diem, A. Mazur, K. Lenau, J. Schubert, B. Bird, M. Miljković, C. Krafft and J. Popp, *J. Biophotonics*, 2013, **6**, 855–886.
2 M. Diem, *Modern vibrational spectroscopy and micro-spectroscopy: theory, instrumentation and biomedical applications*, John Wiley & Sons, 2015.
3 C. Krafft, D. Codrich, G. Pelizzo and V. Sergo, *Analyst*, 2008, **133**, 361–371.
4 P. Lasch, W. Haensch, D. Naumann and M. Diem, *Biochim. Biophys. Acta, Mol. Basis Dis.*, 2004, **1688**, 176–186.
5 D. Sebiskveradze, V. Vrabie, C. Gobinet, A. Durlach, P. Bernard, E. Ly, M. Manfait, P. Jeannesson and O. Piot, *Lab. Invest.*, 2011, **91**, 799–811.
6 J. H. Holland, *Adaptation in Natural and Artificial Systems*, University of Michigan Press, Ann Arbor, 1975.
7 M. Dorigo and C. Blum, *Theor. Comput. Sci.*, 2005, **344**, 243–278.
8 J. Kennedy, *Encyclopedia of Machine Learning*, Springer, 2010, pp. 760–766.
9 M. Dorigo and L. M. Gambardella, *BioSystems*, 1997, **43**, 73–81.
10 B. M. Baker and M. A. Ayechew, *Comput. Oper. Res.*, 2003, **30**, 787–800.
11 J. C. Bansal and K. Deep, *Appl. Math. Comput.*, 2012, **218**, 11042–11061.
12 E. Falkenauer, *J. Heuristics*, 1996, **2**, 5–30.
13 F. Neri, C. Cotta and P. Moscato, *Handbook of memetic algorithms*, Springer, 2012, vol. 379.
14 M. Khanmohammadi, A. B. Garmarudi, K. Ghasemi, H. K. Jaliseh and A. Kaviani, *Med. Oncol.*, 2009, **26**, 292–297.
15 R. Wolthuis, A. Travo, C. Nicolet, A. Neuville, M. P. Gaub, D. Guenot, E. Ly, M. Manfait, P. Jeannesson and O. Piot, *Anal. Chem.*, 2008, **80**, 8461–8469.
16 A. Kohler, N. Kristian Afseth and H. Martens, *Handbook of Vibrational Spectroscopy*, John Wiley & Sons, Ltd, 2006.
17 T. N. Q. Nguyen, P. Jeannesson, A. Groh, D. Guenot and C. Gobinet, *Analyst*, 2015, **140**, 2439–2448.
18 J. MacQueen, *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, 1967, vol. **1**, p. 14.
19 D. R. Jones and M. A. Beltramo, *Proceedings of the 4th International Conference on Genetic Algorithms*, 1991, pp. 442–449.
20 Z. Michalewicz, *Genetic algorithms + data structures = evolution programs*, Springer-Verlag, 1991.
21 K. Krishna and M. N. Murty, *IEEE Trans. Syst. Man Cybern. B Cybern.*, 1999, **29**, 433–439.
22 P. M. Treuting and S. M. Dintzis, *Comparative anatomy and histology: a mouse and human atlas*, Academic Press, 2011.
23 E. Ly, O. Piot, R. Wolthuis, A. Durlach, P. Bernard and M. Manfait, *Analyst*, 2008, **133**, 197–205.
24 E. Ly, O. Piot, A. Durlach, P. Bernard and M. Manfait, *Analyst*, 2009, **134**, 1208–1214.
25 A. K. Jain, M. N. Murty and P. J. Flynn, *ACM Comput. Surv.*, 1999, **31**, 264–323.
26 U. Maulik and S. Bandyopadhyay, *Pattern Recogn.*, 2000, **33**, 1455–1465.

27 R. Goodacre, B. Shann, R. J. Gilbert, E. M. Timmins, A. C. McGovern, B. K. Alsberg, D. B. Kell and N. A. Logan, *Anal. Chem.*, 2000, **72**, 119–127.
28 R. M. Jarvis and R. Goodacre, *Bioinformatics*, 2005, **21**, 860–868.
29 G. N. Elliott, H. Worgan, D. Broadhurst, J. Draper and J. Scullion, *Soil Biol. Biochem.*, 2007, **39**, 2888–2896.
30 J. Trevisan, P. P. Angelov, P. L. Carmichael, A. D. Scott and F. L. Martin, *Analyst*, 2012, **137**, 3202–3215.
31 P. J. Groenen and K. Jajuga, *Fuzzy Set. Syst.*, 2001, **120**, 227–237.
32 D. Graves and W. Pedrycz, *Fuzzy Set. Syst.*, 2010, **161**, 522–543.
33 D.-W. Kim, K. Y. Lee, D. Lee and K. H. Lee, *Pattern Recogn.*, 2005, **38**, 607–611.