Anal Bioanal Chem
DOI 10.1007/s00216-009-2794-9

ORIGINAL PAPER

Pollen discrimination and classification by Fourier transform infrared (FT-IR) microspectroscopy and machine learning

R. Dell'Anna · P. Lazzeri · M. Frisanco · F. Monti ·
F. Malvezzi Campeggi · E. Gottardini · M. Bersani

Received: 13 February 2009 / Revised: 2 April 2009 / Accepted: 3 April 2009

Abstract The discrimination and classification of allergy-relevant pollen was studied for the first time by mid-infrared Fourier transform infrared (FT-IR) microspectroscopy together with unsupervised and supervised multivariate statistical methods. Pollen samples of 11 different taxa were collected, whose outdoor air concentration during the flowering time is typically measured by aerobiological monitoring networks. Unsupervised hierarchical cluster analysis provided valuable information about the reproducibility of FT-IR spectra of the same taxon acquired either from one pollen grain in a 25×25 μm^2 area or from a group of grains inside a 100×100 μm^2 area. As regards the supervised learning method, best results were achieved using a *K* nearest neighbors classifier and the leave-one-out cross-validation procedure on the dataset composed of single pollen grain spectra (overall accuracy 84%). FT-IR microspectroscopy is therefore a reliable method for discrimination and classification of allergenic pollen. The limits of its practical application to the monitoring performed in the aerobiological stations were also discussed.

Keywords FT-IR microspectroscopy · Allergic pollen · Supervised and unsupervised learning methods · Aerobiological monitoring networks

Introduction

Identification of airborne pollen grains is of great importance for allergy studies. Pollen from grasses and trees can cause symptoms of rhino-conjunctivitis and asthma [1, 2]. Pollen seasons can last for several months, and outdoor exposure is difficult to avoid. Aerobiological monitoring networks provide information about airborne pollen concentrations to allergists and allergic sufferers, therefore triggering timely prophylaxis and therapies to prevent or reduce allergic symptoms [3].

For conventional analysis [4, 5], the Hirst spore trap is used: a known outdoor air volume is drawn through an orifice by suction from a vacuum pump, and the airborne particles are collected by impaction on an adhesive surface [6]. The sampling surface is subsequently examined with an optical microscope for the identification and count of captured pollen. The procedure relies on the analysis of morphological information such as grain sizes, shapes, apertures, and surface structures. Because of different flowering times, in a single analysis session some hundreds to some thousands of pollen grains and several different pollen taxa can be considered. The monitoring requires qualified operators and is an extremely time-consuming task. In some cases, this last aspect limits the geographic extension of the aerobiological monitoring network. As a

R. Dell'Anna (✉) · P. Lazzeri · M. Bersani
Center for Materials and Microsystems, Fondazione Bruno Kessler,
Via Sommarive 18,
38100 Trento, Italy
e-mail: dellanna@fbk.eu

M. Frisanco
Fondazione Bruno Kessler & CNR Istituto di Biofisica,
Via alla Cascata 56/C,
38100 Trento, Italy

F. Monti · F. Malvezzi Campeggi
Dipartimento di Informatica, Università degli Studi di Verona,
Strada Le Grazie 15,
37134 Verona, Italy

E. Gottardini
Area Ambiente, FEM-Centro Ricerca ed Innovazione,
Via E.Mach 1,
38010 San Michele all'Adige, Trento, Italy

Published online: 25 April 2009

consequence, especially in regions where local climatic and vegetation conditions are highly variable, detailed information about the exposure of population to allergenic pollen is not possible. Therefore, alternative methods, possibly highly automated, for the rapid identification of airborne pollen grains are extremely desirable.

Fourier transform infrared (FT-IR) and Raman spectra provide chemical rather than morphological information because they enable the investigation of the vibrational dynamics of biochemical components, such as lipids, peptides, proteins, nucleic acids, and sugars. Naumann and coworkers [7] first applied FT-IR and Raman spectroscopies together with multivariate statistical analyses and pattern recognition methodologies to the rapid differentiation, identification, and classification of microorganisms. FT-IR spectroscopy has been shown to be a very useful technique for the analysis of different types of biological samples [8, 9]. It has been successfully used in the past in particular for the characterization of microorganisms [10–14]. FT-IR microspectroscopy has emerged as a key technique for the study of plant growth and development at a cellular level ([15] and references therein). In conjunction with supervised multivariate statistical methods, it has been shown to be a very promising technique for the characterization of cell wall changes in *Arabidopsis mutants*, used as a model for studying plant mutants biology ([16] and references therein).

Both FT-IR and Raman techniques have been employed to discriminate among different allergenic pollens. The first Raman spectroscopy studies [17, 18] demonstrated the possibilities of pollen characterization; however, the spectra presented a strong fluorescence background, limiting the analysis. Pappas et al. [19] obtained FT-IR spectra by sampling macroscopic quantities of pollen of the same taxon. They demonstrated the existence of peculiar spectral features able to discriminate among different species. Aerobiological samples, however, are composed by grains of different taxa, which could be available only in limited quantities. Gottardini et al. [20] confirmed results of [19] and also assessed the potential of FT-IR spectroscopy to discriminate two different pollen taxa in a very unbalanced binary mixture. Ivleva et al. [21] used Raman microscopy, therefore obtaining spectra from single grains, and applied unsupervised multivariate analysis to cluster four different pollen taxa. In [22], the results of the in situ chemical characterization of pollen grains by Raman microspectroscopy were discussed. In addition, unsupervised hierarchical cluster analysis (HCA) was carried out on spectra from pollen samples of 15 different species to investigate taxonomic groups.

In this paper, we applied mid-infrared FT-IR microspectroscopy together with unsupervised (HCA) and supervised (*K* nearest neighbors, *K*-NN classifier) learning methods to discriminate and automatically classify pollen grains from 11 different allergy-relevant species, by only considering the acquired pollen grain spectra. To our knowledge, FT-IR microspectroscopy together with multivariate statistical analyses has not yet been applied to pollen characterization and discrimination.

Our spectroscopic work was part of a more extended project (also including a biomolecular approach), aiming at assessing different innovative methodologies for the rapid identification of airborne allergenic pollen. Besides the interest for the pure spectroscopic and biomolecular challenges, particular attention was devoted in this project to the practical aspect of the real applicability of these alternative techniques to the monitoring currently performed in dedicated aerobiological stations. In this context, FT-IR microspectroscopy appeared particularly promising, since it allowed the acquisition of a spectrum from a single pollen grain in a 25× 25 μm^2 sampling area. Consequently, in our study the automatic classification could be applied to separately identify each single grain from its FT-IR spectrum. This means that, if the grains were collected in an aerobiological station, the identified grains could be straightaway counted (just as the morphological identification of each single grain allows to count them in the conventional analysis), and the concentration in air of the pollen grains of each analyzed taxon could be therefore easily calculated. From this point of view, this paper also discussed the potentialities of FT-IR microspectroscopy as a practical method that could enable the geographic extension of existing aerobiological monitoring networks.

Experimental

Pollen samples

Samples of allergy-relevant pollen were collected at flowering time from each of the following 11 plants: *Alnus glutinosa* L. Gaertner (alder) (AL in the figures), *Artemisia vulgaris* L. (mugwort) (AR), *Betula pendula* Roth (silver birch) (B), *Castanea sativa* Miller (sweet chestnut) (CA), *Corylus avellana* L. (hazel) (CO), *Cupressus arizonica* Greene (Arizona cypress) (A), *Cupressus sempervirens* L. (Italian cypress) (S), *Dactylis glomerata* L. (cocksfoot) (D), *Fraxinus ornus* L. (manna ash) (F), *Olea europaea* L. (olive) (OL), and *Ostrya carpinifolia* Scop. (hop hornbeam) (OS). All samples were dried at 4°C in a desiccator and kept at this temperature until use. The pollen grains of each vegetal species were collected from three different plants located in three different geographical sites of Trentino region. The selected 11 species are typically monitored in the aerobiological stations of this territory. They are reported in Table 1 together with their respective plant family.

Table 1 The plant species considered in this study and their corresponding plant family

Family	Species
Betulaceae	*Alnus glutinosa* L. Gaertner
Betulaceae	*Betula pendula* Roth
Compositae	*Artemisia vulgaris* L.
Fagaceae	*Castanea sativa* Miller
Corylaceae	*Corylus avellana* L.
Corylaceae	*Ostrya carpinifolia* Scop.
Cupressaeae	*Cupressus arizonica* Greene
Cupressaeae	*Cupressus sempervirens* L.
Graminaceae	*Dactylis glomerata* L.
Oleaceae	*Fraxinus ornus* L.
Oleaceae	*Olea europaea* L.

FT-IR measurements

Mid-infrared spectra were acquired for each taxon in transmission mode on a CaF_2 support in the 4,000–850 cm^{-1} range using a Bruker Optics Vertex 70 spectrometer coupled to a Hyperion 3000 vis/IR microscope equipped with a standard photoconductive MCT detector and a ×15 objective. Species by species, five spectra, each one obtained from a different single pollen grain inside a 25×25 μm^2 area, were collected at 4 cm^{-1} resolution by coadding 512 scans (corresponding to an acquisition time of about 220 s for each spectrum). Moreover, three spectra, each one obtained from a group of pollen grains of the same taxon inside a 100×100 μm^2 area, were also acquired, by coadding 64 scans (corresponding to about 27 s of acquisition time) at the same spectral resolution. Figure 1 shows how the microscope allowed to view the sample and to choose the desired aperture. A FT-IR spectrum could be acquired from the selected measuring area. Two data matrices were built by separately collecting the single grain spectra (dataset A) and the multigrain spectra (dataset B).

Data analysis

All spectra were treated working in the R 2.3.0 software environment [23]. Data analysis focused on three distinct phases: (1) development and application of in-house numerical software for a rapid and automatic preprocessing (baseline correction, normalization, smoothing, and first derivative) of all spectra; (2) application of agglomerative HCA to explore both the relationships between spectra and the existence of spectra groups useful in terms of pollen grain discrimination; (3) application of a classifier based on the *K*-NN rule to identify the species membership of unknown single grain spectra. Methods in (2) and (3) were applied using the *stats*, *amap*, and *class* statistical packages of *R*, after mean-centering and scaling to unit variance the feature vectors corresponding to each spectral channel. For all performed multivariate analyses, we tested the whole spectral range and different spectral windows and verified that the maximum information and discrimination power was achieved by only considering the spectral range between 1,800 and 850 cm^{-1}, which provides very specific spectral contributions to distinguish different taxa [19, 20]. Therefore, all presented results were obtained in that range. The preprocessing operations described hereinafter were also carried out in the 1,800–850 cm^{-1} range.

Automatic preprocessing

We investigated FT-IR microspectroscopy as a method that could substitute the current approach of morphological recognition of different pollen taxa. In this context, the automation of the spectral analysis process is crucial to speed-up the FT-IR approach, in order to contribute to the attempt of extension of the aerobiological monitoring network. In particular, manual interventions in the preprocessing step are time-consuming. Therefore, we developed a software program for performing in a completely automatic way baseline shift correction, unit area normalization, and spectra smoothing based on the Savitsky-Golay algorithm (11-point moving fourth degree polynomial) [24]. The program only requires as input all the spectra to be treated. It takes approximately 20 s to preprocess 100 spectra (495 spectral channels). The subsequent calculation of the first derivative of each spectrum is also possible. For 100 spectra, this last operation requires less than 1 s.

Hierarchical cluster analysis

Cluster analysis [25] is the process of grouping objects into clusters that have meaning in the context of a particular

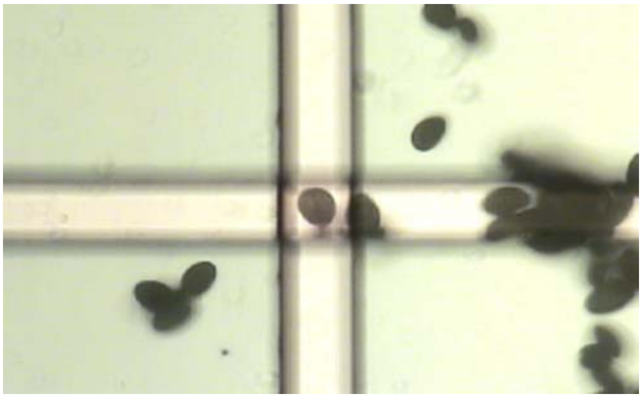

Fig. 1 The photo shows how the microscope allows to view the sample and to choose the desired aperture. A FT-IR spectrum can then be acquired from the selected measuring area. In this image, a single *Fraxinus ornus* pollen grain is centered inside a 25×25 μm^2 aperture

problem. Clustering techniques are unsupervised learning methods because no a priori examples of cluster membership are provided. By repeatedly linking pairs of most similar clusters until every data object is included in the hierarchy, agglomerative hierarchical clustering produces bottom–up a dendrogram. Such an approach allows exploring data on different levels of granularity. A measure of similarity between objects (metric) and one of similarity between clusters (linkage method) must be defined. In this work, we verified that the best results were obtained using a metric based on Pearson's correlation coefficient and the average linkage method. In addition, HCA used the first derivatives of the spectra, as we verified that this allows to maximize information extraction.

K nearest neighbor classifier

The *K*-NN classifier [26] is based on the assumption that the classification of an instance (i.e., a spectrum) is most similar to the classification of other instances that are nearby in the feature space. Therefore, given a new case, its *K* nearest cases in the vector space are found, and the class that appears most frequently among the *K* neighbors is chosen. The classifier performances crucially depend on the choice of *K* and of the distance measure. In this work, the best results were obtained using a metric based on Pearson's correlation coefficient and choosing *K*=1. We also verified that the best classifier performances were achieved using the first derivatives of the spectra.

It is worth noting that this classification procedure is completely automatic. Neither human intervention nor expert evaluation of the classification results as well as of the different computational steps are required. A spectrum is acquired, and in a completely automatic way, it is first preprocessed, and its class membership is subsequently provided. In fact, one goal of our project was the usage of this automatic classification in the aerobiological monitoring stations, where people not expert in the area of IR microspectroscopy and statistical analysis are present.

Results and discussion

As an example, Fig. 2 shows the FT-IR transmission spectra as acquired before preprocessing in the 4,000–850 cm^{-1} range on five different single pollen grains (a) and on three different multigrain samples (b) of *A. vulgaris* (mugwort). Differences are clearly visible. In particular, spectra from single grains are more variable than multigrain spectra. As a matter of fact, single grains of the same taxon present different dimensions and a certain degree of chemical variability. Also, pollen immaturity can produce spectral variations. Spectra from a group of grains, reported in Fig. 2b, on the contrary, are intrinsically averaged and therefore less variable. Similar observations were made when considering the spectra recorded from the other ten pollen taxa of this study.

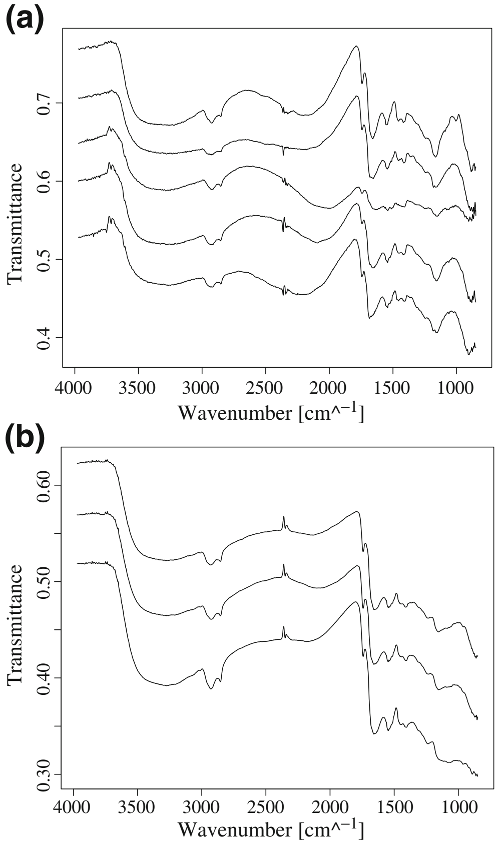

Fig. 2 Comparison among the acquired FT-IR transmission spectra (before preprocessing) of *A. vulgaris* in the whole 4,000-850 cm^{-1} range obtained **a** from five different single pollen grains and **b** on three different multigrain samples. In both panels, the *scale on the vertical axis* shows the transmittance values for the central spectrum. The others are vertically shifted for an easy view

A detailed interpretation of absorption bands and of the biochemical meaning of their diversity among the various species is beyond the scope of our work. As a matter of fact, an advantage of the data mining analyses (HCA and *K*-NN classifier) performed in this study on the acquired FT-IR spectra is that they do not require a priori knowledge or assumptions about the spectral features. In addition, indeed because no spectral interpretation is needed, the discussed classifier could be particularly well suited to be used in aerobiological monitoring stations, where people not expert in the analysis of FT-IR spectra absorption bands work. According to reference [19], where molecular vibrations probed by FT-IR spectroscopy were carefully analyzed, we verified that the 1,800–850 cm^{-1} region allowed to obtain the most significant results as regards supervised and unsupervised analyses. Main absorption bands in this spectral range are around 1,660 cm^{-1} (Amide I) and 1,550 cm^{-1} (Amide II and lignin), around 1,460 and 1,410 cm^{-1} (mainly from lipids and proteins, respectively), and around 1,200 cm^{-1} where a broad and structured absorption band mainly related to carbohydrates appears. As an example, Fig. 3a–c shows one spectrum for each taxon taken from dataset B, i.e. obtained from a group of pollen grains of the same vegetal species inside a 100× 100 μm^2 window after baseline subtraction and area normalization in the 1,800–850 cm^{-1} region.

To further examine the variability of FT-IR spectra, HCA was separately applied to datasets A and B. Obtained results are reported in Fig. 4a, b. In Fig. 4b, the investigated taxa cluster in distinct groups, with the only exception of *O. europaea*. Differently, in Fig. 4a, all *C. sativa*, *C. arizonica*, and *C. sempervirens* spectra form three separate clusters. The results are also satisfying for *A. vulgaris*, *F. ornus*, and *O. carpinifolia* (four out of five spectra are correctly clustered), *A. glutinosa* and *C. avellana* (three out of five spectra), while for the other taxa (*B. pendula*, *D. glomerata*, and *O. europaea*), spectra are spread across the dendrogram as doublets or single spectra. Therefore, we confirmed that spectra from single grains (dataset A) were more variable than multigrain spectra (dataset B), but even more important for the classification results hereinafter presented, we concluded that the degree of variability was different for different plant species.

HCA is an unsupervised method for discovering natural groups of data objects without giving predefined classes but simply identifying potential classes. However in this study, we were particularly interested in building a supervised classification system, which is in general able to extract a decision rule [27] from correctly identified taxon spectra (training set) that will be applicable to classify unknown single-grain pollen spectra (testing set). In fact, this approach could then be used for the classification of collected airborne pollen grains. The supervised method could in principle improve the results of HCA. We therefore completed our study by applying a *K*-NN classifier, with the aim of building a correctly validated, though simple, classifier and to explore the possibility of obtaining an automated procedure for airborne pollen monitoring. The *K*-NN classifier was chosen since it is an uncomplicated yet powerful classification method. In addition, differently from other classifiers, it works even though the data dimensionality is greater than the number of samples, as it was the case of the present study. In this way, it was not necessary to apply a dimension reduction or a feature selection step, but the whole spectral range 1,800–850 cm^{-1} could be used, following in this sense references [19, 20]. Therefore, similarly to [22], we aimed at testing whether the full spectral patterns in the 1,800–850 cm^{-1} region could be successfully used like fingerprint for an overall chemical supervised classification.

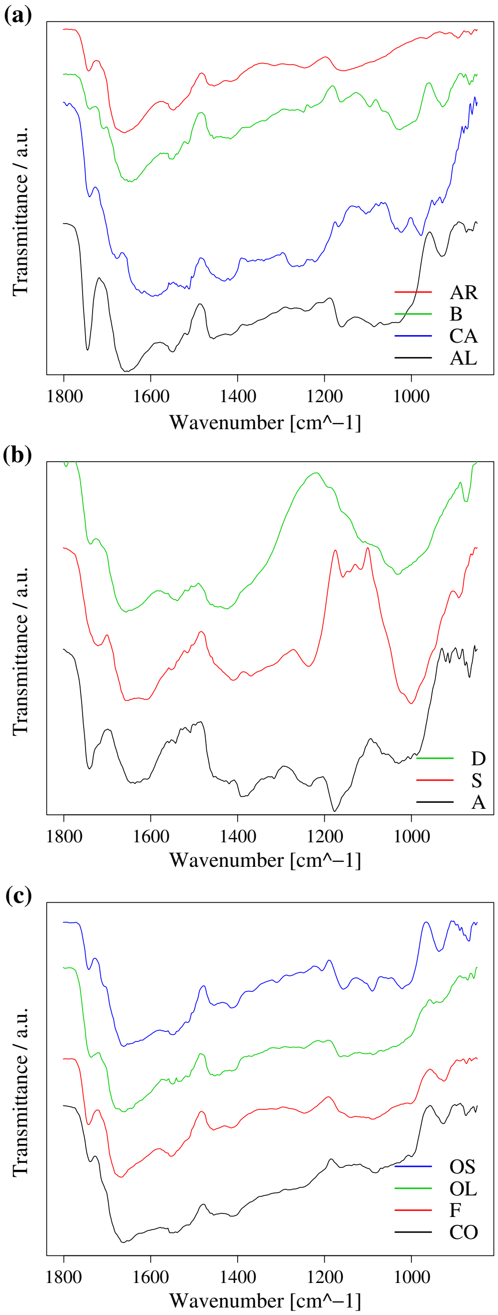

Fig. 3 Eleven FT-IR spectra, each one obtained in the 1,800–850 cm^{-1} spectral region from a group of pollen grains of the same taxon inside a 100×100 μm^2 area. All spectra are baseline corrected and normalized as described in the data analysis subsection. For readability reasons, the spectra are reported in three distinct panels (**a**–**c**) and vertically shifted. *AL*, *Alnus glutinosa*; *AR*, *Artemisia vulgaris*; *B*, *Betula pendula*; *CA*, *Castanea sativa*; *CO*, *Corylus avellana*; *A*, *Cupressus arizonica*; *S*, *Cupressus sempervirens*; *D*, *Dactylis glomerata*; *F*, *Fraxinus ornus*; *OL*, *Olea europaea*; *OS*, *Ostrya carpinifolia*

We investigated two different classification approaches: the first one used a training set (i.e., a reference library of

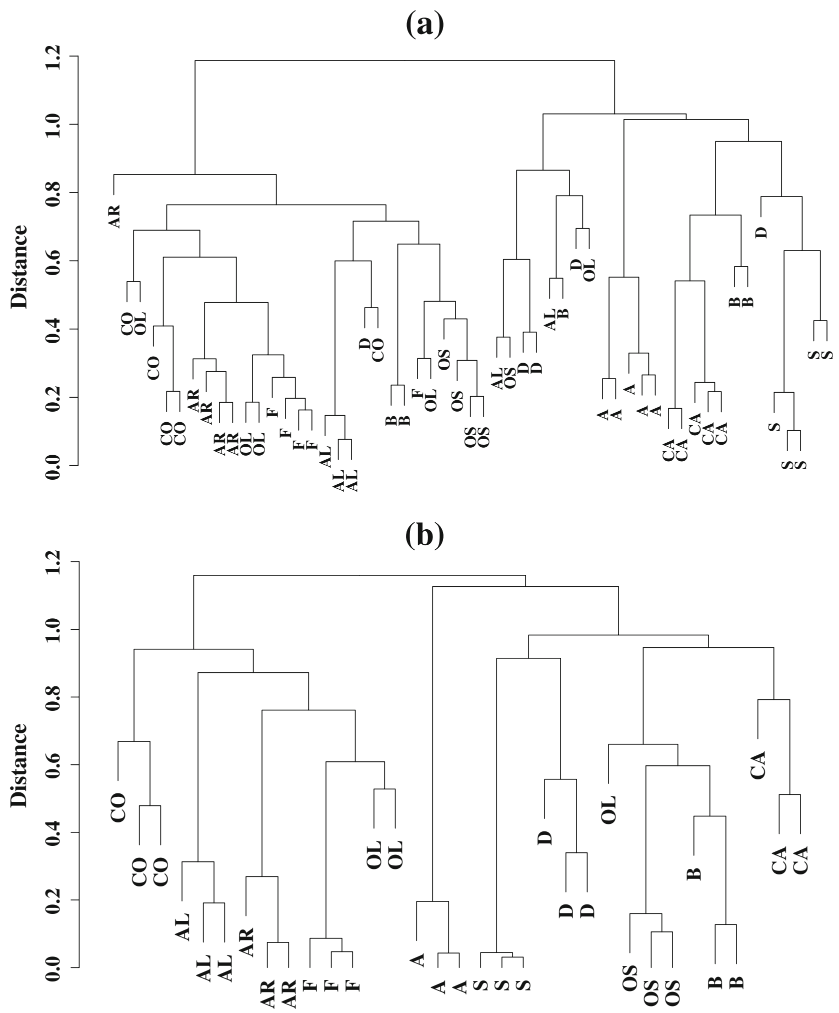

Fig. 4 Agglomerative hierarchical clustering respectively obtained **a** for dataset A, composed of single-grain FT-IR spectra of the 11 plant species studied, and **b** dataset B, composed of multi-grain FT-IR spectra of the same species. Legend, see Fig. 3

identified taxon spectra) of single pollen grain spectra (classifier A), while the training set of the second one was obtained from the multigrain spectra (classifier B).

Classifier A

Classifier A was built using exclusively dataset A. To measure the classifier performance, we estimated its true error rate (or alternatively its overall accuracy) by the leaving-one-out cross-validation method [26]. Therefore, given the n=55 spectra of dataset A, the library was generated using 54 cases, and the single remaining case was considered as unknown and identified by applying the K-NN classifier. This approach was repeated n times. The true error rate is the number of errors on the single test cases divided by n. The overall accuracy of classifier A was 84% (true error rate 16%). This is clearly much better than a random classifier, whose expected accuracy for 11 classes is 9.1%, and as such, FT-IR microspectroscopy is effective for the identification of pollen grains. In addition, this accuracy is greater than that (~80%) currently estimated for the morphological recognition approach used in our aerobiological monitoring network. Nevertheless, we investigated whether these results were precise enough to build an automated classification system alternative to the morphological approach. Therefore, we calculated the confusion matrix of classifier A, reported in Table 2, which gives a more detailed picture of the errors made by the classifier, because instead of simply analyzing the number of correct and incorrect predictions, it shows the type of errors being made. In Table 2, each row of the matrix represents the spectra of an actual taxon (class), while each column represents the spectra in a predicted class. Entries on the diagonal are correct predictions. The last column and last row of Table 2 give respectively the true positive (TP) values for each actual class, i.e., the percentage of spectra of each taxon correctly classified, and the classifier precision (PREC) for each predicted class, i.e., the proportion of correct prediction for each predicted class.

Examining Table 2 and column TP in particular, we concluded that *A. vulgaris*, *C. sativa*, *C. sempervirens*, *C. arizonica*, and *F. ornus* spectra were correctly classified.

Table 2 Confusion matrix for classifier A

	AL	AR	B	CA	CO	A	S	D	F	OL	OS	TP
AL	4										1	0.8
AR		5										1.0
B	1		4									0.8
CA				5								1.0
CO		1			3						1	0.6
A						5						1.0
S							5					1.0
D					1			4				0.8
F									5			1.0
OL					1				1	2	1	0.4
OS	1										4	0.8
PREC	0.67	0.83	1.0	1.0	0.6	1.0	1.0	1.0	0.83	1.0	0.57	

Each row represents the spectra of an actual taxon (class), while each column represents the spectra in a predicted class. The last column gives the percentage of spectra of each taxon correctly classified (*TP* true positive values). The last row gives the proportion of correct prediction for each predicted class (*PREC* precision)

AL, *Alnus glutinosa*; *AR*, *Artemisia vulgaris*; *B*, *Betula pendula*; *CA*, *Castanea sativa*; *CO*, *Corylus avellana*; *A*, *Cupressus arizonica*; *S*, *Cupressus sempervirens*; *D*, *Dactylis glomerata*; *F*, *Fraxinus ornus*; *OL*, *Olea europaea*; *OS*, *Ostrya carpinifolia*

Also, for *A. glutinosa*, *D. glomerata*, *B. pendula*, and *O. carpinifolia*, classification results were good (TP=0.8), while for *C. avellana* and *O. europaea*, the TP values were unsatisfying (TP<0.8). These results in part resembled those of HCA, even if an improvement was clearly visible. However, the spectra variability, already discussed while analyzing the clustering results, clearly affected the classifier behavior, and the different degree of spectrum variability for different plant species caused the failure in classifying two out of 11 taxa. The variability effects were also confirmed by considering the PREC row of Table 2: *A. glutinosa*, *C. avellana*, and *O. carpinifolia* values were particularly low, and this indicated that at least some of the not repeatable measurements resulted distributed across the entire feature space.

In the study reported in [22], Raman microspectroscopy was carried out to acquire 91 spectra from individual pollen grains of 15 different plant species related at the genus level and family level. Unsupervised HCA allowed to discriminate between spectra of different pollen species using the complete Raman spectral signature, and particular attention was focused on the discussion of HCA results in the context of phylogenetic groups. Unfortunately, no supervised classification methods were applied. However, by comparing our results with those discussed in [22], a correspondence between the possibility of classifying pollen of a high number of plant species by FT-IR microspectroscopy and Raman microspectroscopy came out. This was not particularly surprising, as the two techniques provide complementary vibrational information on functional groups or bonds in the biochemical components of the analyzed samples. In the present work, the FT-IR spectra of different species belonged in some cases to the same plant family (but to different genera, except for *C. arizonica* Greene and *C. sempervirens* L.). In fact, as illustrated in Table 1, we considered 11 plant species that were members of seven different families. Therefore, this study allowed to demonstrate that a supervised classification distinguishing FT-IR spectra down to the species level (in different genera) was possible. This result is attested by the confusion matrix of Table 2, where the listed misclassifications were in most of the cases not ascribable to exchanges between species of the same family. In addition, all spectra of two species of *Cupressus* were correctly classified. On the other hand, by comparing the HCA in Fig. 4a with the corresponding result presented in [22], we concluded that, using the whole 1,800–850 cm^{-1} spectral range, the intraspecies variability for FT-IR spectra is more pronounced than for Raman spectra considered in the 1,700–380 cm^{-1} range.

Classifier B

In this case, the training set, which the *K*-NN classifier used to identify the single pollen grain spectra of dataset A, is given by dataset B, and a cross-validation approach is therefore not possible. The overall accuracy of classifier B was 69%, definitely lower than that of classifier A. Table 3 outlines the classifier performances by reporting the obtained true positive and precision values for each class. Examining Table 3, the unsatisfactory performances of classifier B were confirmed. To understand these results,

Table 3 True positive values and precision values (see Table 2) obtained for each plant species using classifier B

	AL	AR	B	CA	CO	A	S	D	F	OL	OS
TP	0.8	1.0	0.6	1.0	0.0	1.0	1.0	0.0	1.0	0.2	1.0
PREC	0.67	0.83	1.0	1.0	0.0	1.0	0.83	0.0	0.45	1.0	0.71

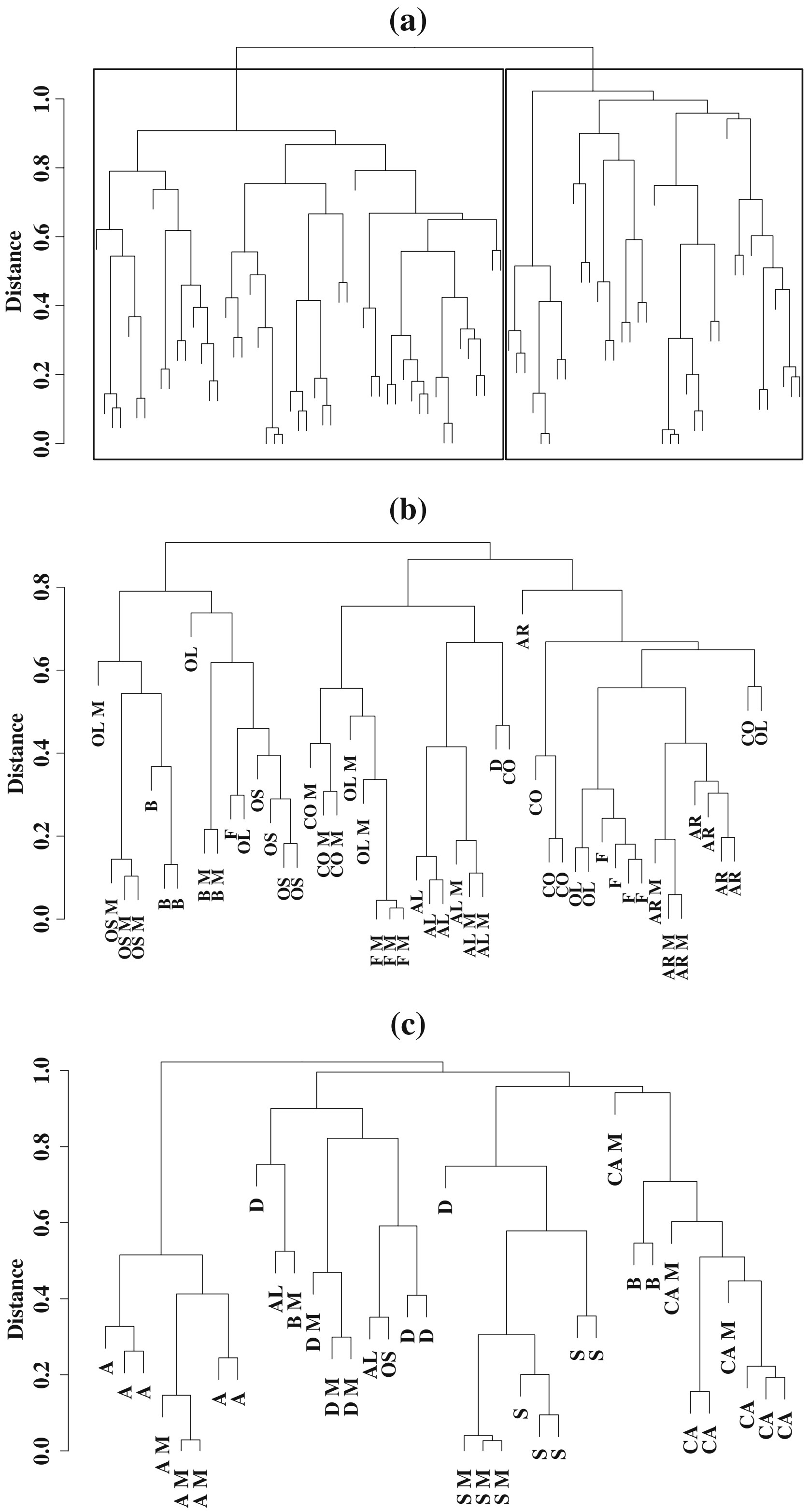

Fig. 5 Agglomerative hierarchical clustering obtained considering both single and multigrain FT-IR spectra of the 11 plant species studied. For readability reasons, **a** illustrates the complete dendrogram; **b**, **c** illustrate the two tree branches highlighted in **a** and obtained from cutting the dendrogram at height =1.1. Legend: see Fig. 3. The "-*M*" suffix means "multigrain"

Fig. 5a–c illustrates the hierarchical clustering simultaneously calculated on datasets A and B. It clearly appears that, for some of the plant species, the single and multigrain spectra are not necessarily close to each other in the feature space. As the *K*-NN classifier chose for each unknown spectrum the class of the single nearest neighbor spectrum ($K=1$), misclassification often occurred. Hence, even though the multigrain spectra, being intrinsically averaged, are probably more representative of the plant species, in order to face single-grain spectrum variability for classification purposes in monitoring stations, a spectrum library accounting for this variability (like dataset A) is surely better. On the other side, in this work, we did not contemplate the classification of multigrain spectra of every plant species, like those collected in dataset B, because we worked as if grains were collected in aerobiological monitoring stations.

Conclusions

This study was meant to verify both (1) the possibility of classifying pollen by FT-IR microspectroscopy and (2) the practical spin-off from this spectroscopic research. The *K*-NN classifier we built got an overall accuracy of 84%, and for nine out of the 11 considered plant species, the obtained accuracy was greater than or equal to 80%. In addition, the overall accuracy was greater than that (~80%) currently estimated for the morphological recognition approach used in our aerobiological monitoring network. Performance of the classifier was then at an adequate level. These results showed, to our knowledge for the first time, that spectra from single pollen grains obtained by FT-IR microspectroscopy can be successfully used to distinguish and classify different allergenic pollen taxa when compared with a library also composed by single grain spectra. Aim (1) of our study was therefore met. As regards aim (2), we were interested in verifying whether an automated classification system, alternative to the current morphological approach, could be really applied to the weekly monitoring performed in the aerobiological stations. The automation of the data analysis procedure (preprocessing and classification) that we obtained in this study was surely good (less of 1 min for obtaining the final results working with 100 spectra). On the other hand, this study showed that the performances of the *K*-NN classifier were not the same for all plant species because the spectra variability was different for different taxa. In particular, the classification results were unsatisfying (accuracy <80%) for two out of 11 taxa: *C. avellana* and *O. europaea.* This result was highlighted because we considered, for the first time in pollen discrimination studies, the combination of a high number of plant species, similar to that considered in the seasonal monitoring, with a cross-validation procedure, which allowed to obtain a more reliable estimation of the true error rate of the classifier. Even though a further assessment of classification accuracy on a larger, independent dataset should be considered, we can conclude that the obtained unbalanced performances of the classifier for two pollen taxa currently discourage the application of FT-IR microspectroscopy to airborne pollen monitoring. In fact, for these plant species, a punctual and correct identification could be compromised, hence preventing an effective risk communication to allergic sufferers.

In addition, although FT-IR measurements do not require prior complex sample preparation, we verified that the acquisition of a typical large number of FT-IR spectra, as required to redact the pollen bulletin, using standard laboratory equipment was a time-consuming step, which did not considerably shorten the time currently necessary for the morphological identification. In our opinion, this actually is the principal obstacle to the practical application of FT-IR microspectroscopy. To speed up the acquisition of FT-IR spectra from a high number of different pollen grains, we intend to explore the possibility of utilizing a 64×64 elements MCT photovoltaic Focal Plane Array detector in the 8×8 pixel binning configuration that would allow the parallel acquisition of 8×8 spectra, each one from a 25×25 μm^2 area.

On the other hand, the automatic identification of pollen grains via FT-IR microspectroscopy can give a significant contribution to other scientific fields, where the simultaneous identification of a large number of plant species is not necessary, such as for example in forensic science or in paleopalinology, or in any other application in which an objective approach for resolving doubts in pollen taxa identification is necessary.

Acknowledgments The authors would like to thank Antonella Cristofori and Fabiana Cristofolini for their precious contribution to selection and collection of pollen samples. This study has been granted by the FONDAZIONE CASSA DI RISPARMIO DI TRENTO E ROVERETO through the CARPOL biennial (2006–2008) project. The FT-IR instrument used in the present work was acquired by Dipartimento di Informatica-Università degli Studi di Verona with the financial support of Fondazione Cariverona.

References

1. Johansson SGO (2002) Clin Exp Allergy 2:2–7
2. Weber RW (1998) Ann Allergy 80:141–145
3. Chuine I, Belmonte J (2004) Grana 43:65–80
4. Faegri K, Iersen J (1989) Textbook of pollen analysis. John Wiley & Sons Ltd, London UK
5. Moore PD, Webb JA, Collinson ME (1991) Pollen analysis. Blackwell Scientific Publications, Oxford UK
6. Hirst JM (1952) Ann Appl Biol 39:257–265
7. Naumann D, Helm D, Labischinski H (1991) Nature 351:81–82

8. Nabet A, Pezolet M (1997) Appl Spectros 51:466–469
9. Petibois C, Gionnet K, Goncalves M, Perromat A, Moenner M, Deleris G (2006) Analyst 131:640–647
10. Orsini F, Ami D, Villa AM, Sala G, Bellotti MG, Doglia SM (2000) J Microbiol Meth 42:17–27
11. Mariey L, Signolle JP, Amiel C, Travert J (2001) Vib Spectros 26:151–159
12. Miguel Gomez MA, Bratos Perez MA, Martin Gil JF, Duenas Diez A, Martin Rodriguez JF, Gutierrez Rodriguez P, Orduna Domingo A, Rodriguez Torres A (2003) J Microbiol Meth 55:121–131
13. Pastuszka JS, Talik E, Hacura A, Sloka J, Wlazlo A (2005) Aerobiologia 21:181–192
14. Al-Holy MA, Lin M, Al-Qadiri H, Cavinato AG, Rasco BA (2006) J Rapid Meth Autom Microbiol 14:189–200
15. Dokken KM, Davis LC, Marinkovic NS (2005) Appl Spec Rev 40 (4):301–326
16. Mouille G, Robin S, Lecomte M, Pagant S, Höfte H (2003) Plant J 35:393–404
17. Laucks ML, Davis EJ (1998) J Aerosol Sci 29:s603–s604
18. Laucks ML, Roll G, Schweiger G, Davis EJ (2000) J Aerosol Sci 31:307–319
19. Pappas CS, Tarantilis PA, Polissiou MG, Harizanis PC (2003) Appl Spectros 57:23–27
20. Gottardini E, Rossi S, Cristofolini F, Benedetti L (2007) Aerobiologia 23:211–219
21. Ivleva NP, Niessner R, Panne U (2005) Anal Bioanal Chem 381:261–267
22. Schulte F, Lingott J, Panne U, Kneipp J (2008) Anal Chem 80 (24):9551–9556
23. R Development Core Team (2006) R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna AU (URL: http://www.R-project.org)
24. Savitzky A, Golay MJE (1964) Anal. Chem 36:1627–1623
25. Gordon AD (1999) Classification (Second edition). Chapman and Hall/CRC Press, London UK
26. Weiss SM, Kulikowski CA (1991) Computer systems that learn. Morgan Kaufmann Publishers, San Mateo CA
27. Ripley BD (1996) Pattern recognition and neural networks. Cambridge University Press, Cambridge UK