Journal of Biomedical Optics

SPIEDigitalLibrary.org/jbo

Identification of fungal phytopathogens using Fourier transform infrared-attenuated total reflection spectroscopy and advanced statistical methods

Ahmad Salman
Itshak Lapidot
Ami Pomerantz
Leah Tsror
Elad Shufan
Raymond Moreh
Shaul Mordechai
Mahmoud Huleihel

SPIE

Identification of fungal phytopathogens using Fourier transform infrared-attenuated total reflection spectroscopy and advanced statistical methods

Ahmad Salman,[a] Itshak Lapidot,[b] Ami Pomerantz,[c] Leah Tsror,[d] Elad Shufan,[a] Raymond Moreh,[e] Shaul Mordechai,[e] and Mahmoud Huleihel[c]

[a]SCE-Sami Shamoon College of Engineering, Department of Physics, Beer-Sheva 84100, Israel
[b]SCE-Sami Shamoon College of Engineering, Department of Electrical and Electronics Engineering, Beer-Sheva 84100, Israel
[c]Ben-Gurion University of the Negev, Department of Virology and Developmental Genetics, Faculty of Health Sciences, Beer-Sheva 84105, Israel
[d]Institute of Plant Protection, Department of Plant Pathology, Agricultural Research Organization, Gilat Experiment Station, M.P. Negev, 85250, Israel
[e]Ben-Gurion University, Department of Physics, Beer-Sheva 84105, Israel

Abstract. The early diagnosis of phytopathogens is of a great importance; it could save large economical losses due to crops damaged by fungal diseases, and prevent unnecessary soil fumigation or the use of fungicides and bactericides and thus prevent considerable environmental pollution. In this study, 18 isolates of three different fungi genera were investigated; six isolates of *Colletotrichum coccodes*, six isolates of *Verticillium dahliae* and six isolates of *Fusarium oxysporum*. Our main goal was to differentiate these fungi samples on the level of isolates, based on their infrared absorption spectra obtained using the Fourier transform infrared-attenuated total reflection (FTIR-ATR) sampling technique. Advanced statistical and mathematical methods: principal component analysis (PCA), linear discriminant analysis (LDA), and *k*-means were applied to the spectra after manipulation. Our results showed significant spectral differences between the various fungi genera examined. The use of *k*-means enabled classification between the genera with a 94.5% accuracy, whereas the use of PCA [3 principal components (PCs)] and LDA has achieved a 99.7% success rate. However, on the level of isolates, the best differentiation results were obtained using PCA (9 PCs) and LDA for the lower wavenumber region (800–1775 cm^{-1}), with identification success rates of 87%, 85.5%, and 94.5% for *Colletotrichum, Fusarium*, and *Verticillium* strains, respectively. [DOI: 10.1117/1.JBO.17.1.017002]

Keywords: Fourier transform infrared-attenuated total reflection; fusarium oxysporum; colletotrichum coccodes; verticillium dahliae; fungal detection; principal component analysis; linear discriminant analysis.

Paper 11504 received Sep. 13, 2011; revised manuscript received Nov. 2, 2011; accepted for publication Nov. 4, 2011; published online Feb. 6, 2012.

1 Introduction

The fungi investigated in this study belong to the same phylum (divisions) named "Ascomycota," a group whose members are also known as Sac fungi, and have evolved from one common ancestor. Living organisms are biologically classified into different units according to their similarity. A species is often defined as a group of organisms capable of interbreeding and producing fertile offspring. More precise measures are often used, such as similarity of DNA, morphology, or ecological niche. Species that are believed to have the same ancestors are grouped together, and this group is called a genus. Each species may include different isolates (strains) which usually result from one or more mutations. In the hierarchy of the binomial classification system, species is above isolate and below genus.

As far as the genus is concerned, these fungi are grouped into three classes: *Fusarium*, *Colletotrichum*, and *Verticillium*. These fungi are pathogens which attack crops, resulting in fungal diseases, which lead to large economic losses.[1,2] For example, *V. ahlia* causes a wilt disease in hundreds of species of eudicot plants. Many economically important plants are susceptible, including cotton, tomatoes, potatoes, eggplants, and peppers. Solanaceous crop may be infected at any age by fungi causing the *Fusarium* wilt and *Verticillium* wilt, with similar symptoms. The diseases therefore cannot be distinguished based on symptoms alone.

Figures 1(a)–1(c)[3,4] show plants infected by the various fungi genera studied here. Figure 1(a) shows potato tubers infected with *F. oxysporum* which causes a wilt symptom characterized by browning of the vascular ring and the stem end. In Fig. 1(b) *Verticillium* wilt symptoms on tomato leaves can be seen. Infected plants usually survive, but both the yields and the fruits may become small, depending on the severity of attack. Figure 1(c) shows a tomato infected with *C. coccodes* which usually causes serious damage to the fruit.

Early detection of phytopathogens is very important for both successful protection and effective treatment,[5] thus enhancing the chances of recovery and simultaneously preventing environmental pollution. By standard physiological methods, it is very difficult to differentiate between closely related species and strains[6,7] based on cultivation. In addition, fungi identification by visual and microscopic observations are usually time consuming and not always very specific.[8]

Address all correspondence to: Ahmad Salman, Department of Physics, SCE-Sami Shamoon College of Engineering, Beer-Sheva 84100, Israel; E-mail: ahmad@sce.ac.il, and Mahmoud Huleihel, Ben-Gurion University of the Negev, Department of Virology and Developmental Genetics, Faculty of Health Sciences, Beer-Sheva 84105, Israel; E-mail: mahmoudh@ bgu.ac.il.

0091-3286/2012/$25.00

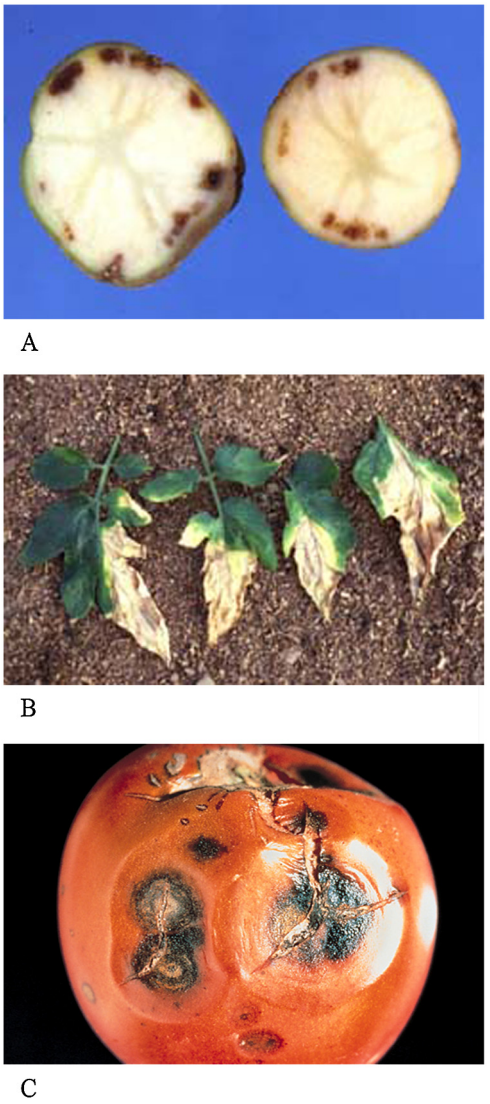

Fig. 1 Different fungi genera affecting crops. (a) Potato tuber discoloration of the vascular ring, caused by the *Fusarium* wilt, (b) *v*-shaped lesions on tomato leaves, associated with *Verticillium* wilt, and (c) *C. coccodes* causes anthracnose on tomatoes and black dot disease on potatoes.

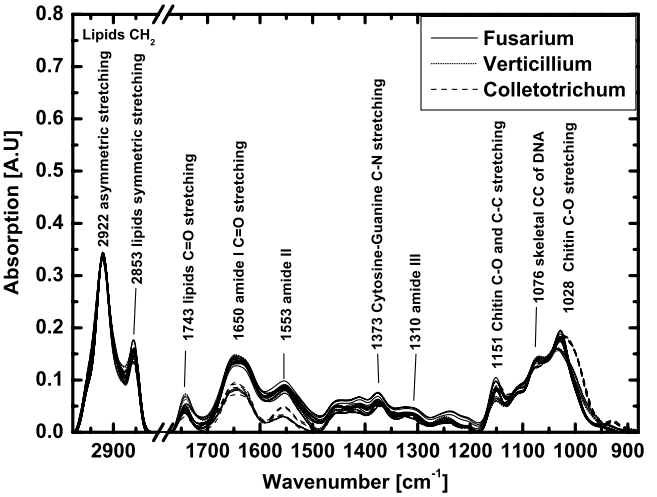

Fig. 2 Infrared absorption spectra of *C. coccodes*, *V. dahliae*, and *F. oxysporum* in the ranges 890 to 1770 cm^{-1} and 2800 to 2990 cm^{-1}.

Serological, molecular biology methods are based on differences in the nucleic acid genome of the tested fungi, whereas the immunological method is based on differences in the proteins constructing the fungi and on the availability of specific monoclonal antibodies against the tested fungi. These methods are very sensitive and rapid relative to physiological methods for the identification of pathogens; however, their use is still limited due to their availability for only a small number of fungal pathogens.[9–11] Developing such methods for different fungi strains is complicated, costly, and not always possible. Therefore, the use of these methods is expected to be limited and not available for the screening of large numbers of samples of different fungal strains.

Many studies have indicated the potential of FTIR spectroscopy methods for the detection and identification of microorganisms, especially in food products.[12,13] The vast information already achieved about spectral bands obtained from FTIR spectra of living cells[14] combined with features of infrared spectroscopy such as sensitivity, rapidity, low expense, and simplicity[15] enabled us to differentiate the fungi even on the levels of species[16,17] and strains.[18–23] The use of Fourier transform infrared-attenuated total reflection (FTIR-ATR) spectroscopy followed by principal component analysis (PCA) and linear discriminant analysis (LDA) for biological classification has gained momentum in the last decades (see Martin et al.).[24–26]

Linker and Tsror[19] showed good results in differentiating between fungi genera and species using FTIR-ATR and applying statistical analysis techniques: PCA, and cluster and canonical variate analysis (CVA). In a different study, A. Naumann[18] classified 26 fungi strains belonging to 24 different species using FTIR-ATR and applying cluster and artificial neural network (ANN) statistical techniques. Fischer et al.[22] tried to develop a method to reproducibly differentiate *Aspergillus* and *Penicillium* species on the generic, species, and strain levels. In their work they used nine different species and only one strain from each species. Shapaval et al.[23] showed that it was possible to differentiate 11 species of five different fungal genera using FTIR spectroscopy. Our former study[21] took the method further into a more challenging stage; by applying the PCA and LDA techniques, the ability of the FTIR-ATR method to classify six different strains of *F. Oxysporum* was examined. The results showed that it is possible to classify and differentiate between the strains with a 81.4% success rate.

The main goal of this study was to test the feasibility of the FTIR-ATR methodology in differentiating 18 isolates from three different fungi genera: *Colletotrichum*, *Verticillium*, and *Fusarium*. From each of these genera, we selected six isolates originating in a single species, namely six isolates of *C. coccodes*, six isolates of *V. dahliae* and six isolates of *F. oxysporum*. The classification procedure was done in two phases. In the first phase, the fungal samples were classified on the genus level, and in the second phase, the samples were classified on the isolates level. The classification procedures were based on infrared absorption spectra of the samples, and used advanced mathematical and statistical techniques; PCA followed by *k*-means and LDA.

The uniqueness of our study lies in the analysis of a larger number of isolates belonging to the same species. Isolates are very similar; their spectra are blended and overlap, which makes it difficult to differentiate between them, especially relative to the species and genera levels as previously mentioned. Moreover, for pattern recognition methods, increasing the

class numbers is a real challenge, especially when the classes are strains of the same species as in this study.

2 Materials and Methods

2.1 *Fungi*

Eighteen isolates of three different fungi genera, namely *C. coccodes*, *V. dahliae*, and *F. oxysporum*, were examined, six isolates of each. The samples were obtained from the Department of Plant Pathology at the Gilat Experiment Station, ARO, Israel.

Fungi samples were isolated from different infected crops obtained directly from the plant stem, root, or tuber. Samples were scratched from the infected areas of the crops, mixed well in 1 ml of a potato dextrose medium and cultivated for several days on potato dextrose agar (PDA, Difco Laboratories, Becton, Dickison and Company, Sparks, MD) at 27°C. Several cultures (5 to 10) of single fungi colonies, obtained by micromanipulation, of each strain were cultivated in different batches. These cultures were grown for 3 to 10 days at 27°C in continuous shaking. Samples of the growing fungi were identified by visual and microscopic observations.

As a first step, the cultivated fungi strains were identified using classical microbiological techniques.[27,28] Next, samples of the fungi were separated and purified by centrifuging about 1.5 mL of the mixture at 13,200 rpm for 4 min, rinsing the extract 4 times with distilled water, and suspending the pellet in an appropriate volume of distilled water (~1 mL) for spectroscopic measurements.

Pure samples enable us to control as many experimental parameters (such as growth conditions, amount of sample examined, duration of growth, etc.) as possible, and to verify that each absorption band in the IR spectrum was due to the specific sample.

2.2 *Sample Preparation*

Due to the complicated structure of the fungi, it is difficult to achieve a homogeneous suspension of the fungi in water or spread them on the ZnSe crystal surface of the attenuated total reflection (ATR) accessory. Therefore, the fungi were mashed into small pieces and mixed as evenly as possible into the distilled water to obtain a suspension. About 500 μL of each fungal suspension sample was spread as homogenously as possible on the surface of the ATR ZnSe crystal to cover the entire crystal surface. These samples were air dried for about 30 min until all water had evaporated, and then measured by ATR spectroscopy. The ATR crystal, of a trapezoid shape 80 mm long, 10 mm wide, and 4 mm thick was obtained from PIKE technologies.

2.3 *FTIR Measurements*

We used a Bruker Tensor 27 spectrometer in the ATR mode with DTGS detector for our measurements. After drying them, the samples were scanned 128 times in the range of 600 to 4000 cm^{-1}, with a 4 cm^{-1} spectral resolution. OPUS software was used for spectral manipulation such as baseline correction, bisecting, and normalization. Our measurements were carried out over several weeks.

2.4 *Spectral Analysis*

Two-hundred fifty six spectra were measured from six different *C. coccodes* isolates, 131 spectra were measured from six different *F. oxysporum* isolates, and 105 spectra were measured from six different *V. dahliae* isolates. All spectra were corrected with OPUS software according to the wavelength dependence of the penetration depth. The spectra were then bisected into two regions (800–775 cm^{-1}) and (2800–2990 cm^{-1}) to exclude the water absorption band and the "dead" region between amide I and the lipid bands. The spectra in each region were baseline corrected using the rubber band method, normalized separately using the vector normalization method, and then offset corrected using OPUS software.

2.5 Statistical Analysis

2.5.1 *PCA*

PCA is an unsupervised[29,30] multivariate analysis tool for dimensionality reduction[31,32] which is widely used in pattern recognition. The common assumption is that the most separable dimensions are those with the highest variance; it frequently is so, but not necessarily.

Here, $X = \{x_1, \ldots, x_N\}$ is the data set, where $x_n \in \Re^{d\times 1}$, and Σ is the covariance matrix of X. Be $\Lambda = [\lambda_1 \quad \cdots \quad \lambda_N]$ the vector of eigenvalues, so that $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_N$ with corresponding eigenvectors matrix: $U = [u_1 \quad \cdots \quad u_d]$; $u_i \in \Re^{d\times 1}$. Then, the data after PCA is:

$$Y = U_q^T X = [u_1 \quad \cdots \quad u_q]^T X; \qquad q \leq d,$$

where T is the transform operator.

Basically, PCA is a mathematical operator that projects the high dimensional data onto a subspace of low dimension which captures the orthogonal directions with the highest variability, i.e., instead of using many variables the variability in the data is described using only a few principal components (PCs).[33]

The first projection coefficient is called the first principal component (PC1). It contains most of the variance. PC2 contains most of the residual variance and is perpendicular to the first. The other PCs obey the same rules. This method allows the reduction of our spectra to three variables in the first phase of classification (genus level) and nine variables in the second phase of classification (isolate level), in the lower wavenumber region that accounts for almost 100% of the variance.[34]

2.5.2 *LDA*

Following PCA, we applied the LDA,[35] where the separation is based on the assumption that different classes sharing the same covariance matrix have different mean vectors for classification. The separation procedure is done by maximizing the probability of a class under these assumptions. The formulation of LDA is as follows[36]: given data which was derived from k classes, with a Gaussian probability density function (PDF) for each class:

$$f_k(x) = \frac{1}{(2\pi)^{d/2}|\Sigma_k|^{1/2}} \exp\left\{\frac{1}{2}(x-\mu_k)^T\Sigma_k^{-1}(x-\mu_k)\right\}, \tag{1}$$

where μ_k is the mean vector and the covariance matrix Σ_k. Assuming that each class's prior probability is π_k, and all the

classes have the same covariance matrix, $\Sigma_k = \Sigma$. The classification of the new input data $x \in \Re^{d\times 1}$ is performed as follows:

$$\hat{k} = \underset{k\in\{1,\ldots,K\}}{\arg\max} \left\{ x^T \Sigma^{-1} \mu_k - \frac{1}{2} \mu_k^T \Sigma^{-1} \mu_k + \log(\pi_k) \right\}. \quad (2)$$

Training and test sets were selected randomly from the database.

The examination of the results performed using two variants of k-fold cross-validation is applied frequently in pattern recognition. The first run was 5-fold, i.e., 20% to 80% when 80% of the data was used for training and 20% for testing. Each time, another 20% (1 of 5) was used for testing and the rest (4 parts) for training. We ran this approach 20 times, each time with a random partition of the data into 5 groups. The other variant "leave-one-out,"[31,32] is when $k = N$, where N is the number of data points. It is usually applied when the amount of data is relatively small.

2.5.3 *K-Means*

K-means is widely used for clustering and vector quantization.[37] The aim is to partition the space into K cells, each represented by a vector, named a centroid or code-word. All code-words compose the codebook of the quantization process $\{c_k\}_{k=1}^K$. The goal is to minimize some predefined average distance:

$$\{c_k^*\}_{k=1}^K = \min_{\{c_k\}_{k=1}^K} \left\{ \frac{1}{N} \sum_{n=1}^{N} D(c_{l_n}, x_n) \right\} \qquad l_n = \underset{l\in 1,\ldots,K}{\arg\min}\, D(c_l, x_n) \quad (3)$$

The most common distance is the square Euclidian distance, $D(x,y) = (x-y)^T(x-y)$. The partitioning was performed using an iterative algorithm which converges to a local minimum.

3 Results

Figure 2 shows the infrared absorption spectra of the three fungi genera investigated in this study, i.e., *Colletotrichum*, *Verticillium*, and *Fusarium*, each with a different color. The spectra for each fungus were measured from six different isolates. The main spectral features in the high wavenumber region (2700–2950 cm^{-1}) are the bands detected at 2853 and 2922 cm^{-1}. These bands derive mainly from phospholipid absorbance.[38] Water absorbance bands in this region were excluded from the spectra as part of the analysis procedure. The main features in the low wavenumber region (800–1775 cm^{-1}), were the amide I and amide II with centroids at 1650 cm^{-1} and 1553 cm^{-1}, respectively. There is a clear peak which arises from lipids absorbance[39] at 1743 cm^{-1}. A large absorbance band at 1076 cm^{-1} was mainly attributed to carbohydrate and nucleic acid vibrations. The centroid of the amid III band was detected at 1252 cm^{-1}. The chitin band, which is specific to fungi, was detected at 1151 cm^{-1} and 1078 cm^{-1} due to its C−O and C−C stretching vibrations, and the glycogen C−O stretching vibration was detected at 1024 cm^{-1}.

As previously mentioned, there are clear differences between the fungi mainly in the lower wavenumber region, namely, in the amide I and II, chitin, and glycogen peaks. These differences enabled us to differentiate the three genera using unsupervised methods, such as k-means, achieving an accuracy of 94.5%. As shown in Fig. 2, the *Colletotrichum* fungus has low intensities in the amide I and II regions which may indicate lower protein content relative to *Fusarim* and *Verticillium* fungi.

Utilizing the PCA and LDA (20 to 80) methods, we classified the sample into three groups with great success. When using one PC we obtained an accuracy of 61.2%, using two PCs a 96.9%

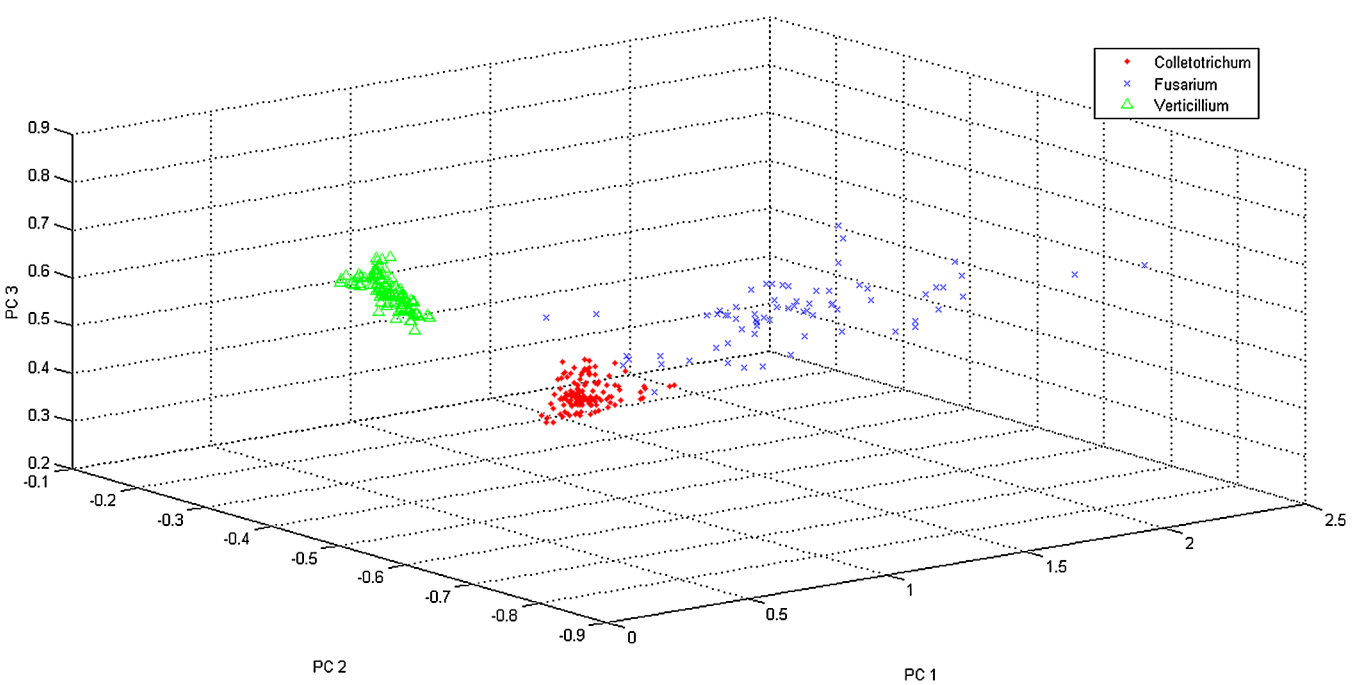

Fig. 3 Complete date sets of *Colletotrichum, Verticillium,* and *Fusarium* presented in a 3-D PCs domain. This figure summarizes the first phase (genus) classification results.

accuracy was achieved, which increased to 99.7% with three PCs.

Figure 3 shows the results for the three fungal groups obtained by LDA based on three PCs of PCA at the lower wavenumber region.

In this figure each spectrum was calculated as a superposition of three loadings (PCs) derived using the PCA algorithm. Each spectrum was identified by three numbers, which are the coefficients of the loadings. As shown in Fig. 3 the data is divided into three groups, red squares (*Colletotrichum*), blue crosses (*Fusarium*), and green open triangles (*Verticillium*). There is an excellent separation between the three groups. To find out which feature contributed to the classification procedure, we plotted two-dimensional figures for PC1 versus PC2, depicted in Fig. 4(a), which provided the best classification between the three studied genera. The projection of the data on the PC1 axis provided a good classification of *Fusarium*. The projection of the *Colletotrichum* and *Verticillium* data on the PC2 axis provided a complete classification between them. The first two PCs are plotted in Fig. 4(b) where the main peaks are labeled. These

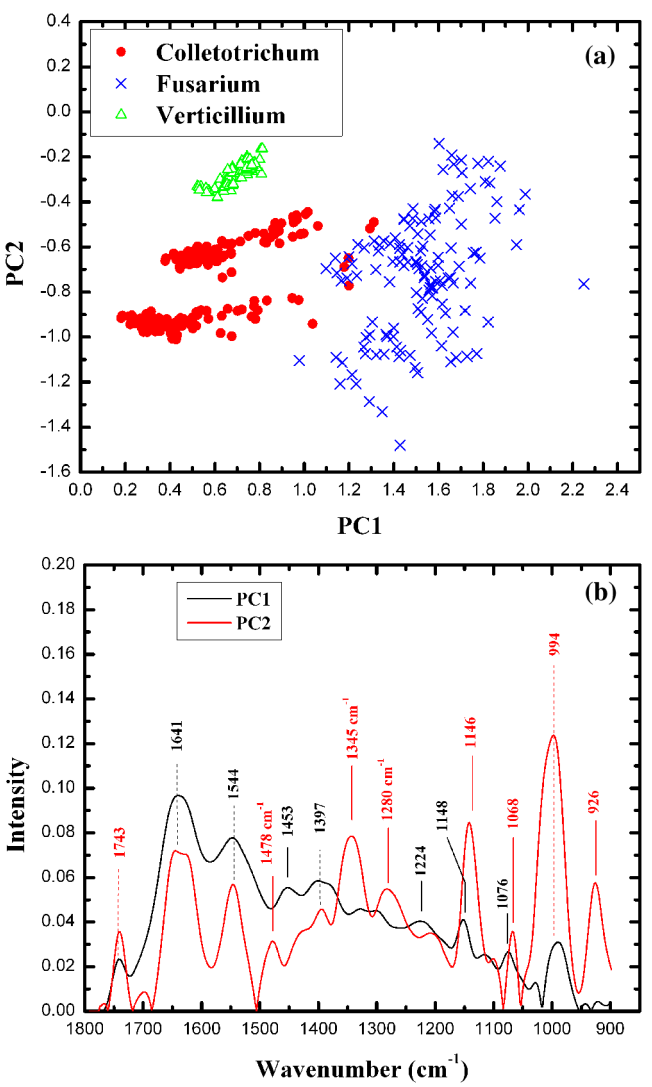

Fig. 4 (a) Two-dimensional plot of *Colletotrichum*, *Verticillium*, and *Fusarium* included in this study. (b) This figure presents the first 2 PCs, derived from PCA, in the lower range. The main peaks of each PC are labeled.

are: for PC1, the C = O stretching mode of lipids (1743 cm^{-1}), the amide I and II bands (1641 cm^{-1} and 1544 cm^{-1}, respectively), the asymmetric and symmetric CH_3 bending modes of the protein methyl groups[40] (1453 cm^{-1} and 1397 cm^{-1}), collagen, the asymmetric stretching of phosphate groups of phosphodiester linkages in DNA and RNA[41] (1224 cm^{-1}), carbohydrates[42] (1148 cm^{-1}), the symmetric phosphate PO_2^- stretching[43] (1076 cm^{-1}), and OCH_3 [polysaccharides-cellulose[44] (989 cm^{-1})]. All the peaks labeled with dashed lines are also common for PC2. The other main peaks of PC2 are: C–N stretching and C–N–H bending of N–H[45] (1478 cm^{-1}), collagen[46] (1345 cm^{-1}), amide III, collagen[47] (1280 cm^{-1}), phosphate, oligosaccharides,[48] and carbohydrates[42] (1146 cm^{-1}), stretching C–O ribose[49] (1068 cm^{-1}), C–C [Ref. 49(994 cm^{-1})], and the phosphodiester stretching bands region [for absorbance due to collagen and glycogen[50] (926 cm^{-1})].

In the second phase we tried to classify the six different isolates within each group (genus). In contrast to the genus level where there were clear differences between the spectra, here the spectra of the different strains were blended and overlap [Figs. 5(a)–5(c)], and in each spectrum all major absorption bands described earlier were detected. However, some differences were found, such as different absorbance intensities in different absorbance bands in the two regions.

The analysis was performed for different regions of the spectrum, and the best results were achieved when the lower wavenumber region (800–1775 cm^{-1}) was used. The *k*-means

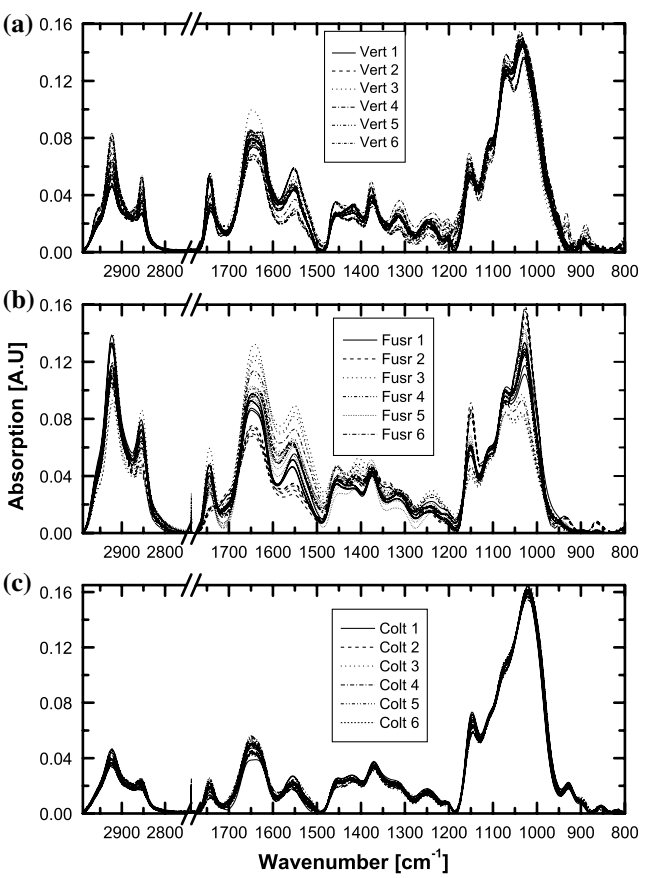

Fig. 5 Infrared absorption spectra of six isolates of *C. coccodes* (a), six isolates of *V. dahliae* (b), and six isolates of *F. oxysporum* (c) in the ranges 800 to 1770 cm^{-1} and 2800 to 2990 cm^{-1}.

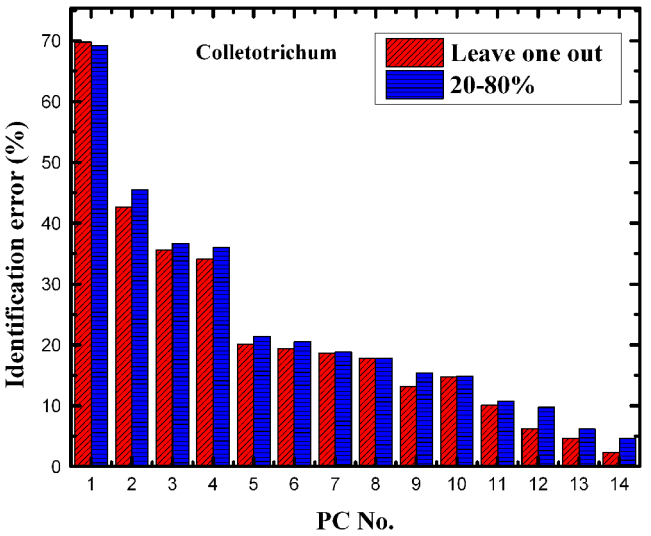

Fig. 6 Identification error vs. PC No. in the lower wavenumber region.

achieved poor results, thus we adopted the PCA and LDA methods. In the second phase where the number of measurements for each strain was low, the "leave-one-out" method was preferred, whereas in the first phase where the statistics were good the 20 to 80 algorithm was used.

Figure 6 shows the identification errors (percentage) derived using LDA together with the "leave-one-out" and 20 to 80 algorithms, as a function of the PC number used in the analysis of *Colletotrichum*. Due to the similarity between the strains of each of the three groups, more PCs had to be used than in the first (genus) phase.[21] Nine PCs were used in the strain phase, which achieved a good classification and simultaneously kept the highest loading (PC) meaningful and noiseless.

The LDA results using the "leave-one-out" algorithm and nine PCs showed that it was possible to differentiate between the *Colletotrichum* isolates with a 86.8% success rate, whereas the classification success rates of the *Fusarium* and *Verticillium* strains were 85.5% and 94.6%, respectively. The LDA results using the 20 to 80 algorithm and 9 PCs showed that it was possible to differentiate between the *Colletotrichum* isolates with a 84.6% success rate, whereas the classification success rates of the *Fusarium* and *Verticillium* strains were 80.7% and 94.3%, respectively.

Table 1 shows the results of the classification of the studied genera on the isolate level using the "leave-one-out" method.

4 Discussion

In principle, the ATR sampling technique is similar to measurements done using remote fiberoptic probes. Thus, evaluation of the ATR potential for fungal classification on large numbers of strains on the levels of genus, species, and isolates is an important step toward *in vivo* measurements using infrared fibers.

Many studies have shown that it is possible to use FTIR-ATR for detection and identification of fungi.[18,19,21] In their study, Linker et al.[19] managed to classify different fungi genera species and to differentiate between two isolates of *Colletotrichum* based on the measured FTIR-ATR spectra and advanced statistical methods: CVA and PCA. Based on the ATR sampling technique, Naumann[18] succeeded to differentiate between 26 fungal isolates that belong to 24 different species using cluster analysis and ANN analysis.[18]

In their work, Fischer et al.[22] used nine different species and only one strain from each species. Shapaval et al.[23] used 11 different species of five different fungal genera. In our previous work, we used the FTIR-ATR to differentiate between six different isolates of *F. oxysporum*.[21]

By enlarging the number of strains, we come closer to reality, where tens of isolates from each species exist. All previous studies focused on just a few species belonging to the same genus, and a few strains belonging to the same species. Thus, success in the classification of large numbers of isolates and species provides a solid base for the future large scale *in vivo* examination of phytofungal pathogens using infrared spectroscopy.

In this study we investigated six different isolates from each of three fungi genera *Colletotrichum*, *Verticillium* and *Fusarium*. Our main objective was to test and evaluate the potential of FTIR-ATR spectroscopy to differentiate between these isolates. This was done in two phases of examinations of fungi samples, the first on the genus level and the second on the isolate level. The fungal isolates to be tested were collected, isolated, and purified to obtain pure samples of the desired strain.

Our analysis was based on the fungal IR evanescent wave absorption spectra. We used a statistical approach to analyze the large database of the obtained IR spectra.

Using the currently practiced classic physiological methods, it is very difficult to accurately differentiate between the strains because of their high morphological similarity. Molecular and serological methods are also not available for all these strains.

Due to the variance in the penetration depth occurring in the ATR mode measurements, absorption band intensities and positions are different than those obtained using the transmission mode.[51–53] Thus, all our spectra have undergone the same penetration depth correction using OPUS software. On the genus level, there were clear differences that have enabled the differentiation between the samples by unsupervised clustering analysis using *K*-means, with a 94.5% success rate. In-house developed software codes based on MATLAB software[54] were used. After PCA calculation, a supervised classifier LDA was applied, which enabled the classification of the data with an accuracy of 99.7%. In this phase of the work (genus level), 3 PCs were sufficient to obtain such high accuracy, whereas in the second phase (isolate level) nine PCs were required. PCA enabled capturing the variability of the fungi using a small number of PCs. In the first phase the examined fungi genera were successfully differentiated with a high accuracy using two PCs, and using three PCs obtained almost a 100% success rate. The second phase, in which the samples were isolates of the same species, the differences between their spectra were minor, and therefore a good differentiation was achieved using a larger number of PCs. It is not guaranteed that the first PCs are the most important for separability; however, they carry most of the variance of the data.

The dependence on the PC number is task-dependent, i.e., number of classes to classify, level of classification (genus, species, or isolates), and also the type of samples.

From Fig. 4(a) we can see that the new directions defined by PC1 and PC2 give a good classification. Using PCA calculation, we obtained a new basis which defines new directions.[55] Figure 4(b) shows PC1 and PC2 vectors in the standard basis. The peaks which are found in PC2 and not in PC1 (labeled by dashed lines) are suspected to be more dominant with respect to the classification. The comparison procedure of the peaks is

Table 1 Successful identification of (a) *Colletotrichum* (Coll), (b) *Fusarium* (Fus) and (c) *Verticillium* (Vert) isolates obtained using LDA calculations and the "leave-one-out" algorithm in the low wavenumber region (900–1775 cm^{-1}).

		Coll 1	Coll 2	Coll 3	Coll 4	Coll 4	Coll 5
(a)	Coll 1	17	0	4	0	0	0
	Coll 2	0	15	1	0	2	0
	Coll 3	0	1	18	1	0	0
	Coll 4	0	0	5	20	0	1
	Coll 5	0	0	0	0	19	1
	Coll 6	0	0	1	0	0	23
		Fus 1	Fus 2	Fus 3	Fus 4	Fus 4	Fus 5
(b)	Fus1	8	0	0	1	1	0
	Fus 2	0	7	1	1	1	0
	Fus 3	0	0	10	0	1	0
	Fus 4	0	2	0	9	1	0
	Fus 5	0	0	0	1	13	0
	Fus 6	0	0	0	0	0	12
		Vert 1	Vert 2	Vert 3	Vert 4	Vert 4	Vert 5
(c)	Vert 1	13	1	0	0	0	0
	Vert 2	1	19	0	0	0	0
	Vert 3	0	0	20	0	0	0
	Vert 4	1	0	0	10	0	0
	Vert 5	0	3	0	0	9	0
	Vert 6	0	0	0	0	0	28

relatively easy when a small number (2 to 3) of the PCs have to be compared.

Smaller numbers of PCs make it easier to relate the changes of the PCs shape to biology, whereas this is much harder when a larger number of PCs is used. In this study, the goal was classification, and thus we should choose the number of PCs which provided the best classification providing that the highest PC was still meaningful for the biological system investigated. This choice of PC number is still not objective in our field of spectroscopy and should be investigated more carefully.

As shown in Fig. 4(a), a good classification of *Fusarium* was achieved. This is probably due to some peaks in PC1 which are absent in PC2 [Fig. 4(b)]. These peaks which contribute mainly to the classification of *Fusarium* are 1453 cm^{-1} (methyl groups of proteins),[40] 1224 cm^{-1} (collagen, asymmetric, stretching of phosphate groups of phosphodiester linkages in DNA and RNA),[41] 1148 cm^{-1} (chitin, carbohydrates)[42] and 1076 cm^{-1} (chitin, symmetric phosphate PO_2^- stretching).[43]

PC2 gives a good classification between *Colletotrichum* and *Verticillium*. It can be seen in Fig. 4(b) that some peaks in PC2 do not exist in PC1, and this is the main contribution to the classification between *Colletotrichum* and *Verticillium*. These peaks are 1478 cm^{-1} (C–N stretching and C–N–H bending N–H),[45] 1345 cm^{-1} (collagen),[46] 1280 cm^{-1} (amide III, collagen),[47] 1146 cm^{-1} (phosphate, oligosaccharides,[48] and carbohydrates)[42] and 1068 cm^{-1} (C–O ribose),[49] 994 cm^{-1} (C–O ribose, C–C)[49] and 926 cm^{-1} [phosphodiester stretching bands region (for absorbance due to collagen and glycogen)].[50]

The LDA calculation was designed applying the two methods ("leave-one-out" and 20 to 80), using different sets for prediction each time, so that the test sets would be statistically independent. The "leave-one-out" method is a common method of cross-validation, extensively explored in machine learning, used to estimate the error in a small sized populations. Using it ensures the validation of results.[56,57]

The 20 to 80 method is used when large numbers of data exist, thus it was used in the first phase of the study. In the second phase of identification the "leave-one-out" was preferred because it achieved better results, but the 20 to 80 was used as well to validate the "leave-one-out" algorithm (Fig. 6).

Our results with a successful classification of the different strains of each genus using the "leave-one-out" method, are presented in Table 1(a)–(c).

In the second phase, nine PCs were used, thus the discrimination power is the overall contribution of the whole set of PCs. It is difficult to outline which PC contributes more to the discrimination between the different isolates (strains).

It is very difficult to discuss the relevance of spectral differences of the various species or isolates, due to the high similarity in the molecular and structural composition of these related isolates, which belong to the same species. According to the FTIR-ATR spectra, the differences are spread over the entire spectrum and not specific to a certain band. One limitation of FTIR is that the spectra obtained from biological samples are complex, and difficult to interpret because it is hard to know which bands arise from which biomolecule. For example, changes in the membrane are reflected in the absorption bands of lipids; however, spectral differences in the lipid absorption bands may not arise solely from membranes, but may be attributed to lipids that are not located in the cell membrane.

The applied method is objective and computerized. Enlarging the database and the strain number will improve the statistics and bring the method closer to reality where large numbers of strains exist. The method shows a great potential for the identification and study of phytofungal pathogens, as far as the level of strain identification.

This method may be useful for studying biological aspects of the different genera as shown in Fig. 2. It shows significantly lower intensities in the amide I and II regions for *Colletotrichum* compared with *Fusarim* and *Verticillium* fungi. These differences may reflect different protein contents between these genera. Certainly, small differences in their structural components (such as lipids, proteins, and sugars) do exist, but unfortunately the origins of these differences are yet unknown and it is very difficult to define the exact biological origin of the observed spectral differences. However, these difficulties do not affect the main objective of this study, which was to examine the potential of spectroscopic techniques for reliable detection and discrimination between different fungi strains.

5 Conclusions

FTIR-ATR spectroscopy in tandem with PCA methods, followed by LDA calculations, enabled the differentiation between the fungal samples studied not only on the genus level but on the level of strains as well (where spectral differences are minute) with a good confidence level.

Enlarging the database to include more species and strains, could improve the statistics and bring the method one step forward toward real conditions, where tens of isolates from the same species exist.

Acknowledgment

Financial support by SCE internal research funding is gratefully acknowledged.

References

1. G. N. Agrios, *Plant Pathology*, Academic Press, New York (1978).
2. J. Katan, "Principles in plant pathology," in *Plant Diseases in Israel*, J. Rotem, J. Palti, and Y. Ben-Yephet, Eds., Volcani Center, Bet-Dagan, Israel (1998).
3. http://www.maine.gov/agriculture/pesticides/gotpests/diseases/tomato-problems.htm.
4. http://ohioline.osu.edu/hyg-fact/3000/3122.html.
5. G. V. Doern et al., "Clinical impact of rapid in vitro susceptibility testing and bacterial identification," *J. Clin. Microbiol.* **32**(7), 1757–1762 (1994).
6. G. H. Kim et al., "Ophiostomatoid fungi isolated from Pinus radiata logs imported from New Zealand to Korea," *Can. J. Bot.* **83**(3), 272–278 (2005).
7. U. Moreth and O. Schmidt, "Investigations on ribosomal DNA of indoor wood decay fungi for their characterization and identification," *Holzforschung* **59**(1), 90–93 (2005).
8. G. L. Schumann and C. J. DArcy, *Essential Plant Pathology*, The American Phytopathological Society, St. Paul, MI (2006).
9. S. Nikkari and D. A. Relman, " Molecular approaches for identification of infectious agents in Wegener's granulomatosis and other vasculitides," *Curr. Opin. Rheumatol.* **11**(1), 11–16 (1999).
10. N. C. Clark et al., "Detection of a streptomycin/spectinomycin adenylyltransferase gene (aadA) in Enterococcus faecalis", *Antimicrob. Agents. Chemother.* **43**(1), 157–160 (1999).
11. M. Vaneechoutte and J. Eldere Van, "The possibilities and limitation of nucleic acid amplification technology in diagnostic microbiology," *J. Med. Microbiol.* **46**(3), 188–194 (1997).
12. M. J. Gupta et al., "Differentiation of food pathogens using FTIR and artificial neural networks," *Trans. ASAE.* **48**(5), 1889–1892 (2005).
13. H. Lamprell et al., " Discrimination of staphylococcus aureus strains from different species of staphylococcus using Fourier transform infrared (FTIR) spectroscopy," *Int. J. Food Microbiol.* **108**(1), 125–129 (2006).
14. D. Naumann, D. Helm, and H. Labischinski, "Microbiological characterizations by FT-IR spectroscopy," *Nature*. **351**(6321), 81–82 (1991).
15. M. Diem, S. Boydston-White, and L. Chiriboga, " Infrared spectroscopy of cells and tissues: shinning light onto a novel subject," *Appl. Spectrosc.* **53**(4), 148–161 (1999).
16. S. H. Beattie et al., "Discrimination among Bacillus cereus, B. mycoides and B. thuringiensis and some other species of the genus Bacillus by Fourier transform infrared spectroscopy," *FEMS Microbiol. Lett.* **164** (1), 201–206 (1998).
17. D. Lefier et al., "Effect of sampling procedure and strain variation in Listeria monocgtogenes on the discrimination of species in the genus Listeria by Fourier transform infrared spectroscopy and canonical variates analysis," *FEMS Microbiol. Lett.* **147**(1), 45–50 (1997).
18. A. Naumann, "A novel procedure for strain classification of fungal mycelium by cluster and artificial neural network analysis of Fourier transform infrared (FTIR) spectra," *Analyst*. **134**(6), 1215–1223 (2009).
19. R. Linker and L. (Lahkim) Tsror, "Discrimination of soil-borne fungi using Fourier transform infrared attenuated total reflection spectroscopy," *Appl. Spectrosc.* **62**(3), 302–305 (2008).
20. T. Udelhoven, D. Naumann, and J. Schmitt, "Development of a hierarchical classification system with artificial neural networks and FT-IR spectra for the identification of bacteria," *Appl. Spectrosc.* **54**(10), 1471–1479 (2000).
21. A. Salman et al., "Distinction of Fusarium oxysporum fungal isolates (strains) using FTIR-ATR spectroscopy and advanced statistical methods," *Analyst*. **136**(5), 988–95 (2011).
22. G. Fischer et al., "FT-IR spectroscopy as a tool for rapid identification and intra-species characterization of airborne filamentous fungi," *J. Microbiol. Methods*. **64**(1), 63–77 (2006).
23. V. Shapaval et al., "A high-throughput microcultivation protocol for FTIR spectroscopic characterization and identification of fungi," *J. Biophotonics*. **3**(8–9), 512–521 (2010).
24. F. L. Martin et al. "Distinguishing cell types or populations based on the computational analysis of their infrared spectra," *Nat. Protoc.* **5**(11), 1748–60 (2010).
25. J. Trevisan et al., "Syrian hamster embryo (SHE) assay (pH 6.7) coupled with infrared spectroscopy and chemometrics towards toxicological assessment," *Analyst*. **135**(12), 3266–3272 (2010).
26. M. J. Walsh et al., "ATR microspectroscopy with multivariate analysis segregates grades of exfoliative cervical cytology," *Biochem. Biophys. Res. Commun.* **352**(1), 213–219 (2007).
27. L. (Lahkim) Tsror et al., "Aggressiveness of Verticillium dahliae isolates from different vegetative compatibility groups to potato and tomato," *Plant Pathol.* **50**(4), 477–482 (2001).

28. L. (Lahkim) Tsror and M. Hazanovsky, "Effect of co-inoculation by Verticillium dahliae and Colletotrichum coccodes on disease symptoms and fungal colonization in four potato cultivars," *Plant Pathol.* **50**(4), 483–488 (2001).
29. A. M. C. Davies and T. Fearn, "Back to basics: the principles of principal component analysis," *Spectrosc. Europe* **4**(17), 20–23 (2004).
30. C. M. Bishop, *Pattern Recognition and Machine Learning*, Springer, New York (2006).
31. F. Camastra and A. Vinciarelli, *Machine Learning for Audio, Image and Video Analysis*, Springer, London (2008).
32. R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification*, 2nd ed., John Wiley & Sons, New York (2001).
33. A. Zwielly et al., "Discrimination between drug-resistant and non-resistant human melanoma cell lines by FTIR spectroscopy," *Analyst*. **134** (2), 294–300 (2009).
34. M. Diem, P. Griffith, and J. Chalmers, *Vibrational Spectroscopy for Medical Diagnosis*, John Wiley & Sons, New York (2008).
35. R. A. Fisher, "The use of multiple measurements in taxonomic problems," *Ann. Hum. Genet.* **7**(2), 179–188 (1936).
36. G. M. James and T. J. Hasti, " Functional linear discriminant analysis for irregularly sampled curves," *J. R. Stat. Soc. Series B Stat. Methodol.* **63**(3), 533–550 (2001).
37. A. Gersho and R. M. Gray, *Vector Quantization and Signal Compression*, Kluwer Academic Press, Dordrecht, the Netherlands (1992).
38. Z. Movasaghi, S. Rehman, and Rehman, "Fourier transform infrared (FTIR) spectroscopy of biological tissues," *Appl. Spectrosc. Rev.* **43**(2), 134–179, (2008).
39. R. K. Dukor, *Handbook of Vibrational Spectroscopy*, John Wiley and Sons, Chichester UK (2001).
40. H. P. Wang, H. C. Wang, and Y. J. Huang, "Microscopic FTIR studies of lung cancer cells in pleural fluid," *Sci. Total Environ.* **204**(3), 283–287 (1997).
41. H. Fabian et al., "A comparative infrared spectroscopic study of human breast tumors and breast tumor cell xenografts," *Biospectroscopy.* **1**(1), 37–45, (1995).
42. M. F. K. Fung et al., "Pressure-tuning Fourier transform infrared spectroscopic study of carcinogenesis in human endometrium," *Biospectroscopy.* **2**(3), 155–165 (1996).
43. B. Rigas and P. T. T. Wong, "Human colon adenocarcinoma cell lines display infrared spectroscopic features of malignant colon tissues," *Canc. Res.* **52**(1), 84–88 (1992).
44. G. Shetty et al., "Raman spectroscopy: evaluation of biochemical changes in carcinogenesis of oesophagus," *Br. J. Cancer* **94**(10), 1460–1464 (2006).
45. R. Eckel et al., "Characteristic infrared spectroscopic patterns in the protein bands of human breast cancer tissue," *Vib. Spectrosc.* **27**(2), 165–173 (2001).
46. N. Fujioka et al., "Discrimination between normal and malignant human gastric tissues by Fourier transform infrared spectroscopy," *Cancer Detect. Prev.* **28**(1), 32–36 (2004).
47. L. Chiriboga et al., "Infrared spectroscopy of human tissue. I. Differentiation and maturation of epithelial cells in the human cervix," *Biospectroscopy.* **4**(1), 47–53 (1998).
48. S. Yoshida et al., "Fourier transform infrared spectroscopic analysis of rat brain microsomal membranes modified by dietary fatty acids: possible correlation with altered learning behavior," *Biospectroscopy.* **3**(4), 281–290 (1997).
49. G. I. Dovbeshko et al., "FTIR spectroscopy studies of nucleic acid damage." *Talanta*. **53**(1), 233–246 (1997).
50. P. G. L. Andrus and R. D. Strickland, "Cancer grading by Fourier transform infrared spectroscopy," *Biospectroscopy.* **4**(1), 37–46 (1998).
51. N. J. Harrick, "Principles of internal reflection spectroscopy," Chapter 2 in *Internal Reflection Spectroscopy*, Harrick Scientific Corporation, Ossining, New York, pp. 13–66 (1979).
52. C. L. Curl et al., "Refractive index measurement in viable cells using quantitative phase-amplitude microscopy and confocal microscopy," *Cytometry* **65A**(1), 88–92 (2005).
53. E. Bogomolny et al., "Attenuated total reflectance spectroscopy: a promising technique for early detection of premalignancy," *Analyst*. **135**(8), 1934–1940 (2010).
54. MATLAB, Version 7.0 (R14), The MatWorks Inc. Natick, MA (2007).
55. M. J. German et al., "Infrared spectroscopy with multivariate analysis potentially facilitates the segregation of different types of prostate cell," *Biophys. J.* **90**(10), 3783–3795 (2006).
56. T. Evgeiou, M. Pontil, and A. Elisseeff, "Leave-one-out error, stability, and generalization of voting combination of classifiers" *Mach. Learn.* **55**(1), 71–97 (2004).
57. A. Elisseeff and M. Pontil, "Leave-one-out error and stability of learning algorithms with applications," in *Advances in Learning Theory: Methods, Models and Applications*, J. A. K. Suykens et al., Eds., NATO Science Series III: Computer and Systems Sciences, Vol. **190**, pp. 111–124, IOS Press, Amsterdam, the Netherlands (2003).