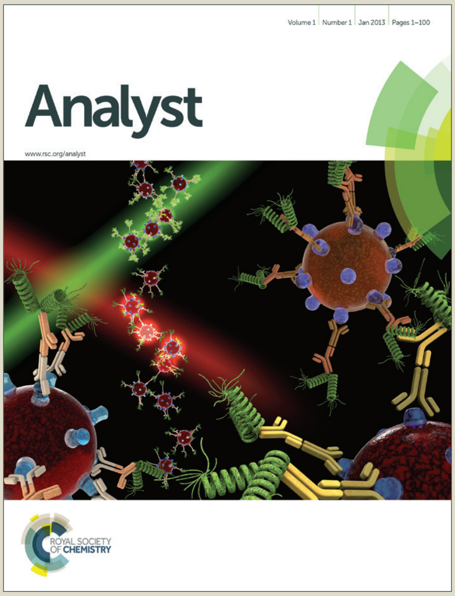
Analyst

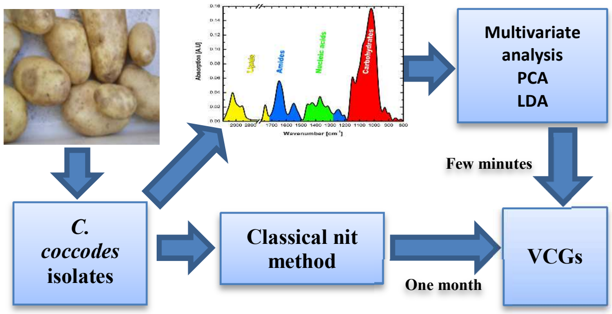
Accepted Manuscript


Volume 1 | Number 1 | Jan 2013 | Pages 1–100
Analyst
www.rsc.org/analyst
ROYAL SOCIETY OF CHEMISTRY

ROYAL SOCIETY OF CHEMISTRY

www.rsc.org/analyst

FTIR spectroscopy may provide a specific, rapid, and inexpensive method for successful classification of *Colletotrichum coccodes* isolates into Vegetative Compatibility Groups.



Assignment of *Colletotrichum coccodes* isolates into vegetative compatibility groups using infrared spectroscopy: a step towards practical application

A. Salman[*a], E. Shufan[a], I. Lapidot[b], L. Tsror[c], R. Moreh[d],

S. Mordechai[d] and M. Huleihel[*e]

[a] *Department of Physics, SCE - Shamoon College of Engineering, Beer-Sheva 84100, Israel.*

[b] *Department of Electrical and Electronics Engineering ACLP- Afeka Center for Language Processing, Afeka . Tel-Aviv Academic College of Engineering, Israel.*

[c] *Department of Plant Pathology, Institute of Plant Protection, Agricultural Research Organization, Gilat Research Center, M.P. Negev, 85250, Israel.*

[d] *Department of Physics, Ben-Gurion University of the Negev, Beer-Sheva 84105, Israel.*

[e] *Department of Microbiology, Immunology and Genetics, Faculty of Health Sciences, Ben-Gurion University of the Negev, Beer-Sheva 84105, Israel.*

Abstract

Colletotrichum coccodes (*C. coccodes*) is a pathogenic fungus that causes anthracnose on tomatoes and black dot disease in potatoes. It is considered as a seed tuber and soil-borne pathogen that is difficult to control. *C. coccodes* isolates are classified into Vegetative Compatibility Groups (VCGs). Early classification of isolates into VCGs is of great importance for a better understanding of the epidemiology of the disease and improving its control. Moreover, the differentiation among these isolates and the assignment of newly-discovered isolates enables control of the disease in its early stages. Distinguishing between isolates using microbiological or genetical methods are time-consuming, and not always available. Our results show that it is possible to assign the isolates into their VCGs and to classify them in the isolate level with a high success rate using Principal Component Analysis (PCA) and Linear Discriminant Analysis (LDA).

Keywords: *C. coccodes,* Infrared spectroscopy, VCG, PCA, LDA.

***Corresponding authors:**

Dr. Ahmad Salman, Fax: +972-8-6475758; Tel: +972-8-6475794; E-mail: ahmad@ sce.ac.il

Prof. Mahmoud Huleihel, Fax: +972-8-6479867; Tel: +972-8-6479867; E-mail: mahmoudh@ bgu.ac.il

Introduction

The use of Fourier Transform InfraRed-Attenuated Total Reflection (FTIR-ATR) spectroscopy in tandem with multivariate analysis and advanced statistical methods for soil-borne fungi classification has increased usage worldwide[1-4]. Investigation of *C. coccodes* is very important because this pathogen attacks a variety of plants and crops, causing premature death of the plants and severe damage to tubers[5-

[7], reducing their marketability and resulting in significant economic loss[8-10]. *C. coccodes* is responsible for up to 50% of crop losses, as indicated by experimental studies in the U.S.A. [11], Israel [12], U.K.[10], and Malaysia[13]. Moreover, *C. coccodes* is responsible for additional economic losses to the seed industry, particularly to the export market [14].

Classic microbiological, biochemical, immunological, and molecular methods are the primary means currently used to detect and characterize fungal pathogens. Classic microbiological methods are based on visual and microscopic observations of the fungus, after it has been cultivated in selective media[15]. These methods are time-consuming[10] (often taking weeks), with low specificity[16]. In addition to these restrictions, these methods have limited success in also differentiating among different fungi at species or isolates levels[17, 18]. Biochemical methods are limited at the isolates level because the specific biochemical variations between different isolates of the same species are not well understood [19]. Immunological methods rely on the interactions between a specific antibody and one of the pathogen proteins[20, 21], and depend on the availability of specific monoclonal antibodies appropriate to the tested fungi. Molecular methods are based on the Polymerase Chain Reaction (PCR). In samples that have available primers, specific DNA fragments are amplified[22]. It is possible to detect *C. coccodes* [9, 23] using PCR and real-time PCR tests. Although molecular techniques are very specific they are expensive and not readily available for different isolates[24-27].

Isolates that can transfer genetic material by contact [28], producing new stable heterokaryons, form subpopulations that tend to be similar due to a common genetic pool and are assigned to the same VCG. *Fusarium oxysporum*[29-31], *Verticillium* [32], and *Colletotrichum* [33, 34] phytopathogens were classified into VCGs in order to study the genetic structure of their populations. Isolates that belong to the same VCG have similar pathogenic aggressiveness, making it important to classify any isolate to a specific VCG in order to improve the understanding of the epidemiology of the disease and enable its control[35].

The main technique for analysis of VCGs is the nitrate non-utilizing (nit) method. This technique relies on pairing between complementary nit mutants[34, 36, 37], which are selected using a potassium chlorate-containing medium. Two isolates are assigned to the same VCG if their mutants can form stable heterokaryons [28, 37]. This method is time-consuming, taking as much as one month to produce usable results.

Encouraging results in detection and characterization of various types of phytopathogens was reported using infrared spectroscopy[1, 2, 4, 38-46]. More recently, many infrared spectroscopic studies have been carried out to investigate the *C. coccodes* soil-borne fungus and classify the samples in the isolate level[38-40]. Using infrared spectroscopy, the vibrational spectrum, which is considered as a fingerprint of the sample, is measured. The FTIR-ATR sampling technique is based on evanescent wave absorption. This technique is of special interest because it could be used to identify specific spectroscopic changes *in situ* and *in vivo*[47-49].

In our previous study, we examined the potential of FTIR-ATR spectroscopy as a tool for assigning 14 *C. coccodes* isolates into five VCGs[50]. The FTIR-ATR spectra were analysed using advanced statistical and mathematical tools, including PCA and LDA.

In the present study we have taken this method one step further towards its practical application in the real world. Thirty five *C. coccodes* isolates were measured and assigned into eight VCGs that are known to exist in Israel; they were classified simultaneously in the isolate level.

Materials and Methods

Fungal isolates

All samples were obtained from the Department of Plant Pathology, the Institute of Plant Protection, Agricultural Research Organization, at the Gilat Research Center, Israel. All tested samples were isolated from infected potato plants and tubers sampled from Southwest Negev, Israel. *C. coccodes* were isolated from surface-sterilized (1% NaCl for 10 min) infected stems or tubers, placed on potato dextrose agar (PDA) plates, incubated in the dark at 27ºC for 7 days, and allowed to sporulate. Sclerotia were placed on potato dextrose agar (PDA) plates and incubated in the dark at 27ºC for 7 days and allowed to sporulate. Conidia were incubated in a medium containing 0.2% sorbose, 15% agar, and 100 ppm streptomycin sulphate (SA) for 24 h at 27ºC in the dark. Monoconidial cultures were obtained from each isolate (by micromanipulation) and maintained on czapek dox agar (CDA) at 6ºC. Assignment of these isolates to VCGs was carried out as previously reported [37, 51].

Five samples of each isolate were grown in different batches at 27°C with continuous shaking for **5** days to achieve comprehensive growth of the samples.

For spectroscopic measurements, fungi were separated and purified by spinning about 1.5 ml of the fungal suspension at 13200 rpm for 4 minutes, washing the pellet 4 times with distilled water, and suspending it with about 1 ml distilled water.

Sample preparation

Special precautions should be taken in preparing homogeneous fungi samples for measurements. Due to their complicated structures, and due to the fact that fungal hyphae have the ability to highly aggregate in water, we encountered some difficulties in preparing a homogeneous suspension of the fungi in water and spreading them evenly on the ZnSe crystal surface of the ATR accessory. We made the fungal sample as homogeneous as possible, by cutting the sample into smaller fragments through repeated pipetting. Thus it was easier to spread the sample across the ATR zinc selenide crystal (trapezoid shape, 80 mm long, 10 mm wide and 4 mm thick) in order to obtain a high signal spectrum. 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), air dried for about 30 minutes, and thereafter measured by ATR spectroscopy.

We measured the spectra that were prepared from different batches and bottles at the same conditions (including the use of the same preparation and isolation techniques, the same methods of drying the sample, and using the same spectrometer).

In this study, 911 measurements were performed from 35 different isolates. At the same time, about 12 different fungal isolates were isolated from the potatoes crops in parallel, purified, grown and measured as described before; one sample from each isolate a day. In one day 12 measurements belonged to 12 different isolates were measured on average. We continued measuring the samples in the next day one measurements from each isolate. In one week, seven measurements from each isolate were obtained. The purified isolates were stored at $4^{\circ}C$ during the entire week of the measurements.

In the successive week, the same 12 isolates were grown in fresh media for 5 days and we repeated the measurements as detailed above until acquiring all the planned measurements.

FTIR measurements

We used Tensor 27 (Bruker Optic Germany) in the ATR mode, attached to DTGS detector. The ATR uses a ZnSe crystal (PIKE technologies) with a trapezoid shape. The samples were air dried before measurements, and were scanned 64 times in the range of 675-4000 cm^{-1}, with a 4 cm^{-1} spectral resolution.

Spectral Manipulation

Table 1 summarizes the VCG, isolates, and the number of measurements acquired from each VCG measured in this study.

All the spectra were manipulated using the same protocol, which employed the “OPUS 7” (Bruker, Germany) commercial software program.

Table 1: Details of the VCGs, the isolates, and the number of measurements included in this study.

	VCG	Number of	Number of
	Number	isolates	measurements
	1	4	102
	2	4	104
	3	5	119
	4	6	154
	5	4	115
	6	4	96
	7	4	111
	8	4	110
Sum	8 VCGs	35 Isolates	911 measurements

ATR correction: The ATR correction is essential and should be to correct for the different penetration depth at different wavelengths of the radiation. The ATR crystal is of a trapezoid shape and is 80 mm long, 10 mm wide, and 4 mm thick (chosen to produce optimum performance). The ATR crystal should be chosen carefully so its a refractive index is much higher than that of the sample. In our case, the angle of incidence is 45°, the critical internal reflection angle is about 33.4° (assuming a refractive index of 2.4 for the ZnSe crystal at 1000 cm^{-1}, and the sample refractive index is 1.35) [52-54]. Thus we insure a total internal reflection inside the crystal. The larger penetration depth yields greater absorption at higher wavelengths and thus, the maxima of the bands are red-shifted in the ATR measurements.

Smoothing: Smoothing was performed using the Savitzky-Golay algorithm with 13 points.

Bisecting: The spectra were then bisected into two regions (900-1775 cm^{-1} and 2800-3000 cm^{-1}), to exclude the water absorption bands (3000-4000 cm^{-1}) and the "dead" region (1775-2800 cm^{-1}).

Baseline correction: Baseline correction was employed by choosing the “concave rubberband” algorithm with the following parameters: number of baseline points equals 64 (i.e., the spectrum is divided into 64 equally sized ranges); and an equal number of iterations.

Vector normalization: The average intensity $\bar{y}$ is calculated and subtracted from the spectrum (

$\tilde{y}_i = y_i - \overline{y}$). A new spectrum is defined by $x_i = \frac{\tilde{y}_i}{\sqrt{S}}$, where $S = \sum_{i=1}^{N} \tilde{y}_i^2$ (variance). The vector norm of the resulting spectrum therefore equals 1 ($\sum_{i=1}^{N} x_i^2 = 1$).

Offset: The spectrum minimum after vector normalization is shifted to zero.

Supplementary Figure1 shows some of our spectra before and after manipulation. Supplementary Figure 1a shows IR absorption spectra as raw data after ATR correction. Supplementary Figure1b shows the same spectra after manipulation before and after vector normalization.

We focused in our analysis on the 900-1775 cm^{-1} region[55] because it gave the best classification results

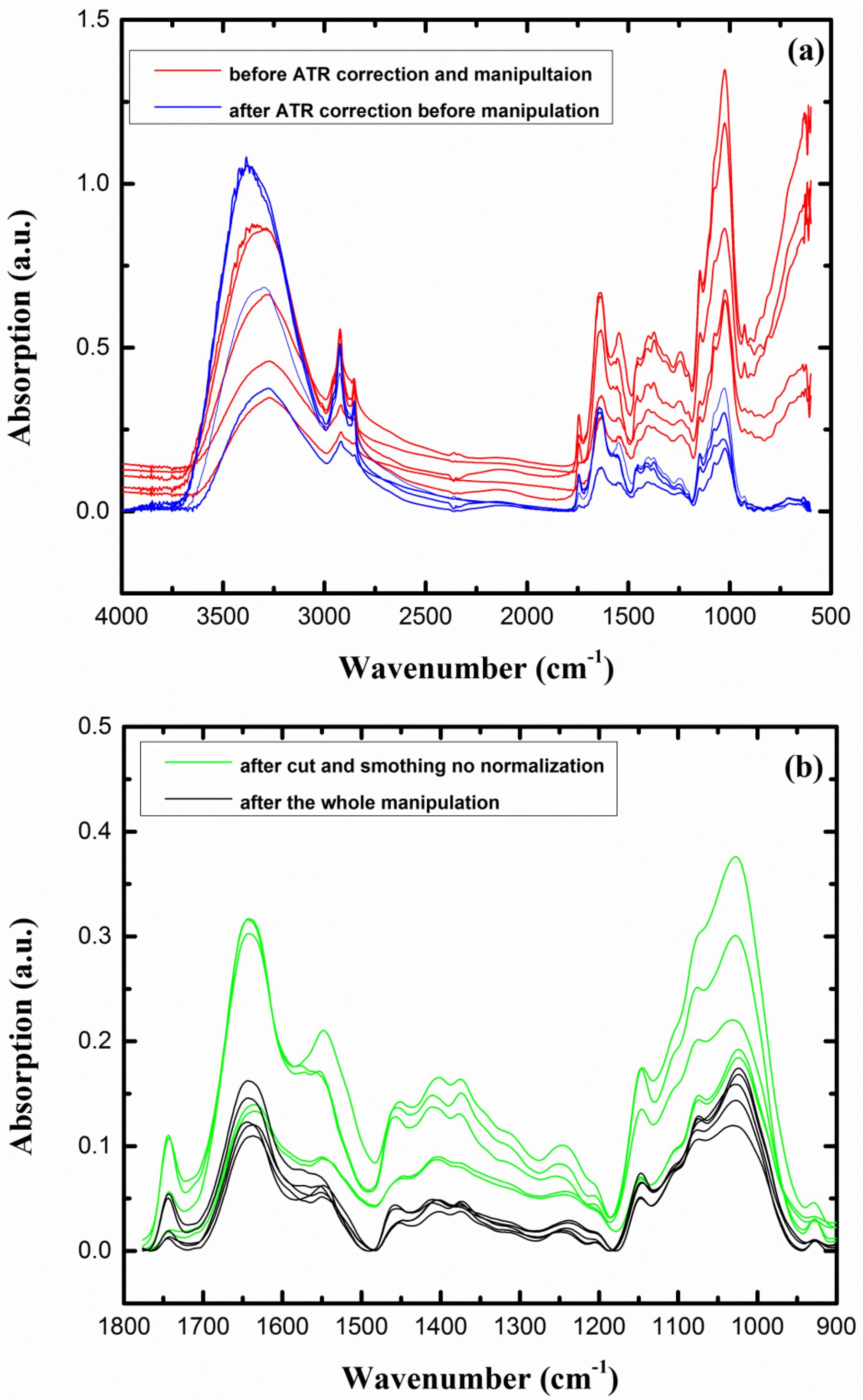
PCA and LDA Statistical Analysis

PCA and LDA are widely used as differentiation tools[56] in problems involving biological samples[38, 39, 54]. PCA is first applied over the spectroscopic measurement. The latter is the raw pattern recognition feature vector. Consider a set of N measurements, each containing D points (intensity versus wave-number) (D=506 in our case). A vector in a D-dimensional space represents each measurement. PCA refers to a linear transformation for which a new basis is chosen to represent this vector. The first basis vector (PC1) is the one with maximal variance; the other PCs are chosen with decreasing variances[57].

The new basis corresponds to the eigenvectors of the covariance matrix $\Sigma = \frac{1}{N}\sum_{n=1}^{N}(x_n - \overline{x})(x_n - \overline{x})^T$,

where $\{x_1, \ldots, x_N\}$ is the set of measurements (column vectors) and $\overline{x}$ is its mean (T denotes the transpose operation). The variance of the data projected to a given principal axis equals to the eigenvalue of the corresponding eigenvector.

Applying PCA leads to dimensional reduction[58, 59] and may lead to recognition improvement: in the PCA procedure only a partial set of the basis vectors are chosen to represent the sample (in this work



Supplementary Figure 1: Five absorption spectra of different isolates. (a) region 675-4000 cm^{-1}, as raw data before and after ATR correction. (b) region 900-1775cm^{-1}, after manipulation (cutting and smoothing) before and after vector normalization.

PC1-PCd, with $1 \le d \le D$), and therefore the new feature vector has less dimensions than the original one. Moreover, in many problems the variability is correlated with the separability[38, 39, 54], so that properly choosing d will lead to a better feature vectors in the separability sense. After applying the PCA procedure, we use Fisher linear classifier: the probability density of each class c (out of C classes) is assumed to be a Gaussian centered around the mean μ_k, and all the classes share the same covariance matrix Σ. The category of a given measurement x is then given by $\underset{c \in \{1,\ldots,C\}}{\arg\max} \left\{ x^T \Sigma^{-1} \mu_c - \frac{1}{2} \mu_c^T \Sigma^{-1} \mu_c + \log\left(\pi_c\right) \right\}$, where $\pi_c = N_c / N$, and N_c is the number of measurements that belong to category c.

The identification success was calculated using two variants of k-fold cross-validation frequently applied in pattern recognition. Leave one out (LOO) [58] was the first variant used and obtained when $k = N$, where N is the number of data points. The LOO approach is generally used for small amounts of data. The second was 5-folds, i.e., 20-80% when 80% of the data was used for training and 20% for testing.

We performed PCA on the spectra followed by Fisher linear classifier after all the manipulations (Supplementary Figure 1b). We applied Fisher linear classifier with the LOO method 911 times (the number of spectra), where each time we took 910 spectra for training, and we predicted the type of the 911th spectrum. With the 20-80% method we applied the Fisher linear classifier calculations 100 times. Each time 80% of the results were used for training and 20% for validation. The training sets were chosen randomly from the results with an aim to include spectra from all categories.

Results

Typical infrared absorption spectra of *C. coccodes* isolates are shown in Figure 1. The spectra are averages of 11 isolates belonging to 4 different VCGs, chosen arbitrarily from the isolates investigated in this study. Each VCG is represented with a different color. The standard deviations of four isolates, one for each VCG, are plotted as error bars in the figure.

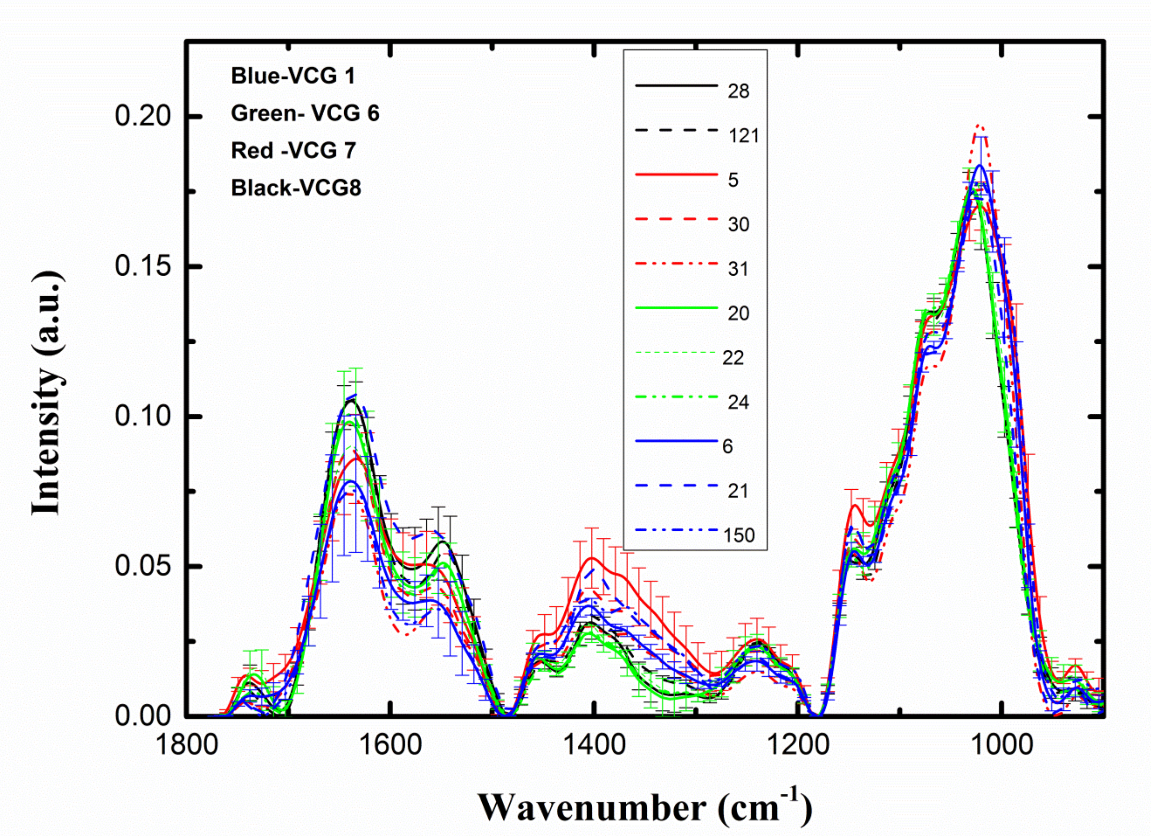


Figure 1: Average spectra of 11 *C. coccodes* isolates belonging to 4 different VCGs are shown in the 900-1775 cm^{-1} range. The VCGs are plotted with different colors. The numbers in the text box refer to the serial number labels chosen for the different isolates. The error bars are the standard deviation of the four isolates associated with different VCGs as labelled in the figure.

The spectra are dominated by a large peak in the 1185–900 cm^{-1} range[2]. Polysaccharide carbohydrates and nucleic acid vibrations are the main contributors in this region. Chitin, a molecule specific to fungi, absorbs at 1151 cm^{-1} and 1078 cm^{-1}, due to its C-O and C-C stretching vibrations[60, 61]. C-O stretching vibration of glycogen[62] attributes at 1024 cm^{-1}. Amide III bands of proteins[61] attribute at 1240-1310 cm^{-1}. The P=O functional groups of proteins, nucleic acids and phospholipids attribute at 1085-1090 cm^{-1} due to their symmetric stretching and at 1220-1250 cm^{-1} due to their anti-symmetric stretching [63]. Lipids[63], sugar rings[64, 65], and phosphate compounds[66, 67] attribute to the absorption bands in the 1185-1485 cm^{-1} range.

Amide I (C=O carbonyl stretch) and Amide II (C-N stretching and a CNH bending vibrations)[68] with centroids, at 1650 cm^{-1} and 1553 cm^{-1} respectively, are considered to be the main features in the 800-1775 cm^{-1} region. The principal spectral features in the 2800-3000 cm^{-1} region (data not shown) are contributed mainly from phospholipids[62] proteins due to the CH_2 and CH_3 functional group (symmetric and anti-symmetric stretching) with vibrational bands centered at 2853 and 2922 cm^{-1}.

As can be seen from Figure 1, the spectra of the different isolates are overlapped with their error bars, although there remain small but identifiable differences in the intensities and shapes at different wavenumbers when they are compared as couples. For example, in the range centered at 1400 cm^{-1}, the red isolates are clearly different than the green isolates, whereas the latter overlap with the black isolates in this range. Looking carefully at the figure, there are some differences between the isolates that belong to the same VCG 7 (labelled in red) in this range, but still they are significantly different from the other tested VCGs.

The measured spectra were transformed by PCA. Graphs based on two or three PC dimensions can provide a distinction between several groups. We tried many 2D and 3D plots of the PCA scores of the eight VCGs but the results were poor. In order to obtain good separation between all eight VCGs we should consider much higher PCs following by Fisher linear classifier.

An interesting question is how many PCs should optimally be used in the multivariate analysis? Using PCA calculations we find a new orthogonal basis by which all the spectra are calculated. For example, in Figure 2(a), each spectrum was presented as a linear combination of ten loadings (labelled “calculated”), in order to demonstrate the power of PCA in our study (where instead of using 506 components we need only up to 39 components). The measured spectrum (blue) and the calculated spectra were compared using different PC numbers. A good agreement was achieved using 39 PCs illustrating the feasibility and the high accuracy of the PCA method employed in the present study.

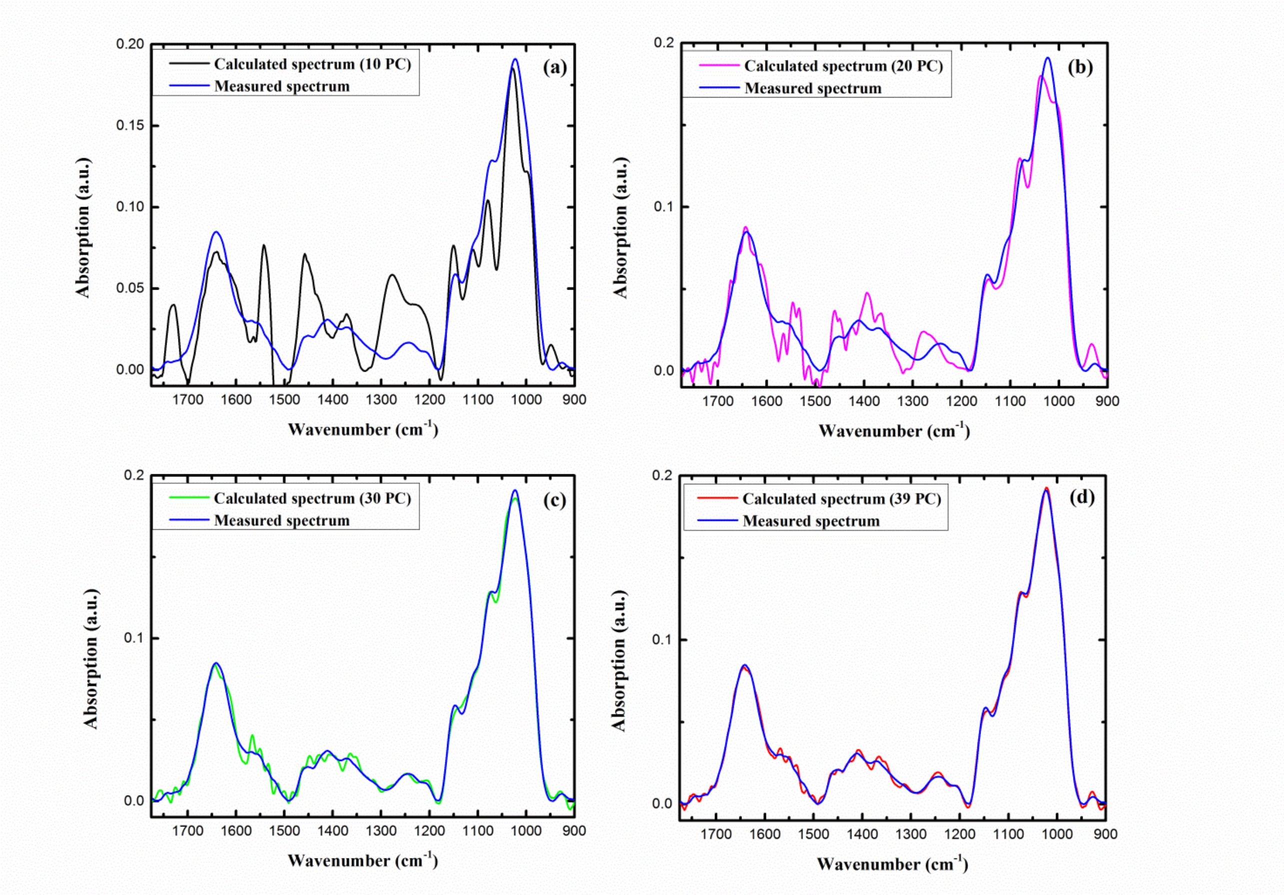


Figure 2: Comparison between the measured and the calculated spectra based on different PCs number selected in the range of 900-1775 cm^{-1}. The measured spectrum is the same in the four panels, while the calculated spectra were calculated using: (a) 10 PCs, (b) 20 PCs, (c) 30 PCs and (d) 39 PCs. The best agreement between the measured and the calculated spectra was achieved in (d) revealing a good agreement and illustrating the feasibility and the high accuracy of the chosen analysis.

Figure 3 (a&b) shows the identification success rate, stated in percentage, as a function of PC number for assigning the isolates into their VCGs and for the isolate level classification. The identification success rate was calculated by Fisher linear classifier using the LOO and 20-80% algorithms. The identification success for assigning the isolates into their VCGs was $85.7\% \pm 2.9\%$ using the 20-80% method and 86.3% using the LOO method with 39 PCs.

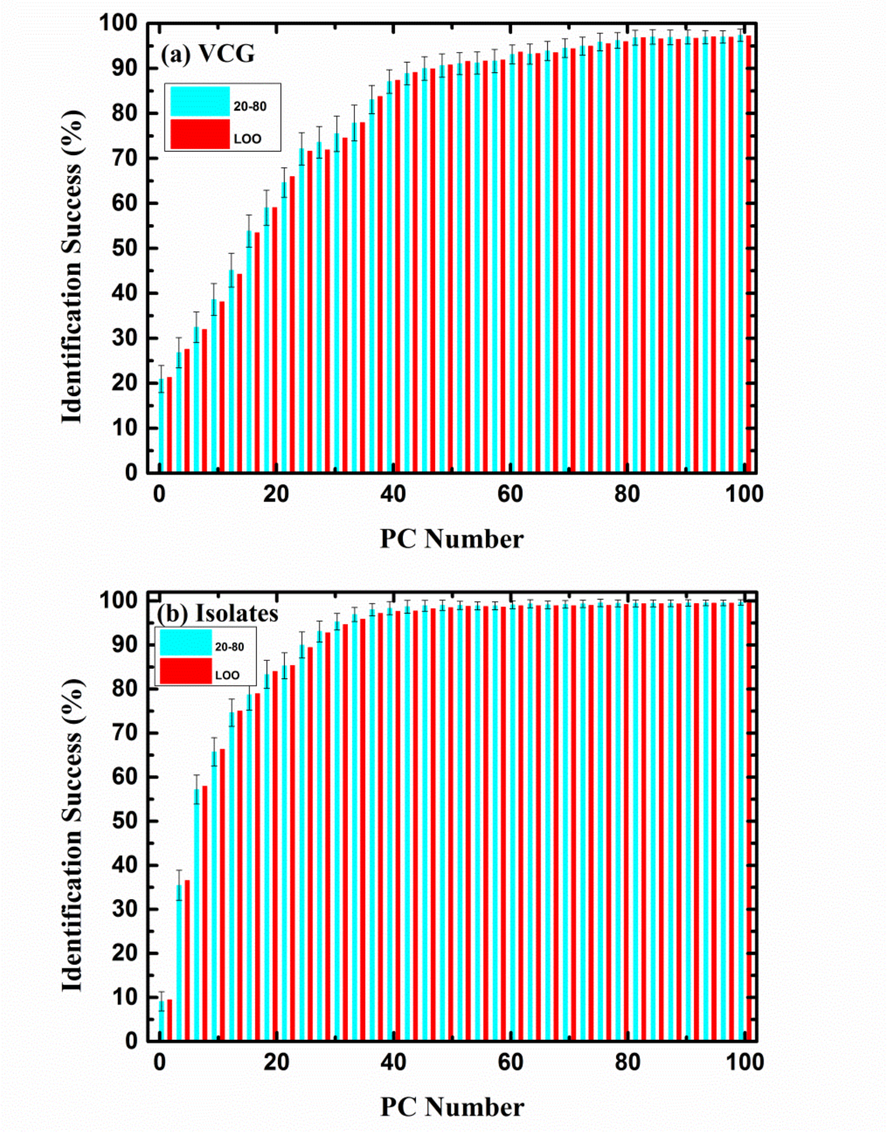


Figure 3: Identification success rates versus PCs number for the two classification procedures. The identification success was calculated using the LOO and the 20-80% approaches. a) VCGs level and b) isolate level.

The identification results obtained using the LOO method, for classification of the different samples into VCGs, and differentiation between them at the isolates level, are shown in Table 2 (a) and 2 (b), respectively. It is important to mention that the choice of 39 PCs was based on variance consideration; however for the classification purposes it is better to increase the number of PCs as can be seen in Figure 3a.

Table 2 (a, b): Success identification rates presented as a confusion matrix of (a) *Colletotrichum coccodes* VCGs, (b) *Colletotrichum coccodes* isolates (labelled as C). Identifications were obtained using Fisher linear classifier calculations and the LOO algorithm in the 900-1775 cm^{-1} low wavenumber region.

a)

	VCG 1	**VCG 2**	**VCG 3**	**VCG 4**	**VCG 5**	**VCG 6**	**VCG 7**	**VCG 8**
VCG 1	99	0	2	0	0	0	1	0
VCG 2	0	92	5	0	0	1	6	0
VCG 3	0	0	107	6	5	0	1	0
VCG 4	0	0	5	144	0	0	4	1
VCG 5	7	0	0	8	80	2	18	0
VCG 6	0	0	7	0	17	71	1	0
VCG 7	12	0	1	0	1	0	96	1
VCG 8	0	1	0	0	10	0	2	97

Discussion

The classification of *C. coccodes* samples in the isolates level and assigning them into their VCGs was carried out using the FTIR-ATR spectroscopic method, in tandem with multivariate analysis with PCA followed by Fisher linear classifier [55]. Simple methods like K-means, clustering, and spectral biomarkers were not able to achieve these goals. Using the FTIR-ATR method takes only a few minutes to determine the VCG of a certain isolate. Moreover, the method is fully computerized, highly objective, and the examined isolates can be assigned into VCGs and classified at the isolate level simultaneously. These achievements provide distinct advantages over classical methods, because the pairing between complementary nit mutants[34, 36, 37] for VCGs, can take about one month to reach the same determination.

In the PCA calculations, the spectra were transformed to another domain by mathematical transformation. In this domain, the transformed spectra are easier to use for further analysis.

The biological samples used in this study are isolates that belong to the same species. They are similar in their genome, components, and structure[39, 40]. This similarity is reflected in their infrared absorption spectra, as shown in Figure 1. The spectra overlap in their error bars and have only minute differences. When the differences among the classes are large, as, for example, in different generic samples, 2D figures based on projections of two PCs were sufficient to differentiate among them[38]. In this study, the large number of isolates, taken from the entire group of eight known VCGs, and the similarity of these isolates, makes the differentiation into VCGs between these samples a challenge for multivariate analysis. We tried different projections at different directions, but the differentiation results were poor. The conclusion was that a more sophisticated classifier is needed to achieve the classification for all the investigated groups simultaneously. Therefore, we used Fisher linear classifier with LOO and 20-80% algorithms combined with the PCA calculation. The LOO algorithm is a common method used for cross-validation in small populations[59], and has been extensively explored in machine learning for estimating the error. There is an excellent correlation between the results of the two approaches as can be seen from Figure 2.

Choosing the number of PCs is an interesting issue. The interclass variance in the isolates level was much less than the interclass variance in the VCG level. VCG class consists of few isolates. The b)

	101	104	105	107	11	121	124	133	138	145	14	150	154	15	166	177	190	192	19	20	21	24	25	27	2	30	31	4	56	5	69	6	70	83	88
101	21	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	3	0	0	0	8	0	0	0	0	0	0	0	0	0
104	0	27	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
105	0	0	21	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
107	0	0	0	22	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
11	0	0	0	0	26	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
121	0	0	0	0	0	18	3	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
124	0	0	0	0	0	0	31	0	0	0	0	6	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	5	0	0	0	0	0
133	0	0	0	0	0	0	0	21	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
138	0	0	0	0	0	0	0	0	40	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
145	0	0	0	0	0	0	0	0	0	23	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
14	0	0	0	0	0	0	0	0	0	0	20	0	0	0	0	0	0	0	0	0	0	7	0	0	0	3	0	0	0	0	0	0	0	0	0
150	0	0	0	0	0	0	2	0	0	0	0	23	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0
154	0	0	0	0	0	0	0	0	5	0	0	0	39	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
15	0	0	0	0	0	0	0	0	0	0	0	0	0	18	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
166	0	0	0	0	0	0	0	0	0	0	0	0	0	0	19	0	0	0	0	0	0	0	0	0	0	0	0	5	0	0	0	0	0	0	0
177	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	20	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
190	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	24	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
192	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	19	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
19	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	29	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
20	0	0	0	0	0	0	5	0	0	0	0	0	0	0	0	0	0	0	0	17	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
21	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	16	0	0	0	0	0	0	0	0	0	0	0	0	0	4
24	0	0	0	0	0	0	5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	18	0	0	0	0	0	0	0	0	0	0	0	0	0
25	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	22	0	0	0	0	0	0	0	0	0	0	0	0
27	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	22	0	0	0	0	0	0	0	0	0	0	0
2	0	0	0	0	22	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	22	0	0	0	0	0	0	0	0	0	0
30	2	0	0	0	0	0	0	0	0	0	0	4	0	0	0	0	0	0	0	0	0	0	0	0	0	18	0	0	0	0	0	3	0	0	0
31	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	22	0	0	0	0	0	0	0	0
4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	8	0	0	0	0	0	0	0	0	0	21	0	0	0	0	0	0	0
56	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	22	0	0	0	0	0	0
5	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	3	0	0	0	0	0	0	0	2	0	0	0	14	0	0	0	0	0
69	0	0	0	0	0	0	2	0	0	2	4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	10	0	0	0	0	23	0	0	0	0
6	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	4	0	10	0	0	0
70	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	21	0	0
83	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	26	0
88	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	22

changes in the interclass variance was reflected in the number of PCs that were needed to achieve good classification (Figure 3a and 3b). Using 39 PCs, it was possible to achieve ~ 97% success rate in the classification in the isolate level and ~ 85% success rate in assigning these samples into eight VCGs.

As mentioned in material and methods, the loadings which accounts for the high variance are not always the most separable directions. At the same time, the fact that some high PC loading has small variance does not necessarily indicate that it has no positive impact on the separability of the data. There is no reason to assume that there is a correlation between preserving the signal's variance and the classification abilities of the projected data. Such a correlation might be found empirically on specific data for a specific task. There are two popular approaches to define the number of principal components: 1. Dimensionality reduction task while preserving most of the signal’s variance: in this approach the variance to be preserved is determined and the minimal number of principle components (with the highest eigenvalues) which satisfy the variance constraint is chosen.

2. Cross validation: When the goal is to classify the data, as in our case, the number of principal components needed to preserve the signal's energy is not relevant. In this case, the data is divided into three sets: training, development, and test. The eigenvalues and the eigenvectors are calculated for the training set and sorted in decreasing order. Then, for the first d components, where d=1,2,3, ..., the classifier is trained. Each classifier is tested on the development set. The best classifier, i.e., one that gives the lowest classification error, is tested on the test set.

When the total amount of data is small, it is difficult to have a sufficient amount of data for training, development, and test sets. Instead, the k-folds approach is applied, and in the extreme cases, the method turns to LOO. The LOO method has a drawback in that the training sets are highly correlated. Nevertheless, what is more important is that the test sets are disjoint sets (and as the samples are assumed to be statistically independent, therefore all the test sets are statistically independent). For a very small amount of data this assumption is the price that must be paid in order to have a sufficient amount of data for training (correlation matrix for PCA and the classifier).

In addition, for the VCG classification where we had enough data, we performed cross-validation by dividing the data into three sets: 60% for training, 20% for development (determining the PCs number to be used) and 20% for test validation. The training, development, and test sets were determined randomly. We repeated these experiments 10 times; the results are listed in the following Table 3. The average error between identification success for the development and the test set was less than 2.0 %. It is important to emphasized that the original vector dimension was 506.

Table 3: Successful identification rates of assigning *C. coccodes* samples into their VCGs, using cross validation approach based on the 900-1775 cm^{-1} low wavenumber region. The data was divided into 3 sets 60% for PCA calculation 20% for development (determining the PCs number to be used) and 20% for test validation.

	Optimal PC number	Success rate for the development set	Success rate for the test set.	Error in percentage
1	80	98.9	96.2	2.7
2	95	98.4	97.3	1.1
3	80	98.4	94.5	3.9
4	82	99.5	95.1	4.4
5	89	95.1	95.6	0.5
6	94	89.4	95.6	2.8
7	92	99.5	98.4	1.1
8	95	98.4	95.6	2.8
9	97	99.5	99.5	0
10	79	97.8	96.2	1.6

The aim of the PCA is not to improve the classification performance (sometimes it could happen) but to preserve the classification performance using features' vectors which have lower dimensions. As a consequence, such dimensionality reduction can save computer resources as memories and processing time and simplifying the classifier[69].

In this field, choosing the number of PCs depends on the number of classes, the level of classification (genus, species or isolates), and the sample type. For example, when investigating three biological systems belonging to different genera, with typically large class differences, 3 PCs were found enough to yield excellent success rate[39]. In other study of differentiating among 15 isolates of *Colletotrichum coccodes*[40], 16 PCs were used to achieve 90% success rate in the region 800-1775 cm^{-1}. In the present study we used 35 isolates of the same species of *Colletotrichum coccodes* the feature vector has 506 dimensions. Employing PCA, the vector dimension was reduced to 39 (Figure 2). Thus the vector dimension after dimensionality reduction is ~8% relative to the original vector size compared to speaker recognition field where the ratio was up to 50%[70].

The prediction of the validation sets was done by our system and was based on different features of the classified categories derived from the training sets. This method is always improved by enlarging the number of spectra; because the number of training sets is enlarged, thus the results of the differentiation are improved.

The differences derived by our system, which enabled us to differentiate among the various classes, could not be related to specific IR absorption bands, but were instead spread over the entire region. This feature is a limitation of infrared spectroscopy. Nevertheless, this limitation does not affect the practical advantages of the FTIR-ATR spectroscopy system together with multivariate analysis, because the main issue of the VCG classification procedure can be achieved with a good success rate of 85 %. It will be interesting to find out and identify the biomarkers which might lead to higher success rate. Trevisan[71] et al. suggest a method based on a general frame work for biomarker identification applicable to the FTIR datasets. It is worthwhile to test this method and make a correlation with biology in a further study.

We used a model for the ATR correction developed by Bruker Optic Germany to correct for the different penetration depth at different wavelengths of the radiation which may lead to the red shift[52-54]. Although the above Bruker ATR correction model is simple, it helps to partially compensate for this phenomena. This issue is much more significant when the analyzed spectra were measured using different sampling techniques such as transmission and ATR. In this study, however, all the spectra were measured using the same ATR sampling technique.

It is very important to develop the FTIR-ATR spectroscopic method in this field of research as it is similar to the remote fiber-optic probes technique which may lead in the future to the highly desired *in vivo* measurements.

Conclusions

In this study, we showed that the method of FTIR-ATR spectroscopy in tandem with multivariate PCA and LDA calculations could be a practical method for assigning *C. coccodes* isolates into their VCGs, while simultaneously classifying the sample in the isolate level. In fact, all known VCGs of *C. coccodes* in Israel were used in this study and were successfully classified by this method. The technique was proved as an effective and promising method for classifying the samples into VCGs and for a rapid identification of various fungal isolates, with some notable advantages over the standard microbiological methods. Further examination of all the known isolates of *C. coccodes* in Israel is required in order to fully establish the potential of this spectroscopic method for accurate classification of fungal isolates into VCGs, and for a rapid identification of these isolates. Moreover, this method should be applied for different genus and species samples, in attempts to assign them into their VCGs.

Acknowledgments

Financial support by SCE – Shamoon College of Engineering internal research funding is gratefully acknowledged.

References

1. R. Linker and L. Tsror Lahkim, *Applied spectroscopy*, 2008, **62**, 302-305.
2. A. Naumann, *The Analyst*, 2009, **134**, 1215-1223.
3. A. Naumann, M. Navarro-Gonzalez, S. Peddireddi, U. Kues and A. Polle, *Fungal genetics and biology : FG & B*, 2005, **42**, 829-835.
4. A. Salman, A. Pomerantz, L. Tsror, I. Lapidot, A. Zwielly, R. Moreh, S. Mordechai and M. Huleihel, *The Analyst*, 2011, **136**, 988-995.
5. M. L. Powelson and R. C. Rowe, *Annual review of phytopathology*, 1993, **31**, 111-126.
6. W. R. Stevenson, R. J. Green and G. B. Bergeson, *Plant Disease Reporter*, 1976, **60**, 248–251.
7. J. Ingram and D. Johnson, *American Journal of Potato Research*, 2010, **87**, 382-389.
8. Y. Katan, ed., *Principles in Plant Pathology*, Volcani Center, Bet-Dagan, Israel, 1998.
9. G. N. Agrios, *Plant pathology*, Academic Press, Orlando, 1988.
10. A. K. Lees and A. J. Hilton, *Plant Pathology*, 2003, **52**, 3-12.
11. D. A. Johnson, *Plant disease*, 1994, **78**, 1075-1078.
12. L. Tsror, M. Aharon and O. Erlich, *Phytoparasitica*, 1999, **27**, 215-226.
13. M. Sariah, *Biotrop Spec. publ.* , 1994, **54**, 103-120.
14. P. J. Read and G. A. Hide, *Annals of applied biology*, 2008, **126**, 437-447.
15. A. W. Barkdoll, J.R. Davis, *Plant Disease*, 1992, **76**, 131–135.
16. L. Tsror, *Plant Pathology*, 2004, **53**, 288-293.
17. G.-H. Kim, J.-J. Kim, Y. W. Lim and C. Breuil, *Canadian Journal of Botany*, 2005, **83**, 272-278.
18. U. Moreth and O. Schmidt, *Holzforschung* 2005, **59**, 90-93.
19. M. R. Bonde, G.L. Peterson, J.L. Maas, *Phytopathology*, 1991, **81**, 1523–1528.
20. S. E. Maddison, *Clinical microbiology reviews*, 1991, **4**, 457-469.
21. A. F. C. da Silva, M. L. Rodrigues, S. E. Farias, I. C. Almeida, M. R. Pinto and E. Barreto-Bergter, *FEBS letters*, 2004, **561**, 137-143.
22. R. K. Saiki, S. Scharf, F. Faloona, K. B. Mullis, G. T. Horn, H. A. Erlich and N. Arnheim, *Science*, 1985, **230**, 1350-1354.
23. C. Torres-Calzada, R. Tapia-Tussell, A. Quijano-Ramayo, R. Martin-Mex, R. Rojas-Herrera, I. Higuera-Ciapara and D. Perez-Brito, *Molecular biotechnology*, 2011, **49**, 48-55.
24. S. Nikkari and D. A. Relman, *Current opinion in rheumatology*, 1999, **11**, 11-16.
25. N. C. Clark, O. Olsvik, J. M. Swenson, C. A. Spiegel and F. C. Tenover, *Antimicrobial agents and chemotherapy*, 1999, **43**, 157-160.
26. M. Vaneechoutte and J. Van Eldere, *Journal of medical microbiology*, 1997, **46**, 188-194.
27. E. B. Wilson, J. C. Decius and P. C. Cross, *Molecular vibrations : the theory of infrared and Raman vibrational spectra*, Dover Publications, New York, 1980.
28. J. F. Leslie, *Annual review of phytopathology*, 1993, **31**, 127-150.
29. T. Katan, *Phytoparasitica*, 1999, **27**, 51-64.
30. A. M. Nogales Moncada, R. M. Jiménez Díaz and E. Pérez Artés, *Journal of Phytopathology*, 2009, **157**, 729-735.
31. A. Somrith, N. Singburaudom and O. Piasai, *Kasetsart J. (Nat. Sci.)*, 2011, **45**, 451 - 460.
32. N. Korolev, J. Katan and T. Katan, *Phytopathology*, 2000, **90**, 529-536.
33. N. Nitzan, L. Tsror and D. A. Johnson, *Plant disease*, 2006, **90**, 1287-1292.
34. N. Nitzan, M. Hazanovsky, M. Tal and L. Tsror, *Phytopathology*, 2002, **92**, 827-832.
35. L. Tsror, M. Hazanovsky, S. Mordechi-Lebiush and S. Sivan, *Plant Pathology*, 2001, **50**, 477-482.
36. M. Fiers, V. Edel-Hermann, C. Chatot, Y. Hingrat, C. Alabouvette and C. Steinberg, *Agron. Sustain. Dev.*, 2012, **32**, 93-132.

37. S. Shcolnick, A. Dinoor and L. Tsror, *Plant Disease*, 2007, **91**, 805-808.

38. A. Salman, I. Lapidot, A. Pomerantz, L. Tsror, Z. Hammody, R. Moreh, M. Huleihel and S. Mordechai, *Spectroscopy: An International Journal*, 2012, **27**, 551-556.

39. A. Salman, I. Lapidot, A. Pomerantz, L. Tsror, E. Shufan, R. Moreh, S. Mordechai and M. Huleihel, *Journal of biomedical optics*, 2012, **17**, 017002.

40. A. Salman, A. Pomerantz, L. Tsror, I. Lapidot, R. Moreh, S. Mordechai and M. Huleihel, *The Analyst*, 2012, **137**, 3558-3564.

41. A. Salman, L. Tsror, A. Pomerantz, R. Moreh, S. Mordechai and M. Huleihel, *Spectroscopy: An International Journal*, 2010, **24**, 261-267.

42. M. J. Gupta, J. M. Irudayaraj, C. Debroy, Z. Schmilovitch and A. Mizrach, *Transactions of the ASABE*, 2005, **48**, 1889-1892.

43. H. Lamprell, G. Mazerolles, A. Kodjo, J. F. Chamba, Y. Noël and E. Beuvier, *International Journal of Food Microbiology*, 2006, **108**, 125-129.

44. G. Fischer, S. Braun, R. Thissen and W. Dott, *Journal of microbiological methods*, 2006, **64**, 63-77.

45. V. Shapaval, T. Moretro, H. P. Suso, A. W. Asli, J. Schmitt, D. Lillehaug, H. Martens, U. Bocker and A. Kohler, *Journal of biophotonics*, 2010, **3**, 512-521.

46. F. L. Martin, J. G. Kelly, V. Llabjani, P. L. Martin-Hirsch, Patel, II, J. Trevisan, N. J. Fullwood and M. J. Walsh, *Nature protocols*, 2010, **5**, 1748-1760.

47. M. A. Mackanos, J. Hargrove, R. Wolters, C. B. Du, S. Friedland, R. M. Soetikno, C. H. Contag, M. R. Arroyo, J. M. Crawford and T. D. Wang, *Journal of biomedical optics*, 2009, **14**, 044006.

48. M. J. Walsh, M. N. Singh, H. M. Pollock, L. J. Cooper, M. J. German, H. F. Stringfellow, N. J. Fullwood, E. Paraskevaidis, P. L. Martin-Hirsch and F. L. Martin, *Biochemical and biophysical research communications*, 2007, **352**, 213-219.

49. M. Khanmohammadi, A. B. Garmarudi, K. Ghasemi, H. K. Jaliseh and A. Kaviani, *Medical oncology*, 2009, **26**, 292-297.

50. A. Salman, E. Shufan, L. Tsror, R. Moreh, S. Mordechai and M. Huleihel, *Methods*, 2014.

51. B. Ben-Daniel, D. Bar-Zvi, D. Johnson, R. Harding, M. Hazanovsky and L. Tsror Lahkim, *Phytopathology*, 2010, **100**, 271-278.

52. C. L. Curl, C. J. Bellair, T. Harris, B. E. Allman, P. J. Harris, A. G. Stewart, A. Roberts, K. A. Nugent and L. M. Delbridge, *Cytometry. Part A : the journal of the International Society for Analytical Cytology*, 2005, **65**, 88-92.

53. E. Bogomolny, M. Huleihel, A. Salman, A. Zwielly, R. Moreh and S. Mordechai, *The Analyst*, 2010, **135**, 1934-1940.

54. N. J. Harrick, in *Internal reflection spectroscopy*, Harrick Scientific Corporation, Ossining, New York, 1979, pp. 13-66,.

55. J. Trevisan, P. P. Angelov, P. L. Carmichael, A. D. Scott and F. L. Martin, *The Analyst*, 2012, **137**, 3202-3215.

56. R. A. Fisher, *Annals of Eugenics*, 1936, **7**, 179-188.

57. C. M. Bishop, *Pattern recognition and machine learning*, Springer, New York, 2006.

58. F. Camastra and A. Vinciarelli, *Machine learning for audio, image and video analysis : theory and applications*, Springer, London, 2008.

59. R. O. Duda, P. E. Hart and D. G. Stork, *Pattern classification*, Wiley, New York ; Chichester, 2001.

60. J. Brugnerotto, J. Lizardi, F. M. Goycoolea, W. Argüelles-Monal, J. Desbrières and M. Rinaudo, *Polymer*, 2001, **42**, 3569-3580.

61. D. Naumann, D. Helm, H. Labischinski and P. Giesbrecht, *Modern techniques for rapid microbiological analysis*, 1991, 43-96.

62. Z. Movasaghi, S. Rehman and D. I. ur Rehman, *Applied Spectroscopy Reviews*, 2008, **43**, 134-179.

63. N. Fujioka, Y. Morimoto, T. Arai and M. Kikuchi, *Cancer Detection and Prevention*, 2004, **28**, 32-36.

64. Y. Yang, J. Sule-Suso, G. D. Sockalingum, G. Kegelaer, M. Manfait and A. J. El Haj, *Biopolymers*, 2005, **78**, 311-317.

65. L. M. McIntosh, M. Jackson, H. H. Mantsch, M. F. Stranc, D. Pilavdzic and A. N. Crowson, *The Journal of investigative dermatology*, 1999, **112**, 951-956.

66. S. Yoshida, M. Miyazaki, K. Sakai, M. Takeshita, S. Yuasa, A. Sato, T. Kobayashi, S. Watanabe and H. Okuyama, *Biospectroscopy*, 1997, **3**, 281-290.

67. G. I. Dovbeshko, V. I. Chegel, N. Y. Gridina, O. P. Repnytska, Y. M. Shirshov, V. P. Tryndiak, I. M. Todor and G. I. Solyanik, *Biopolymers*, 2002, **67**, 470-486.

68. G. Kos, H. Lohninger and R. Krska, *Analytical chemistry*, 2003, **75**, 1211-1217.

69. I. Lapidot and J.-F. Bonastre, presented in part at the Speech Processing Conference, Tel-Aviv, Israel, June 19-20, 2012, 2012.

70. N. Dehak, P. Kenny, R. Dehak, P. Dumouchel and P. Ouellet, *Audio, Speech, and Language Processing, IEEE Transactions on*, 2011, **19**, 788-798.

71. J. Trevisan, J. Park, P. P. Angelov, A. A. Ahmadzai, K. Gajjar, A. D. Scott, P. L. Carmichael and F. L. Martin, *Journal of biophotonics*, 2014, **7**, 254-265.