Journal of Chromatography, 644 (1993) 259–267
Elsevier Science Publishers B.V., Amsterdam

CHROM. 25 189

Reversed-phase high-performance liquid chromatography and chemometrics, a combined investigation tool for complex phytochemical problems

C. Baiocchi*, E. Marengo, G. Saini, M.A. Roggero and D. Giacosa

Dipartimento di Chimica Analitica, Università di Torino, Via P. Giuria 5, 10125 Turin (Italy)

(First received January 6th, 1993; revised manuscript received April 13th, 1993)

ABSTRACT

The phenolic content in the bark of poplar trees was analysed by RP-HPLC with the aim of finding some evidence of a relationship between the presence of phenols (either the total amount or the amount of an individual specific compound) and the differential resistance to the fungus *Dothichiza populea* that is found in different clones of these trees. Direct comparison of chromatographic results did not allow any useful information to be gleaned on this subject. On the other hand, the application of principal component analysis and linear discriminant analysis methods to the quantitative chromatographic data gave very promising results, allowing discrimination between resistant and susceptible poplars and the identification of phenolic compounds that are important for such discrimination.

INTRODUCTION

A connection between the resistance of poplar trees to infection by the fungus *Dothichiza populea* and the presence in their bark of some phenolic compounds providing a fungistatic activity has already been found [1–3].

The aim of this study was to determine the possibility of classifying genetically controlled hybrids of poplar trees as resistant or susceptible to the fungal infection on the basis of the phenol content of their bark.

To this end, phenolic compounds present in poplar bark were extracted, separated by reversed-phase HPLC and identified by comparing their retention times with those of suitable standards. The chromatograms provided qualitative and quantitative information that was not easily exploitable to produce a definite classification criterion, thus it became necessary to resort to multivariate chemometric treatments, even though the original sampling programme was not formulated with a subsequent statistical treatment in mind.

For the characterization of aromatic natural products, gas chromatographic techniques coupled with such treatments have already proved to be effective in studies on olive oil [4], wine [5,6], coffee [7], tea [8] and honey bees [9]. No report of statistical treatments concerning classification of resistant and susceptible hybrids of poplar trees towards *Dothichiza* infection is currently available. In this paper principal component analysis (PCA) [10,11] and linear discriminant analysis (LDA) [12–14] methods were applied to the chromatographic data concerning poplar hybrids sampled in two different places and at regular intervals between December 1989 and June 1990. Our objective was to discover the most suitable chemical variables to perform an effective classification of poplar trees of different

* Corresponding author.

0021-9673/93/$06.00

infection resistance and to evaluate the influence of seasonal and geographic factors on such a classification.

EXPERIMENTAL

Instrumentation

The HPLC equipment consisted of a Varian 5560 liquid chromatograph equipped with a UV 200 spectrophotometric detector and a 4290 integrator. The detector was operated at 270 nm.

The column was a LiChrospher RP-18 (250 × 4.6 mm I.D.), 5 μm particle size (Merck, Darmstadt, Germany). The injection was 10 μl, and the flow-rate 1.0 ml/min.

Reagents

HPLC-grade methanol and acetonitrile and a 0.57% solution of acetic acid in Millipore Milli-Q water were used as mobile phase constituents.

Chromatographic conditions

A gradient programme based on a ternary mobile phase gave the best results: acetonitrile (A), acetic acid (0.57%) water solution (B) and methanol (C). The starting conditions were 6% A, 88% B, 6% C. At 40 min the eluent composition was 6% A, 48% B and 46% C.

Sampling

The bark of three clones known to differ in resistance to *Dothichiza populea*—(a) *S. MARTINO* (S.M., resistant), (b) *LUISA AVANZO* (L.A., susceptible) and (c) I-214 (intermediate) —were sampled, in duplicate, in two different places in Italy (Casale Monferrato and Scottine). The sampling programme started in December 1989 and continued monthly until June 1990. The total number of samples collected was 96: two samples of each clone from each geographic site for eight different sampling periods.

Sample preparation

The release of phenolic compounds from poplar bark and the preparation of solutions for analytical HPLC were HPLC were accomplished by performing the following procedure. About 30 mg of previously liophilized bark were added to 2.0 ml of 1.0 *M* sodium hydroxide in a filter tube. The air was removed from the tube by flushing with nitrogen and the stopper secured. The suspension was shaken at 20°C for 20 h and subsequently filtered. The residue was washed with water (total volume of filtrate *ca.* 2.0 ml). The filtrate was acidified to pH 2.5 with 6.0 *M* hydrochloric acid and diluted with water to a final volume of 5.0 ml [1].

Standards

As standard substances we used the phenolic compounds more frequently proposed in previous works [2–4] as involved in the mechanism of defence against disease in poplar trees. They were: 4-hydroxy-3-methoxybenzaldehyde, 4-hydroxy-3,5-dimethoxybenzaldehyde, 4-hydroxybenzaldehyde, 4-hydroxybenzoic acid, 4-hydroxy-3-methoxycinnamic acid, 4-hydroxycinnamic acid, benzoic acid, 2-hydroxybenzoic acid, 3,5-dimetoxy-4-hydroxybenzoic acid, cathecol, pyrogallol, 3,4-dihydroxybenzoic acid, 4-hydroxy-3-methoxybenzoic acid and 4-methoxybenzoic acid (Fig. 1).

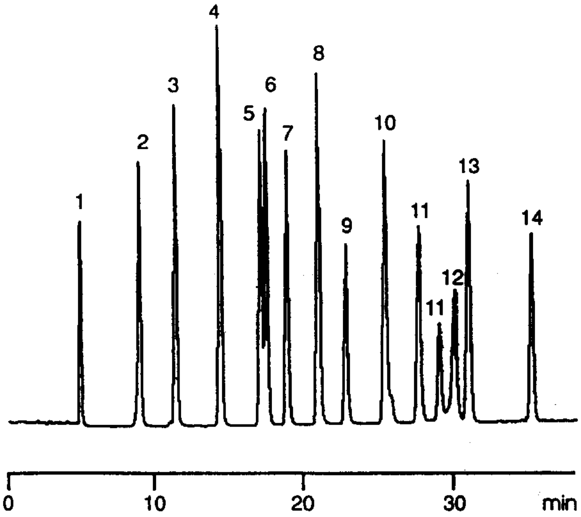

Fig. 1. Chromatographic separation of phenolic standard substances. (For elution conditions, see Experimental section.) Peaks: 1 = 4-hydroxy-3-methoxybenzaldehyde; 2 = 4-hydroxy-3,5-dimethoxybenzaldehyde; 3 = 4-hydroxybenzaldehyde; 4 = 4-hydroxybenzoic acid; 5 = 4-hydroxy-3-methoxycinnamic acid; 6 = 4-hydroxycinnamic acid; 7 = benzoic acid; 8 = salicylic acid; 9 = 4-hydroxy-3,5-dimethoxybenzoic acid; 10 = cathecol; 11 = pyrogallol; 12 = 3,4-dihydroxybenzoic acid; 13 = 4-hydroxy-3-methoxybenzoic acid; 14 = 4-methoxybenzoic acid.

Chemometric methods

The data was analysed by PCA and LDA. The former is a pattern recognition method that generates new orthogonal variables, the principal components (PCs), linear combination of the original variables, so that the maximum possible amount of variance of the data is compressed in few PCs. In fact, it is theoretically possible to determine as many principal components as original variables, however they are obtained in order of decreasing contribution to the total variance, so it is usually sufficient to consider the first principal components and still retain most of the variance, that is most of information present in the original data.

PCA is a useful tool to perform variable reductions for high-dimensional complex problems. Moreover, the analysis of the new variable space (PCs space) often provides important information on the pattern of the data (*i.e.* clusters, systematic trends, etc.).

In LDA, as in the PCA technique, the aim is to reduce the number of features. However, while PCA selects a direction that retains maximal structure in a lower dimension among the data, LDA selects a direction that achieves maximum separation among the given classes. The discriminant function obtained in this way leads to a new variable which, as in principal components, is a linear combination of the original variables.

Once this direction has been found, it is possible to perform a statistical test of the significance of the separation of the groups two by two and to assign new objects to any of the two classes.

In the study we also applied a new classification algorithm that generates orthogonal discriminant directions, the orthogonal discriminant analysis (ODA) [15]. This algorithm is advantageous since it allows correct graphic visualization of the discriminant directions. This graphic visualization generally cannot be achieved with the classical LDA algorithm, whose discriminant directions are not required to be orthogonal one to each other.

The data were autoscaled before any PCA treatment in order to attribute the same *a priori* importance to the variables.

RESULTS AND DISCUSSION

Fig. 1 shows the separation of fourteen standard phenolic compounds, whereas Fig. 2 reports typical chromatographic runs regarding the analyses of the bark extracts of the resistant (a and c) and susceptible (b and d) clones. They refer to the first and the last sampling periods December and June, respectively. By comparing the chromatographic profiles obtained from the winter samples (Fig. 2a and b), definite quantitative differences between the clones can be seen. The differences are apparently remarkably reduced in the late spring samples (Fig. 2c and d). Chromatographic runs of the clone of the intermediate resistance (I-214) are not reported because they were not significantly different from those of the susceptible one (L.A.). Ten peaks identified by comparison with the retention times of the pure standards indicated in the chromatograms were used in the chemometric study. They correspond to compounds 2–11 of Fig. 1. A small number of other peaks remain unknown.

By examining the area variations of the identified peaks among the three different clones, no definite trends in the result could be found by visual inspection. So, in order to set up a useful correlation between phenolic contents and infection resistance of the poplar hybrids considered, a chemometric study was performed, though the sampling design was not programmed with a subsequent statistical treatment in mind. The chromatographic peak areas are not reported here because their number is very high.

Since there was a large sampling variation, in the total amount of phenolic compounds, owing to the various combinations of clone, geographic origin and date, but with an apparent conservation of the proportion between the areas, the calculations were performed using the percentage areas. The starting data set contained 96 samples described by ten variables each (the phenolic compounds identified and listed in Fig. 1 from 2 to 11).

The starting analysis was conducted on the whole data set containing samples of different origin, sampling period and genetic origin which, in principle, could significantly affect the phenol

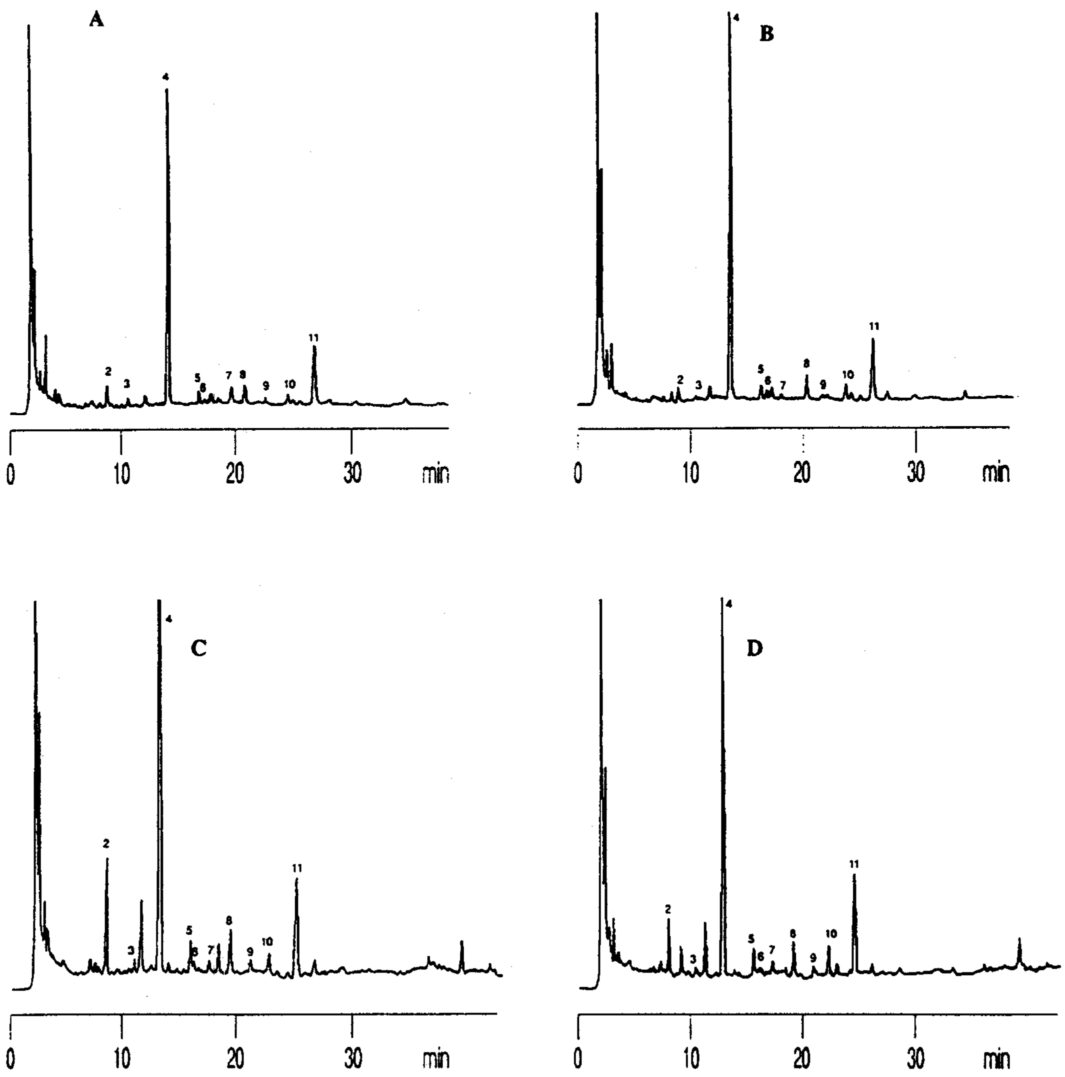

Fig. 2. Chromatographic separations of real samples (resistant and susceptible clones) at the start and the end of the sampling period. (For elution conditions, see Experimental section.) (A) Luisa Avanzo (Dec. 1989), 0.0312 AUFS; (B) S. Martino (Dec. 1989), 0.0312 AUFS; (c) Luisa Avanzo (June 1990), 0.0156 AUFS; (D) S. Martino (June 1990), 0.0156 AUFS. Peak numbers as in Fig. 1.

composition of the poplar bark extracts. After autoscaling, the complete set of the original data was subjected to PCA. The plot of the scores of the first PCs allowed the detection of two outliers (*i.e.* two anomalous samples far from any other sample), which were eliminated from the successive analysis. The new data set containing 94 samples was again autoscaled and submitted to PCA. The variance explained by the first four PCs (eigenvalue greater or around 1.0) is listed in Table I, while Table II reports the loadings of

TABLE I

VARIANCE EXPLAINED BY THE FIRST FOUR PCs (CLONES 1, 2 AND 3)

PC	Variance (%)	Total variance (%)
1	40.2	40.2
2	14.9	55.1
3	10.7	65.8
4	9.7	75.5

TABLE II

LOADINGS OF THE FIRST FOUR PCs (CLONES 1, 2 AND 3)

Compounds	PC_1	PC_2	PC_3	PC_4
4-Hydroxy-3,5-dimethoxybenzaldehyde	0.35	−0.02	−0.21	−0.46
4-Hydroxybenzaldehyde	0.28	−0.39	−0.26	−0.12
4-Hydroxybenzoic acid	−0.47	−0.22	0.11	0.01
4-Hydroxy-3-methoxycinnamic acid	0.38	−0.21	−0.07	−0.12
4-Hydroxycinnamic acid	0.30	−0.23	0.38	0.13
Benzoic acid	0.07	−0.27	−0.42	0.78
Salicylic acid	0.33	0.13	0.50	0.30
4-Hydroxy-3,5-dimethoxybenzoic acid	0.26	−0.43	0.24	−0.04
Cathecol	0.26	0.33	−0.47	0.04
Pyrogallol	0.30	0.55	0.12	0.17

the same PCs, *i.e.* the contribution of every original variable to the definition of the PCs. The scatter plots of the samples projected on these PCs show that the PCs containing information about the difference between the clones are the second and the fourth ones. A plot of PC_2 and PC_4 presents two clusters of objects corresponding to the two clones, S.M. and L.A., slightly overlapping in the middle, and a third cluster (clone I-214) largely overlapping with the other two (Fig. 3). Since the main interest lies in the separation of resistant and susceptible clones, PCA was performed on a reduced data set containing only the hybrids S.M. and L.A. (classes 1 and 3).

The variance explained by the first four PCs (eigenvalue greater or around 1.0) is given in Table III and the corresponding loadings in Table IV. The scatter plot of the scores of PC_2 and PC_4 is shown in Fig. 4. The situation is clearer and there is an improvement in the separation between the two classes, because of the elimination of the data from the samples of the intermediate resistance clone.

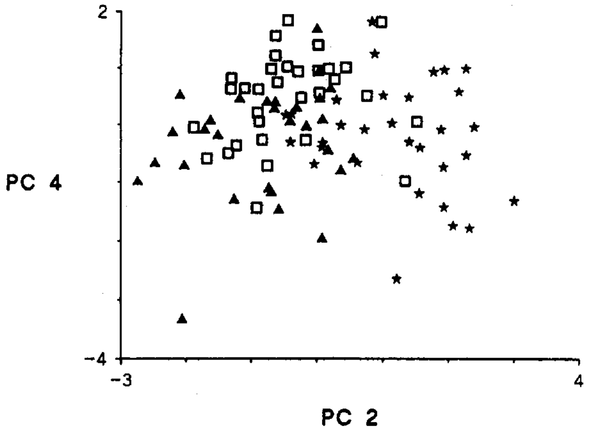

Fig. 3. Scatter plot of the scores of the objects (three types of clones) on the PCs. ★ = S. Martino; ▲ = 214; □ = Avanzo.

In order to better isolate the information traceable back to a seasonal factor, we broke down the previous general plot into a sequence of eight distinct plots in order of sampling time (Fig. 5A and B). Every plot is updated with respect to the preceding one according to the sequence of sampling period, *i.e.* it shows the preceding samples (open markers) plus the samples of the updated period (black markers). The sample origin was not considered in this step since it was found to be irrelevant in PC_2 and PC_4. As can be seen from Fig. 5, the separation of the two clones remains satisfactory for almost all sampling times. Only the data of the late

TABLE III

VARIANCE EXPLAINED BY THE FIRST FOUR PCs (CLONES 1 AND 3)

PC	Variance (%)	Total variance (%)
1	37.41004	37.41004
2	18.27317	55.68321
3	10.54750	66.23071
4	8.550346	74.78105

TABLE IV

LOADINGS OF THE FIRST FOUR PCs (CLONES 1 AND 3)

Compounds	PC_1	PC_2	PC_3	PC_4
4-Hydroxy-3,5-dimethoxybenzaldehyde	0.20	−0.59	−0.25	0.01
4-Hydroxybenzaldehyde	0.22	−0.17	0.72	0.00
4-Hydroxybenzoic acid	0.09	0.60	0.11	−0.32
4-Hydroxy-3-methoxycinnamic acid	0.39	−0.15	0.10	−0.06
4-Hydroxycinnamic acid	0.34	0.24	0.19	−0.06
Benzoic acid	0.26	0.09	0.14	0.83
Salicylic acid	0.46	−0.02	−0.24	−0.05
4-Hydroxy-3,5-dimethoxybenzoic acid	0.23	−0.27	0.30	−0.42
Cathecol	0.37	0.29	−0.11	0.08
Pyrogallol	0.41	0.05	−0.43	−0.13

spring tend to mix together, as expected. It is worth noticing that, within each sampling time, the discrimination between the two clones is very good.

However, as sampling time proceeds there is a progressive shift in the relative and absolute positions of the objects along the axes of the diagrams. Such a shift should be caused by changes in the extract composition as the season advances. In fact, the change is not random and can be explained by considering the composition of the two PCs considered. In particular, by taking into account the relative movements of the two clones, it can be observed that there are shifts along the axes attributable to changes, equal for both clones, in the percentage of the chemical components that mainly define the two PCs. At the same time there are rotations of the relative positions, which may be interpreted as different behaviours of the same chemical compounds in the two clones.

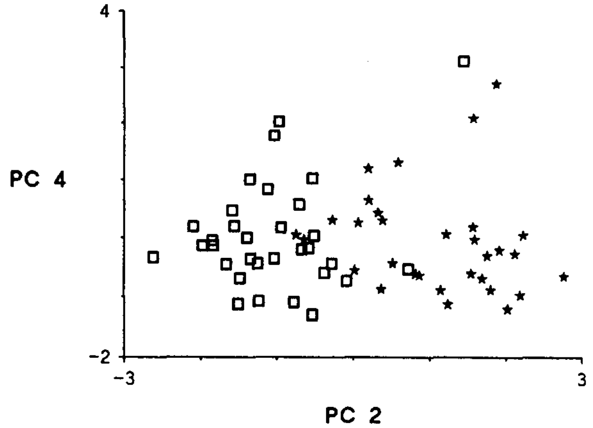

Fig. 4. Scatter plot of the scores of the reduced space of the objects (only resistant and susceptible clones) on the PCs. ★ = S. Martino; □ = Avanzo.

The most discriminant principal component is PC_2, which is mainly constituted by the variables corresponding to 4-hydroxy-3,5-dimethoxybenzaldehyde and 4-hydroxybenzoic acid, while PC_4 is mainly constituted by benzoic acid and 4-hydroxy-3,5-dimethoxybenzoic acid.

Very little information is available to achieve a geographic discrimination. It is apparent that the geographic variable poorly affects the characterization of the clones, the main difference being the genetic diversity and secondly the sampling period effect.

After the PCA treatment, which provides information on the pattern of the data and on the seasonal changes in the composition of the samples, we applied LDA to find the best directions for classifying the samples.

LDA was applied in turn to separate each of the clones from the others. This provided three discriminant directions. Table V reports the statistical features of the three separations and the normalized composition of the discriminant directions. The angle between the discriminating directions, easily determined from the scalar product of the vectors along the directions, is 66.6° between $R_{1,2}$ and $R_{1,3}$, 132.6° between $R_{1,2}$ and $R_{2,3}$ and 76.7° between $R_{1,3}$ and $R_{2,3}$. These

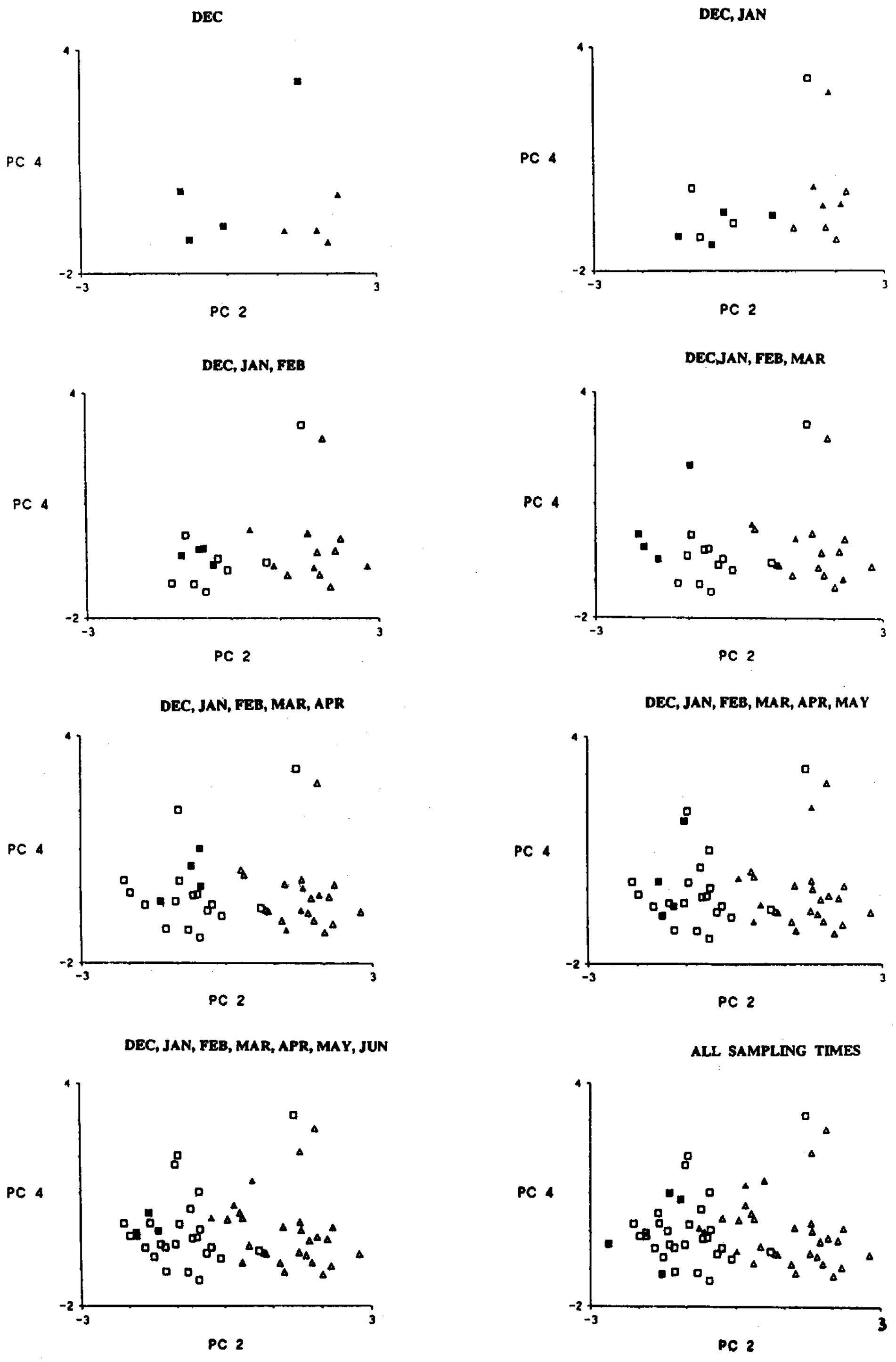

Fig. 5. Scatter plots of the reduced set of the objects of the various sampling periods in PCs. △ = S. Martino; □ = Avanzo.

TABLE V

STATISTICAL FEATURES OF LDA SEPARATIONS AND NORMALIZED COMPOSITION OF DISCRIMINATING DIRECTIONS

Discrimination between classes 1 and 2: Mahalanobis distance: 16.20, $F_{10,49} = 20.51$
Discrimination between classes 2 and 3: Mahalanobis distance: 10.60, $F_{10,50} = 13.67$
Discrimination between classes 1 and 3: Mahalanobis distance: 20.60, $F_{10,52} = 20.60$

Compounds	Autoscaled director cosines:	Autoscaled director cosines:	Autoscaled director cosines:
4-Hydroxy-3,5-dimethoxybenzaldehyde	−0.60	0.16	−0.89
4-Hydroxybenzaldehyde	−0.02	0.12	0.06
4-Hydroxybenzoic acid	0.06	0.27	0.23
4-Hydroxy-3-methoxycinnamic acid	−0.34	−0.03	−0.09
4-Hydroxycinnamic acid	0.15	0.02	0.01
Benzoic acid	0.02	0.10	0.05
Salicylic acid	0.59	−0.58	−0.13
4-Hydroxy-3,5-dimethoxybenzoic acid	0.12	−0.11	0.13
Cathecol	−0.37	0.71	0.32
Pyrogallol	−0.06	−0.13	−0.03

directions are shown independently in Fig. 6, from which it can be seen that all classifications are satisfactory. The least separated classes are 2 and 3, for which the F-test provides a value of 13.67, which is still highly significant ($F_{50,10} = 4.2$ at the 99% confidence level); this confirms the results of the PCA step.

LDA was also applied to the separation of the samples of different geographic origin, but a very poor classification was obtained.

The application of a different LDA algorithm, namely the ODA algorithm, which provides orthogonal discriminant directions, led to the results shown in Fig. 7. It is clear that all three clones can be effectively separated on the basis of the chemical composition of the extracts.

Summing up, it was impossible to discriminate between the different clones by a direct visual inspection of the raw data alone, these being too complex. The discriminant features obtained by the chemometric treatments are not single original variables but linear combinations of variables. It is evident that the use of some multivariate statistical tools greatly improved the possibility of analysing the data structure and performing an effective classification of the clones with different resistance to fungal infection.

In conclusion, it is worth noting that such

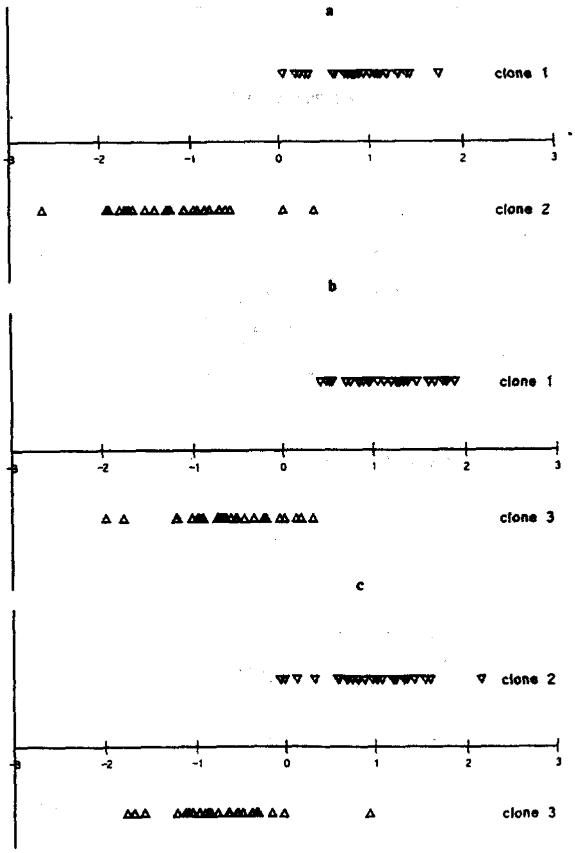

Fig. 6. Classification ability of the three clones by the discriminant directions determined by the LDA procedure. (a) Direction $R_{1,2}$; (b) direction $R_{1,3}$; (c) direction $R_{2,3}$.

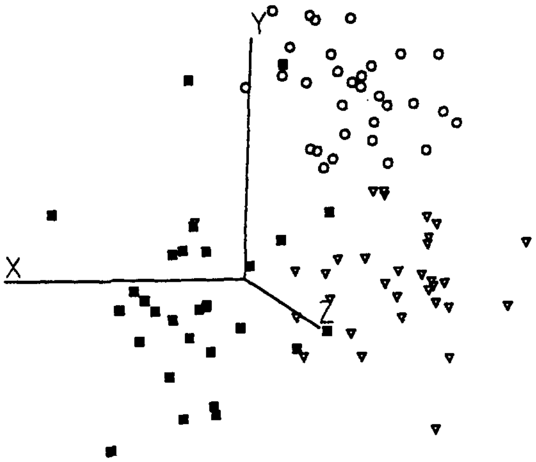

Fig. 7. Plot of the objects in the space of the three discriminant directions obtained by a different LDA algorithm.

information was obtained from data not particularly suitable for treatment with the method adopted. A new sampling plan much more tailored on the statistical requirements is almost accomplished and a further chromatographic and chemometric study is in progress in our laboratory.

REFERENCES

1 S. Pukacka, *Arbor. Korn.*, 20 (1975) 227.
2 S. Pukacka, *Arbor. Korn.*, 25 (1980) 257.
3 Y. Chuanhe, Y. Wang and Z. Zhongming, presented at *International Poplar Commission, XVIIIth Session, Beijing, September 5–8, 1988.*
4 M. Forina and E. Tiscornia, *Ann. Chim. (Rome)*, 72 (1982) 143.
5 I.E. Frank, B.R. Kowalski, *Anal. Chim. Acta*, 162 (1984) 241.
6 W.E. Kwan and B.R. Kowalski, *J. Food Sci.*, 43 (1978) 1320.
7 K. Wada, S. Oghama, H. Sasaki and M. Shimoda, *Agr. Biol. Chem.*, 51 (1987) 1745.
8 L. Xiande, P. Van Espen and F. Adams, *Anal. Chim. Acta*, 200 (1987) 421.
9 R. Valcarce and G.G. Smith, *Chemom. Intell. Lab. Syst.*, 6 (1989) 157.
10 J.C. Davis, *Statistics and Data Analysis in Geology*, Wiley, New York, 2nd ed., 1986, p. 524.
11 S. Wold, *Technometrics*, 20 (1978), 397.
12 D.L. Massart, B.G.M. Vandeginste, S.N. Deming, Y. Michotte and L. Kaufman, *Chemometrics: a Textbook*, Elsevier, Amsterdam, 1988, p. 330.
13 K.V. Mardia, J.T. Kent and J.M. Bibby, *Multivariate Analysis*, Academic Press, London, 1979.
14 M. James, *Classification Algorithms*, Collins, London, 1985.
15 E. Marengo and R. Todeschini, presented at *X congresso Nazionale della Divisione di Chimica Analitica della Societá Chimica Italiana, Turin, Sept. 16–19, 1991.*