Journal of Chromatography, 466 (1989) 371–378
Elsevier Science Publishers B.V., Amsterdam — Printed in The Netherlands

CHROM. 21 177

Note

Cluster analysis in the comparison of two-dimensional chromatograms

DANIELA HEIMLER*
Dipartimento di Scienza del Suolo e Nutrizione della Pianta, Università degli Studi di Firenze, Piazzale delle Cascine 28, Florence (Italy)
and
VIERI BODDI
Istituto di Patologia Generale, Università degli Studi di Firenze, Viale G.B. Morgagni 50, Florence (Italy)
(Received December 5th, 1988)

Two-dimensional chromatography, notwithstanding its undoubted advantages from a theoretical point of view, is the least studied separation technique owing to the difficulty in interpretating the experimental results. The technique has two main drawbacks: the possibility of analysing only one sample on one sheet without the simultaneous spotting of test compounds[1] and the difficult interpretation of a two-dimensional chromatogram on the basis of R_F values obtained by one-dimensional developments. Even the quantitative analysis of two-dimensional chromatograms is difficult, as the spots are not arranged in vertical strips but occupy the whole layer. Nevertheless two-dimensional chromatography can separate very complex mixtures which are difficult to resolve by means of other techniques[2].

This paper describes a method which allows the use of chromatographic data in order to calculate similarity criteria without having quantitative data. The problem of sample classification on the basis of chromatographic results is generally carried out by means of pattern recognition techniques using the quantitative data from gas chromatography or high-performance liquid chromatography[3].

EXPERIMENTAL

Flavonoids aglycones of elm and iris leaves were examined. The crushed leaves (1 g) were treated with 25 ml of boiling hydrochloric acid for 30 min. The flavonoids were extracted with 10 ml of ethyl acetate, which was evaporated to dryness under vacuum. The residue was dissolved in 2 ml of methanol and 2 μl of this solution were spotted on Sil C_{18}-50 plates (Macherey, Nagel & Co.) and eluted in the first direction with *n*-hexane–ethyl acetate–acetic acid (72:27:1) and in the second direction with 1 *M* acetic acid in 50% methanol. The spots were sprayed with a 1% methanolic solution of ethanolamine diphenylborate and a 5% ethanolic solution of polyethylene glycol. The spots were observed under UV light (360 nm). Under these conditions the flavonoids give fluorescent spots of different colours. The spots were characterized by their positions on the layer by means of two coordinates, obtained by dividing the

0021-9673/89/$03.50

distance of the spot from the origin lines by the distance of the two solvents from the same lines, and by their colour under UV light.

For elm leaves we considered only those compounds which migrate in both eluents, because compounds which remain at the starting point with the first eluent and migrate with the second eluent are difficult to identify owing to their incomplete separation. In this way every spot in all of the chromatograms were assigned to a definite group and the results in Table I were obtained. It should be noted that only in a few instances could the spots be identified[4]. The aim of this work, however, was to compare several two-dimensional chromatograms, regarded as "fingerprints" of different plants, in order to ascertain whether the thin-layer chromatographic (TLC) profile of phenolic compounds could be of help in the determination of differences among populations, provenances and species (that is, intra- and inter-specific differences).

The elm leaves were obtained from the germoplasm collection of the Centre for Forest Pathology of the National Research Council of Florence. The data for the two-dimensional chromatograms are reported in Table II.

The iris leaves belong to spontaneous species and were sampled in the Giardino dell'Iris in Florence. In this instance we also considered the compounds lying on the *y*-axis (that is, those compounds which migrate with the first eluent but remain at the origin with the second), as they are better characterized than in the case of elm leaves. The data are reported in Tables III and IV.

TABLE I

COORDINATES AND COLOURS UNDER UV LIGHT OF ALL THE SPOTS OBSERVED IN THE TWO-DIMENSIONAL CHROMATOGRAMS OF ELM LEAVES

The values of the coordinates are the means of 10–32 determinations. Letters A–U represent the different spots in the chromatograms.

	Coordinates × *100*	*Colour*	*Name*
A	18–18	Orange	Quercetin
B	12–51	Red	–
C	13–37	Red	Myricetin
D	25–39	Red	–
E	13–50	Orange	–
F	29–74	Light blue	–
G	31–53	Light blue	–
H	39–70	Light blue	–
I	24–84	Light blue	–
L	29–61	Light blue	–
M	25–67	Yellow	Caffeic acid
N	25–75	Yellow	Caffeic acid
O	44–45	Yellow	–
P	14–48	Yellow	–
Q	19–55	Yellow	–
R	30–10	Green yellow	Kaempferol
S	23–35	Light blue	–
T	63–81	Light blue	–
U	42–82	Blue	–

TABLE II
DISTRIBUTION OF SPOTS IN THE ELM LEAVES
A–U as in Table I.

Elm leaves	*A*	*B*	*C*	*D*	*E*	*F*	*G*	*H*	*I*	*L*	*M*	*N*	*O*	*P*	*Q*	*R*	*S*	*T*	*U*
U. pumilia:																			
(1) S1	+	+	+			+	+	+	+	+	+					+			
(2) S12	+		+				+		+		+					+			
(3) S15	+						+	+	+		+					+			
(4) PU1	+	+	+		+		+	+	+		+	+	+			+			
(5) 73P	+	+	+			+	+	+	+		+	+				+	+		
(6) 182P	+	+	+				+	+	+		+	+	+	+	+	+			
U. parvifolia:																			
(7) PA1.1, PA1.2	+					+	+		+		+	+				+			+
(8) PA2	+		+				+		+		+	+		+		+			
(9) 157P	+	+	+	+			+		+		+	+	+			+			+
(10) NA33	+		+				+				+	+	+	+		+			
U. japonica:																			
(11) 3P	+	+	+		+		+	+	+		+	+				+			
(12) 2P	+	+	+	+	+		+	+	+		+	+				+			
(13) 127P	+	+	+		+	+	+	+	+		+	+	+			+			
(14) 23P	+	+	+	+	+	+	+		+		+	+	+			+			
(15) 57P	+	+	+	+			+		+	+	+	+	+	+		+			+
U. carpinifolia:																			
(16) C3	+	+	+				+	+			+	+	+			+			
(17) C6	+	+	+			+	+	+			+	+	+			+			
(18) 6-11	+	+	+			+	+	+			+	+				+			
U. xhollandica:																			
(19) 274, P38, 275	+		+				+		+	+	+	+	+	+		+	+		+
(20) 405	+		+	+			+		+	+	+	+	+	+	+	+			+
U. chemnoui:																			
(21) 176P.2	+	+	+		+	+	+	+	+	+	+	+	+	+	+	+			
(22) 176P.5	+	+	+		+		+	+	+		+	+	+	+	+	+			
U. villosa:																			
(23) VII,P554	+						+		+		+	+		+		+			+
(24) *U. laevis*	+		+				+		+		+	+		+		+	+		+
(25) *U. glabra*	+		+				+		+		+	+		+	+	+			+
(26) *U. elliptica*	+		+	+			+	+	+		+	+		+		+			
(27) *U. laciniata*	+		+				+		+		+	+		+	+	+			
(28) *U. wilsoniana*	+	+	+				+	+	+	+	+	+				+		+	+

RESULTS AND DISCUSSION

The question of the comparison of qualitative data regarded as indicating the presence or absence of a compound can be solved by means of numerical indices expressed by equations which may change slightly from one case to another. The information in each line in Tables II and IV can be codified as either 0 or 1 (absence or presence of a spot). The result of the comparison of two sequences is characterized by four values: N_{11} (number of positive agreements), N_{00} (number of negative agreements), N_{10} and N_{01} (number of disagreements, that is, the presence of a spot in

TABLE III

COORDINATES AND COLOURS UNDER UV LIGHT OF ALL THE SPOTS OBSERVED IN THE TWO-DIMENSIONAL CHROMATOGRAMS OF IRIS LEAVES

	Coordinates × *100*	*Colour*
A	9–27	Green
B	12–18	Green
C	17–10	Green
D	18–64	Light blue
E	18–77	Orange
F	24–0	Red
G	25–60	Light blue
H	30–10	Light blue
I	30–83	Blue
L	51–0	Red
M	56–0	Red
N	76–0	Red
O	12–15	Green
P	86–60	Light blue
Q	18–69	Light blue
R	37–49	Light blue
S	65–28	Light blue
T	51–73	Light blue

TABLE IV

DISTRIBUTION OF SPOTS IN THE IRIS LEAVES

A–T as in Table III.

Iris leaves	*A*	*B*	*C*	*D*	*E*	*F*	*G*	*H*	*I*	*L*	*M*	*N*	*O*	*P*	*Q*	*R*	*S*	*T*
(1) *I. pallida*[a]				+	+	+	+		+	+		+						
(2) *I. pallida*[b]				+	+	+	+		+	+		+			+			
(3) *I. pallida*				+	+	+			+	+		+			+			
(4) *I. cengialti*	+	+	+	+		+			+	+		+		+	+			
(5) *I. florentina*	+	+	+	+	+	+			+	+	+	+			+			
(6) *I. germanica*	+		+	+	+	+			+	+		+						+
(7) *I. lutescens (Quercianella)*		+	+			+				+	+	+						
(8) *I. lutescens (Monte Marcello)*	+	+	+			+				+	+	+					+	
(9) *I. squalens*	+	+	+	+	+	+	+		+	+	+	+				+		
(10) *I. kockii*			+	+		+			+	+	+	+	+		+			
(11) *I. sambucina*	+	+	+	+	+	+			+	+		+			+			
(12) *I. aphilla*	+	+	+			+			+	+			+	+	+			
(13) *I. uinguicularis*	+	+	+	+	+	+	+	+	+	+	+	+	+					

[a] Fertile form.
[b] Sterile form.

one sequence and its absence in the other). The most general similarity index of two sequences (simple matching coefficient) is[5]

$$S_{SM} = (N_{11} + N_{00})/(N_{11} + N_{00} + N_{10} + N_{01})$$

The Jaccard–Sneath coefficient does not consider the negative agreements (N_{00}):

$$S_{JS} = N_{11}/(N_{11} + N_{10} + N_{01})$$

From the point of view of our data (TLC data), we deemed the Jaccard–Sneath coefficient to be more useful, as TLC can give information on the presence of one

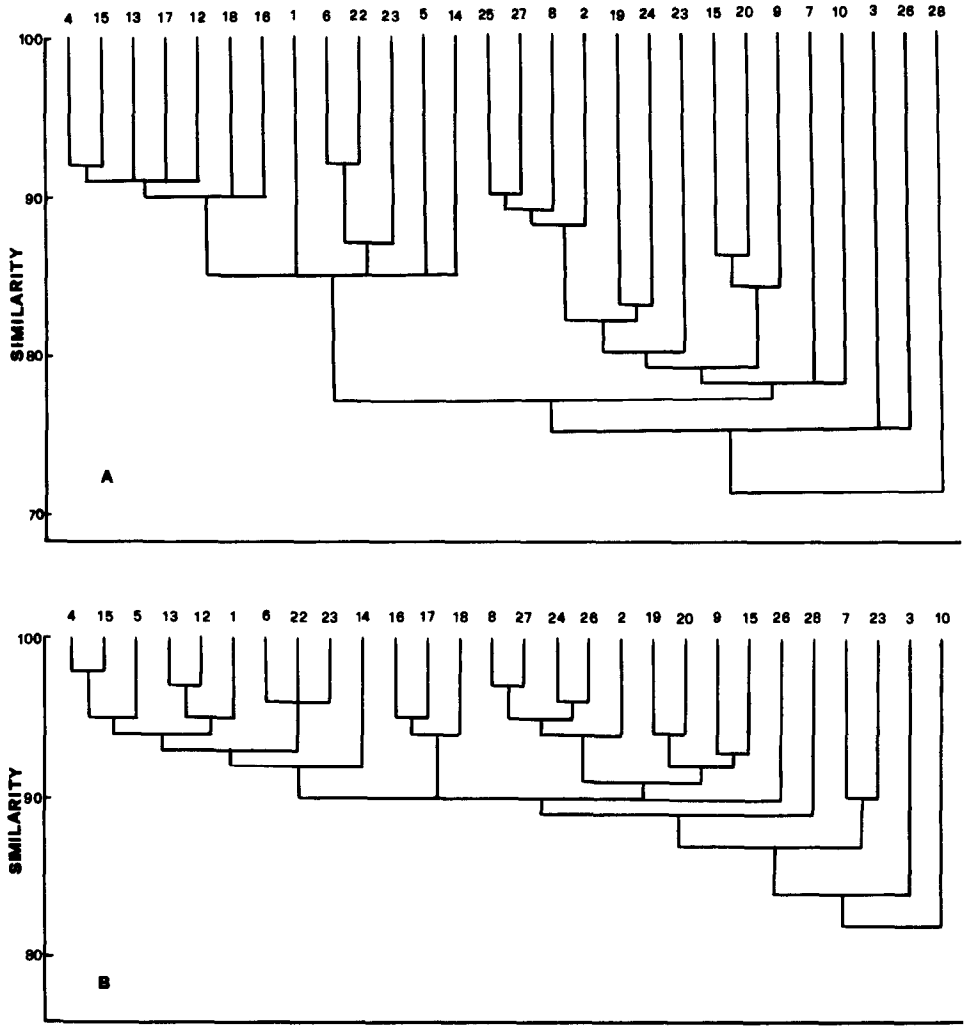

Fig. 1. Dendrograms obtained by the cluster analysis of: (A) S_{JS} coefficient and (B) S_w coefficient. The data refer to elm leaves. Numbers as in Table II.

compound but not on its absence. The similarity coefficients were calculated from all the pairs of sequences in Table II and were used for the cluster analysis. We used one of the simplest classification methods, that is, the single linkage cluster analysis[5]. The results of this kind of analysis are visualized by means of a dendrogram (see Fig. 1A). The data for 30 different two-dimensional chromatograms can be easily correlated.

A general consideration should, however, be made before discussing in detail the results in Fig. 1. The Jaccard–Sneath coefficient (S_{JS}) ascribes the same weight to each positive agreement and to each disagreement. It seemed interesting to ascribe a different weight to each spot depending on its frequency in the whole data matrix. In this way we consider the presence in one sequence of a compound which is present in a large number of sequences to be more important than the presence of one compound which rarely appears in the whole data matrix. For this reason we attributed a weight to each spot equal to the number of times that the spot appears in the whole sequences matrix. The resulting coefficient is

$$S_w = W_{11}/(W_{11} + W_{10} + W_{01})$$

where W_{11} is the sum of the weights of the spots present in both sequences and W_{01} and W_{10} are the sums of the weights of the spots present in one of the two sequences considered.

The value of a similarity coefficient S_w, such as that of Jaccard–Sneath, changes from 0 to 1. The S_w coefficient allows the introduction into each coefficient of information concerning the whole data matrix, in contrast to all the other similarity coefficients which consider only two sequences.

In order to test the validity of the S_w coefficient, Fig. 2 shows the correlation

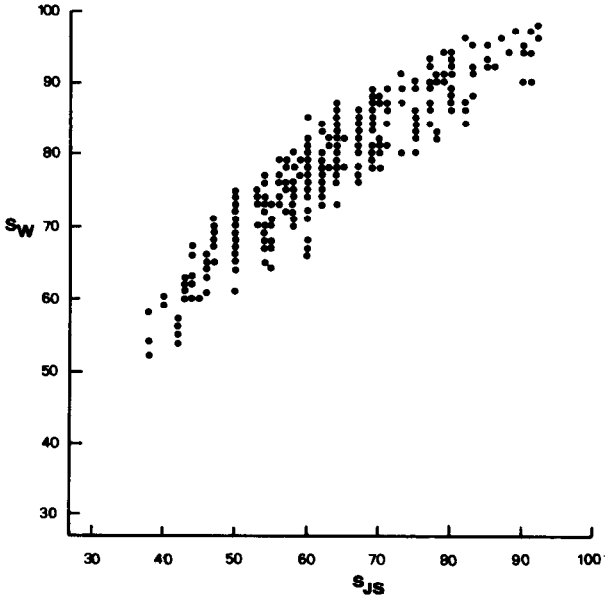

Fig. 2. Correlation between S_{JS} and S_w coefficients of the elm data set.

between the values of the S_{JS} and S_w coefficients for the same data matrix (Table II). It should be noted that the S_w coefficients generally have a higher value and exhibit a better differentiation. In fact, in many instances, one value of the S_{JS} coefficient corresponds to different values of the S_w coefficient; this occurrence could be of help in giving a better differentiation overall in those instances in which the chromatographic data are very similar. Fig. 1B shows the dendrogram obtained by the cluster analysis of the S_w coefficients. Comparison of the two dendrograms in Fig. 1A and B indicates that in Fig. 1B the similarity among the sequences due to the higher mean values of the S_w coefficients is increased with respect to Fig. 1A. However, from a general point of view, such an occurrence does not affect the dendrogram, as Fig. 1 must be considered as a whole.

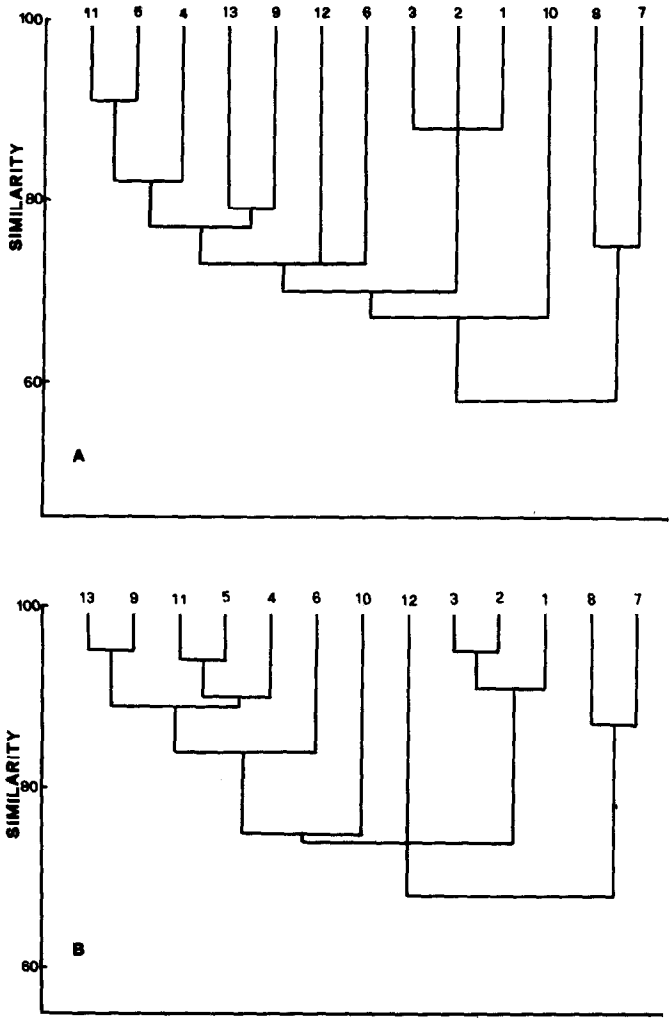

Fig 3. Dendrograms obtained by the cluster analysis of (A) S_{JS} coefficient and (B) S_w coefficient. The data refer to iris leaves. Numbers as in Table IV.

Let us now see what kind of information can be obtained from these dendrograms. *U. pumila* and *U. parvifolia* exhibit a high intra-specific variation owing to their provenance from a wide geographic area; in both instances, however, all *U. parvifolia* populations are in the cluster on the right. *U. japonica* and *U. carpinifolia* are very similar species from a botanical point of view[6]; in the dendrogram in Fig. 1A they appear in the same cluster. In Fig. 1B, however, the three populations of *U. carpinifolia* are gathered in one cluster, offering more detailed information in this instance where the differences between the samples are very small. The species *U. xhollandica* is very similar in all its populations (it should be noted that three of the four populations studied are identical) and in Fig. 1B (but not in Fig. 1A) the four populations are in one cluster.

As a further demonstration of the higher resolving power of the S_w coefficient for very similar sequences, it is interesting to consider the data for two populations of *U. japonica* (2P and 3P) which come from a restricted area of southern Japan; in Fig. 1B the two populations are linked in a more evident way than in Fig. 1A. From a botanical point of view other considerations could be made on the way in which the different species are linked, but this is beyond the aims of this paper.

As can be seen from the data in Table II, in all elm samples five spots were constantly found; in order to go deeper into the question of the interpretation of chromatographic data by means of cluster analysis, we considered the matrix in Table II without the five common columns. Apart from an expected translation towards lower values of similarity, there are no substantial differences with respect to Fig. 1A and B.

Fig. 3 shows the data relating to the iris leaves; Fig. 3A refers to the cluster analysis of the S_{JS} coefficients and Fig. 3B to that of the S_w coefficient. The only notable difference between the two dendrograms is found where the similarity between the sequences is higher, resulting in a better differentiation of clusters in Fig. 3B.

ACKNOWLEDGEMENT

This work was performed with financial support from the National Research Council (CNR).

REFERENCES

1 F. Geiss, *Fundamentals of Thin Layer Chromatography*, Hüthig, Heidelberg, 1987.
2 G. Guiochon, M. F. Gonnord, A. Siouffi and M. Zakaria, *J. Chromatogr.*, 250 (1982) 1–20.
3 S. D. Brown, T. Q. Arker, R. J. Larivee, S. L. Monfre and H. R. Wilk, *Anal. Chem.*, 60 (1988) 252R–273R.
4 D. Heimler and V. Vidrich, *Agrochimica*, in press.
5 P. H. A. Sneath and R. R. Sokal, *Numerical Taxonomy*, Freeman, San Francisco, CA, 1973.
6 G. Gambi, R. Gellinni and L. Brogi, *Inf. Fitopatol.*, 30 (1980) 27–47.