Journal of Chromatography, 173 (1979) 19–32

CHROM. 11,703

CLASSIFICATION OF FUNGI BY MEANS OF PYROLYSIS–GAS CHROMATOGRAPHY–PATTERN RECOGNITION

GÖRAN BLOMQUIST

National Board of Occupational Safety and Health, Department of Occupational Health, S-901 85 Umeå (Sweden)

ERIK JOHANSSON

Research Group of Chemometrics, Institute of Chemistry, Umeå University, S-901 87 Umeå (Sweden)

BENGT SÖDERSTRÖM

Department of Microbiological Ecology, University of Lund, S-223 62 Lund (Sweden)

and

SVANTE WOLD*

Research Group of Chemometrics, Institute of Chemistry, Umeå University, S-901 87 Umeå (Sweden)

(Received November 29th, 1978)

SUMMARY

Repetitive samples of three strains of the mould *Penicillium* were subjected to pyrolysis–gas chromatography (Py–GC). From the chromatograms, 26 peak heights were used in a subsequent SIMCA pattern recognition analysis. This data analysis gives a marked improvement in the classification of the samples (100% correct, 85% unique) in comparison with the traditional analysis based on the average chromatogram of each class (92% correct, 45% unique).

The data analytical method is described in detail using the Py–GC data as an illustration.

INTRODUCTION

Pyrolysis–gas chromatography (Py–GC) can be used to obtain a chemical "fingerprint" of complex samples in terms of chromatograms (see Fig. 1), *e.g.*, microorganisms[1,2], polymers[3] and oils[4]. Efforts to use these chromatograms for the classification of new samples have often been difficult, however, owing to the large apparent variability of repetitive chromatograms measured on the same type of samples[5].

In a previous paper[6], we showed that part of the variability between repetitive chromatograms measured on the same type of mould was systematic. This systematic variability could be modelled mathematically by means of a principal components (PC) model with two product terms (eqn. 1 below with $A = 2$). Thus, the "precision"

* To whom correspondence should be addressed.

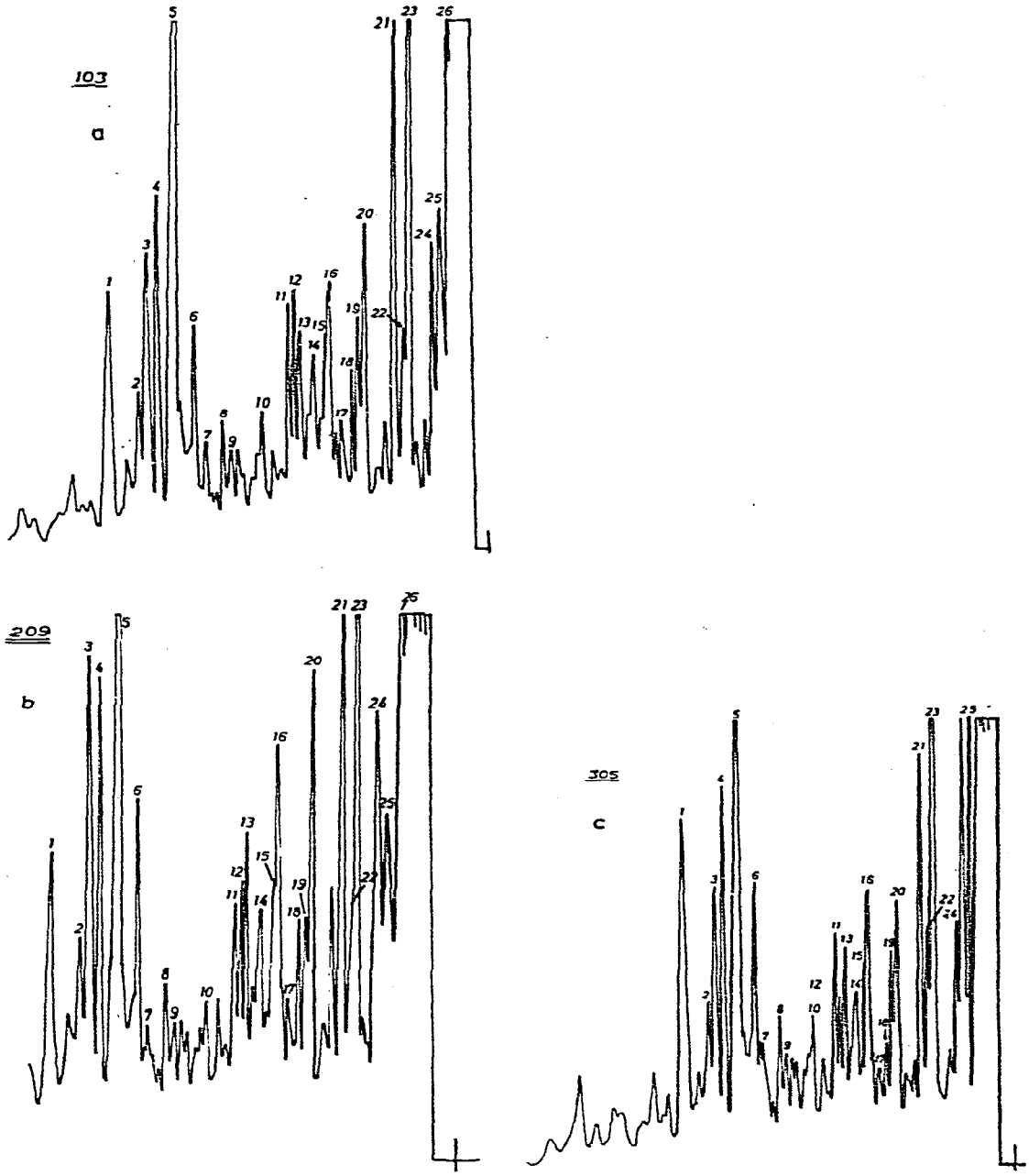


Fig. 1. Three Py–GC runs on samples of PB1, PF and PB2.

of the Py–GC runs improved by more than a factor of 2. In this paper we extend the investigation to three varieties of the mould *Penicillium*. We shall compare the classification of samples of this mould based on the traditional concept of reproducibility with the classification based on reproducibility of the second kind using PC models as introduced in the previous paper[6].

THEORY; CLASSIFICATION BASED ON MULTIVARIATE DATA

Class models

A gas chromatogram obtained from a pyrolysis of a mould (Fig. 1) can be digitized into a vector of numbers by describing each reoccurring peak by its height or integral. Thus, for run k the vector Y_k with the elements y_{ik} (peak i; $i = 1,2,\ldots,M$) is obtained. For a group or class of samples of the same type ($k = 1,2,\ldots,N$), it can be shown that the corresponding data matrix Y can be described by a principal components (PC) model[7], for class q:

$$y_{ik} = \bar{y}_i^{(q)} + \sum_{a=1}^{Aq} \beta_{ia}^{(q)} \theta_{ak}^{(q)} + \varepsilon_{ik}^{(q)} \qquad (1)$$

where the parameters $\bar{y}_i$ define the mean vector of the class and the parameters β and θ describe the correlation structure between the data in the class. The residuals ε_{ik} describe the "random" variation in the data. When the number of product terms A is significantly larger than zero, model (1) gives a better precision ("reproducibility") than the traditional model where the variability of the data is simply measured around the mean vector $\bar{y}_i^{(q)}$. The reproducibility of the second kind is measured by the residual standard deviation (S.D.):

$$s_0 = [\Sigma \varepsilon_{ik}^2/(M - A)\,(N - A - 1)]^{\frac{1}{2}} \qquad (2)$$

In the previous paper[6] we showed that indeed part of the variability among 10 Py–GC runs on *Penicillium brevi-compactum* was described by model (1) with two product terms ($A = 2$).

Further details on the estimation of the parameters $\bar{y}$, β and θ and the dimensionality, A, are given below and in refs. 7–10.

If now different classes of moulds are pyrolysed in replicate, we obtain a training set containing classes of pyrograms of "known type". One can then describe each class of mould replicates by means of a separate PC model (1). If the data contain information which differentiates between the different kinds of moulds, this results in different values of the parameters $\bar{y}^q$ and β^q for the different classes of replicates (class index q). These differences can then be used later to identify a new mould sample (a sample from the test set) on the basis of its pyrogram. Thus, the new pyrogram is digitized and normalized in the same way, giving data denoted by y_i^*. These data are fitted to each of the class PC models by means of linear regressions:

$$z_i^* = y_i^* - \bar{y}_i^q = \Sigma t_a^* \beta_{ia}^q + \varepsilon_i^{*q} \qquad (3)$$

The new mould is identified as belonging to the class the PC model of which shows the best fit, *i.e.*, the smallest residuals ε_i^{*q}. These residuals ε_i^{*q} should also be so small as to have an S.D. comparable to that estimated for the class by model (1), s_0^q (eqn.2).

Geometrically, this classification scheme has a straightforward interpretation. If each variable i is given an orthogonal coordinate axis, one obtains an M-dimensional space (M-space) where each pyrogram is represented as a point. Fig. 2 shows

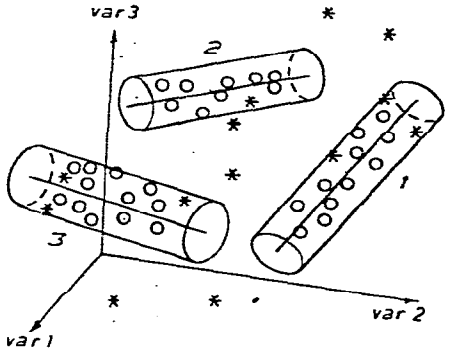


Fig. 2. A three-dimensional M-space with a training set containing three groups of mould sample points (rings), each group being described by a one-dimensional PC model, eqn. 1 with $A = 1$. The confidence cylinder around each PC model is constructed on the basis of the class residual S.D. eqn, 2. and the distribution of the parameters $\theta_k^{(q)}$ (see further refs. 7 and 8). New sample points from the test set (asterisks) are classified as belonging to a class if they fall inside the class "cylinder".

a simplified case with $M = 3$ and a training set consisting of three classes of moulds, each class represented by a one-dimensional PC-model, eqn. 1 (1) with $A = 1$.

The residual SD of each class, s_0^q (eqn. 2) is used to define a confidence interval as a cylinder around the PC model of the class. A new sample from the test set is identified as belonging to a class if it falls inside the corresponding confidence interval. Sample points falling outside all class cylinders are of a "new type"; they belong to hitherto unrepresented classes.

In practice, the number of variables, M, is usually larger than three and the dimensionality of the class PC models, A, is often larger than one. Most concepts involved can, however, be discussed with reference to Fig. 2, remembering that higher dimensional spaces have analogous properties to spaces with three dimensions.

This method of classification on the basis of disjoint PC models, called the SIMCA method (*s*oft *i*ndependent *m*odels describing *c*lass *a*nalogy), has been described in detail elsewhere[7–10] and applied to numerous classification problems of a chemical nature[11–19]. It suffices here to say the method has general applicability provided that the following assumptions are fulfilled: (1) the data should be of a continuous nature; (2) inside each group, the "objects" (in the present case the pyrograms) should be similar, *i.e.*, generated by a process undergoing small fluctuations. Both of these assumptions are well fulfilled in the present application and therefore the SIMCA classification method should work, provided that the data contain class-distinguishing information. Further details of the analysis will be discussed in connection with the actual analysis of the fungal data.

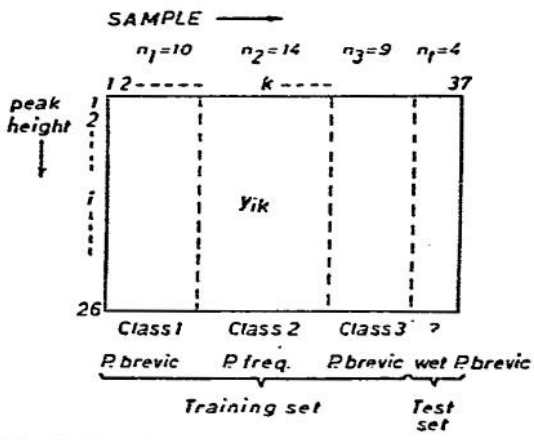
PYROLYSIS–GAS CHROMATOGRAPHY OF THREE VARIETIES OF *PENICILLIUM*

Three different fungal isolates were used to illustrate the methodology. *Penicillium brevi-compactum* Dierkx (CBS 210.28), *Penicillium frequentans* Westling (CBS 787.70) and a newly isolated *Penicillium brevi-compactum*. The identity of the last mould was confirmed by the Centraal Bureau voor Schimmelcultures (CBS), Baarn, The Netherlands. These three types will henceforth be referred to as PB1, PF and PB2.

They were separately grown in Oxoid malt extract broth for 5 days on a rotary shaker (100 rpm) at 22°. Very few conidia were formed during the incubation. The mycelium was harvested by filtration, freeze dried, ground in a mortar and stored in glass tubes in desiccators at room temperature.

Samples (approximately 0.5 mg) were pyrolysed (510°) in a Curie-point pyrolysis unit connected directly to the inlet of a gas chromatograph as described previously[6]. Ten samples of PB1 (class 1), 14 samples of PF (class 2) and 9 samples of PB2 (class 3) were taken as the training set. A small amount of water was added to each of four additional samples of PB1 and these samples were used as a test set.

The pyrolysis chromatogram obtained from each sample was digitized using the peak heights of 26 peaks occurring in all 37 chromatograms (see Fig. 1). Each sample data vector was then normalized to the sum 1000 over the 26 peaks. Thus, a 37 × 26 data matrix (Fig. 3) constitutes the empirical material for the data analysis described in the next section.



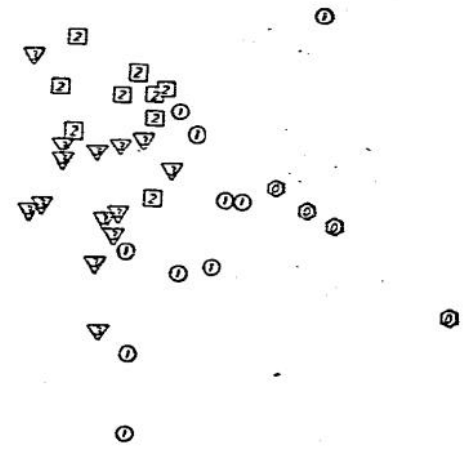


Fig. 3. The data matrix Y with elements y_{ik} denoting the observed height (normalized) of peak i of sample k.

Fig. 4. Eigenvector projection of the normalized data down on the plane corresponding to the third and fourth eigenvector of the whole date set. The four test samples (zeros) clearly are outside the three class domains.

To give graphical pictures of the data set, eigenvector projections of the data set[20,21] are useful. Fig. 4 shows one such projection from the 26-dimensional M-space down on a plane. We see a partial separation of the classes indicating a good separation in M-space.

DATA ANALYSIS AND RESULTS

Fitting PC models to each class

The data were first normalized by subtraction of the variable means and division by their standard deviation. This gave each variable a zero mean and unit variance over the data set and corresponds to giving them equal weight in the subsequent analysis.

Determination of the number of components, A, by cross-validation. When

describing the data matrix of a separate class, in the present instance a separate strain, by the PC model defined in eqn. 1, the first problem is to estimate the optimal number of product terms, A. This corresponds to an estimation of how much of the variation in the class data matrix is systematic, signal, and how much is "random" noise. In the present investigation, this estimation was made by means of cross-validation[10].

Briefly, elements are deleted from the class data matrix. PC parameters for values of $A = 0,1,2,\ldots$ are then estimated from the elements remaining in the matrix. These PC parameters are used to calculate "predicted" values for the deleted elements. Different predicted values are obtained for $A = 0,1,2\ldots$ and the predicted values are then compared with the actual values of the deleted elements. This deletion procedure is repeated until each element in the class data matrix has been deleted once and once only. The optimal value of A is that which gives the smallest prediction error on average. In the present instance the use of cross-validation shows that $A = 2$ for all three classes, *i.e.*, each class is best described by a two-dimensional hyper-plane in the 26-dimensional measurement space.

The parameters $\bar{y}$, β and θ. Once the dimensionality of each class model has been determined as discussed in the previous section, the estimation of the parameters in each class model, $\bar{y}$, β and θ is a routine numerical problem[21,22]. We use the NIPALS procedure of Wold[22] as described elsewhere[7]. The resulting parameters for this initial data analysis are not given as some variables were found to be irrelevant and deleted (see below). Hence, the values shown in Table I are those obtained from the second PC analysis with irrelevant variables deleted. The parameters θ shown in Table II describe the position of each sample in relation to its class model. In the present instance these values are of no particular interest as all samples in one class

TABLE I

RESULTING VALUES OF PARAMETERS $\bar{y}$, β_{1i} FOR EACH CLASS MODEL 1-3

$s_i^{(1)}$ are residual S.D.s (eqn. 4a) for each variable after the initial data analysis with all variables included, $s_i^{(2)}$ from the second data analysis with only relevant variables. $s_i^{(0)}$ denote the S.D.s for each variable around its class mean, *i.e.*, the variability related to the traditional reproducibility of the first kind. The discrimination power of each variable is given in the d_i row. The data have been scaled by subtracting the total mean (row 1) and dividing by the total S.D. (row 2).

Parameter	*Variable (i)*											
	1	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	*11*	*12*
Total mean	28	21	32	37	78	26	21	21	18	23	30	27
Total S.D.	8.2	4.7	6.3	5.9	17	5.6	6.5	4.9	3.3	6.1	5.2	2.4
$\bar{y}^{(1)}$	0.54	0.62	0.02	0.19	0.66	0.10	0.02	−0.18	0.22	0.19	0.57	0.94
$\beta_{1i}^{(1)}$	−0.11	—	0.01	−0.08	0.06	0.00	−0.27	−0.21	−0.37	−0.31	−0.32	−0.05
$\beta_{2i}^{(1)}$	0.47	—	0.28	0.33	0.16	0.16	0.19	0.20	0.09	0.09	0.08	−0.39
$\bar{y}^{(2)}$	−0.30	−0.31	0.26	0.02	−0.78	0.14	0.14	0.26	0.02	−0.11	−0.34	−0.75
$\beta_{1i}^{(2)}$	0.09	—	0.01	0.04	−0.01	−0.05	0.31	0.32	0.26	0.30	0.27	−0.14
$\beta_{2i}^{(2)}$	0.33	—	0.52	0.44	0.22	0.43	0.03	0.17	0.11	−0.04	−0.03	−0.03
$\bar{y}^{(3)}$	−0.14	−0.21	−0.43	−0.25	0.48	−0.33	−0.24	−0.22	−0.28	−0.05	−0.10	0.12
$\beta_{1i}^{(3)}$	−0.36	—	−0.28	−0.41	−0.37	−0.39	0.08	−0.04	−0.04	0.08	0.00	0.11
$\beta_{2i}^{(3)}$	−0.13	—	−0.14	−0.15	−0.05	−0.16	−0.30	−0.28	−0.26	−0.33	−0.27	0.16
$s_i^{(0)}$	0.98	0.93	0.94	1.0	0.81	0.99	0.99	0.96	0.98	1.0	0.90	0.71
$s_i^{(1)}$	0.44	0.65	0.35	0.28	0.42	0.52	0.33	0.25	0.32	0.29	0.27	0.60
$s_i^{(2)}$	0.37	—	0.29	0.28	0.45	0.48	0.32	0.26	0.31	0.27	0.26	0.53
d_i	1.4	1.4	2.1	1.4	3.0	1.3	1.6	2.7	1.3	1.2	2.2	2.6

originate from precisely the same fungus. In other instances when, for example, one class contains several strains, a plot of θ_{1k} against θ_{2k} for a given class gives information about the similarities between strains inside the class, clustering tendencies of samples and so forth.

Grouping of variables. Plots of the parameters β_1 against β_2 for each class (in the general case β_a against $\beta_{a'}$; $a \neq a'$) give information about the grouping of the variables. In the previous paper[6] we showed this plot for class 1 (PB1). The other two classes display similar plots, here shown for class 2 (PF) in Fig. 5. Six groups of variables are seen to cluster together in all classes, namely group 1, variables 1–6; group 2, variables 7–11 and 15; group 3, variables 12, 13, 22 and 24; group 4, variable 14; group 5, variables 16–19 and 25; and group 6, variables 20, 21, 23 and 26.

This grouping might provide valuable information about the origin of the pyrolysis fragments in the microorganism macromolecules, in particular if the chemical structures of the fragments are known. We have not investigated this aspect further, however, but show the plots to give a complete picture of the SIMCA methodology and its interpretative possibilities.

The residuals ε. The residuals ε_{ik} can be used to calculate standard deviations over each class (s_{iq}) and over the whole training set (s_i) for all variables i (n_q is the number of samples in class q and A_q the class model dimensionality):

$$s'_{iq} = \left[\sum_{k}^{n_q} \varepsilon_{ik}^2/(n_q - A_q)\right]^{1/2} \tag{4a}$$

$$s'_i = \left[\sum_{q=1}^{Q} s_{iq}^2/Q\right]^{1/2} \tag{4b}$$

13	14	15	16	17	18	19	20	21	22	23	24	25	26
28	24	29	22	21	26	30	37	64	31	193	49	51	133
2.7	3.1	4.8	4.5	4.4	4.9	4.8	5.0	15	12.5	32	8.5	17	26
−0.67	0.05	−0.19	−0.23	0.18	−0.12	0.46	−0.69	−0.22	−0.13	−0.66	0.16	0.22	−0.07
−0.03	–	−0.15	−0.31	−0.31	−0.26	−0.26	–	–	–	0.23	–	−0.15	0.32
−0.34	–	−0.01	−0.18	−0.19	−0.11	−0.17	–	–	–	−0.10	–	−0.22	0.07
−0.29	−0.21	0.43	0.22	−0.04	0.51	−0.63	0.47	0.35	−0.15	0.39	0.24	−0.60	0.19
0.02	–	0.31	0.29	0.22	0.26	0.24	–	–	–	−0.27	–	0.20	−0.26
−0.03	–	−0.06	−0.14	−0.13	−0.19	−0.10	–	–	–	−0.06	–	−0.10	−0.21
1.2	0.28	−0.45	−0.08	−0.14	−0.67	0.47	0.04	−0.30	0.38	0.12	−0.56	0.69	−0.28
0.10	–	−0.08	0.23	0.27	0.20	0.21	–	–	–	0.00	–	0.27	0.07
0.20	–	−0.20	−0.17	−0.30	−0.12	−0.04	–	–	–	0.31	–	−0.33	0.23
0.71	1.0	0.90	0.96	1.0	0.88	0.85	0.93	0.91	0.90	0.89	1.0	0.87	1.0
0.66	0.59	0.52	0.24	0.48	0.20	0.32	0.63	0.71	0.68	0.42	0.87	0.34	0.41
0.67	–	0.54	0.22	0.48	0.14	0.33	–	–	–	0.39	–	0.32	0.44
2.3	1.4	1.8	3.9	1.1	5.0	3.2	1.5	1.3	1.3	1.7	1.2	3.3	1.7

TABLE II

PARAMETERS θ_{1k} AND θ_{2k} FOR EACH SAMPLE TOGETHER WITH ITS RESIDUAL S.D., s_k, WITH RESPECT TO ITS "OWN" CLASS MODEL AND THE NEXT CLOSEST CLASS MODEL (NUMBER OF CLASS IN PARENTHESES)

Also shown are the corresponding values $s_k^{(0)}$ for the traditional model (eqn. 1 with $A = 0$).

Class	*Sample*	θ_{1k}	θ_{2k}	$s_k^{(own)}$	$s_k^{(next)}$	$s_k^{(0)}$	$s_k^{(0,\ next)}$
1	1	−4.7	1.8	0.35	0.92 (3)	1.0	1.4 (3)
	2	−2.0	−1.7	0.59	0.91 (3)	0.77	1.1 (3)
	3	3.6	0.4	0.45	0.66 (2)	0.78	0.83 (2)
	4	−2.6	−3.1	0.35	0.67 (3)	0.84	0.97 (3)
	5	0.8	1.2	0.38	0.62 (3)	0.41	0.73 (3)
	6	2.4	1.7	0.42	0.89 (2)	0.75	0.86 (2)
	7	−2.9	0.7	0.54	0.71 (3)	0.74	1.1 (3)
	8	4.2	−0.8	0.42	0.72 (3)	1.0	0.95 (3)
	9	3.2	−0.9	0.34	0.57 (3)	0.83	0.92 (3)
	10	−1.9	0.6	0.48	0.57 (3)	0.84	1.2 (2)
2	1	4.9	−1.2	0.22	0.90 (3)	1.0	1.2 (1)
	2	3.3	2.3	0.44	0.90 (1)	0.95	1.0 (1)
	2	8.1	0.1	0.36	1.2 (3)	1.7	1.7 (1)
	4	−2.0	−1.6	0.14	0.81 (3)	0.63	0.95 (3)
	5	5.2	−1.6	0.38	0.88 (3)	1.2	1.3 (1)
	6	−2.4	−3.2	0.24	0.99 (3)	0.92	1.2 (3)
	7	−2.3	−3.3	0.22	0.96 (3)	1.5	1.5 (3)
	8	−2.1	0.4	0.17	0.69 (1)	0.47	0.86 (3)
	9	−2.9	2.1	0.32	0.83 (1)	0.75	0.99 (3)
	10	−2.8	−0.9	0.17	0.74 (1)	0.62	0.94 (3)
	11	−2.2	2.8	0.45	0.79 (1)	0.81	1.1 (1)
	12	−0.9	4.2	0.49	0.99 (3)	1.1	1.1 (1)
	13	−2.2	−2.8	0.31	0.78 (3)	0.80	0.98 (3)
	14	−1.7	2.2	0.28	0.80 (1)	0.61	0.90 (1)
3	1	3.5	2.0	0.25	0.80 (1)	0.81	1.1 (2)
	2	1.7	−4.8	0.40	0.81 (1)	1.0	1.1 (1)
	3	1.7	1.1	0.46	0.83 (1)	0.48	1.0 (2)
	4	−4.8	−1.3	0.27	0.85 (1)	0.92	0.99 (1)
	5	2.4	−2.2	0.68	0.82 (1)	0.72	0.87 (1)
	6	0.6	2.0	0.38	0.54 (1)	0.49	0.82 (2)
	7	−3.6	0.5	0.30	0.64 (1)	0.75	0.80 (1)
	8	−1.4	0.90	0.27	0.59 (1)	0.37	0.74 (1)
	9	−0.2	1.8	0.28	0.66 (1)	0.39	0.91 (2)
Test	1	—	—	—	1.3 (1)	—	1.2 (1)
	2	—	—	—	1.5 (1)	—	1.4 (1)
	3	—	—	—	1.9 (1)	—	1.8 (1)
	4	—	—	—	1.8 (1)	—	1.9 (1)

The former give information about the "modelling power" of each variable in each class and are given in Table I for the three classes.

The residual S.D. for each class, s_0^q (eqn. 2), gives a measure of the "typical" distance between a mould of class Q and its class PC model. The residual S.D.s for each sample, s_k, with respect to a class model q describes the distance between the sample vector point (M variables) and the class model q:

$$s_k^q = \left[\varphi \sum_{i=1}^{M} (\varepsilon_{ik}^q)^2/(M - A_q)\right]^{1/2} \tag{5}$$

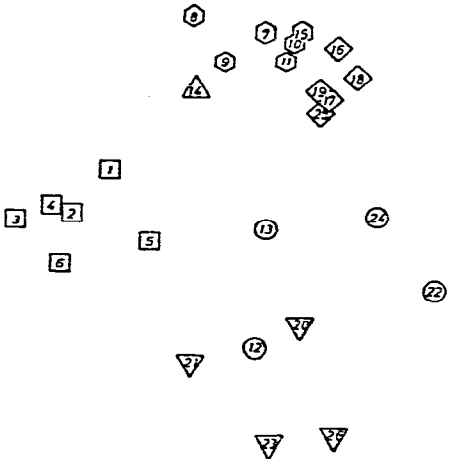


Fig. 5. Values of β_{2i} plotted against values of β_{1i} for class 2 (PF). The parameters β are not taken from Table I but correspond to the initial analysis with all variables included.

The correction factor φ equals 1.0 if sample k is not in the training set of class q. (Note that in this instance ε_{ik} is calculated using eqn. 3). When sample k is a member of the class q training set (with n_q samples), the correction factor φ becomes

$$\varphi = n_q/(n_q - A_q - 1) \tag{6}$$

The values of s_k for each mould sample are given in Table III. We can see that indeed the moulds fit their "own" class models much better than they fit the other models. These results are discussed further in the classification section.

Relevance of variables

After fitting separate PC models to each group of sample data vectors (the training set), one can calculate various measures of relevance for the variables i. The modelling power measures how much a variable i participates in the modelling of the groups. It is calculated from the S.D. of the residuals of the variable i (eqn. 4a and 4b) in relation to the S.D. of the corresponding data (see Table I).

The discrimination power of a variable measures its degree of class separation ability. This is calculated from the residuals of variable i obtained when all sample vectors are fitted to class models other than their "own" in relation to the corresponding residuals when the same vectors are fitted to their "own" class models (see Table I). Values close to 1 correspond to "bad" and values larger than 3 correspond to "good" discriminatory power. The mathematical details are given in refs. 7 and 8.

Selection of variables

In a given classification problem, it is often found that several variables are "irrelevant" to the problem. In other instances one wishes to reduce the number of variables to a more manageable set. One then often, without much thought, selects those variables which show the largest differences between the classes. If the number of samples in each class is very large compared with the number of variables (M), this procedure is not unsound. However, in the common case when the number of

variables approaches or even exceeds the number of cases in the training set (in the present instance the former is 26 and the latter is 33), this is a dangerous and often misleading procedure. The reason is that there is always a certain chance that a variable, even if totally irrelevant to the problem, will show a substantial difference between the classes. If the number of variables is large, the total chance is fairly large that a few of the variables by accident will display substantial differences between the classes.

Empirically it has been found that the selection of variables must not be based on differences between classes if the number of variables exceeds a third of the total number of samples in the training set[23,24]. In such cases, which include the present case, other selection criteria must be used, which are not based on the separation of classes as such.

The SIMCA methodology provides two measures of relevance for the variables. One, the discrimination power (see above), is based on the class separation and cannot, therefore, be used here. The second, the modelling power, is based on how much each variable participates in the modelling of the classes. This measure does not utilize the class separation and is therefore useful in the present instance.

Table I gives the modelling power for each variable in each class in terms of its residual S.D., $s_i^{(q)}$. The variables 2, 14, 20–22 and 24 are seen to have low relevance in some class and were therefore deleted and the PC analysis repeated with the reduced data matrix with 20 variables.

Class distances

A measure of the distance between two classes r and q is calculated from (a) the residuals obtained when all objects in class r are fitted to class model q and *vice versa* in comparison with (b) the residuals when all objects in classes q and r are fitted to their "own" class models[7,8].

Table III gives the class distances for (i) the PC analysis with all 26 variables, (ii) the PC analyses with only the 20 relevant variables included and (iii) for the traditional reproducibility, eqn. 1 with $A = 0$. It can be seen that the classes are fairly well separated and that the separation increases when irrelevant variables are deleted. The class separation based on the traditional model of reproducibility is seen to be substantially smaller. We also see that the two *P. brevi-compactum* strains (classes 1 and 3) are closer to each other than the *P. frequentans* strain (class 2).

Validation

A very important step in a classification data analysis is to validate the results. For many data analytical methods, the classification of the training set gives a "success rate" which is highly over-optimistic[19,23,24]. Therefore, it is necessary to make a check of the classification rate on the basis of a test set which has not been involved in the training phase of the data analysis but which still has a "known" classification.

The SIMCA method gives classification results for the training set which are little biased towards optimism. The reason is that the SIMCA method in its training phase calculates the separate class models independently, not directly using the information of class assignment of the objects in the training set to maximize the class separation. Even so, it is advisable always to perform a validation to confirm the lack of bias. This was done in the present study by using a repeated partial validation,

TABLE III

CLASS DISTANCES

The residual S.D. when samples in class S are fitted to class model r. (i) 26 variables and (ii) 20 relevant variables. Case (iii) shows the class distances obtained using the traditional model, eqn. 1 with $A = 0$, *i.e.*, describing each class by its variable averages.

Case	r	S		
(i)	1	*1*	*2*	*3 (own)*
(i)	1	0.57	0.97	0.72
	2	0.94	0.49	0.87
	(fitted) 3	0.80	0.95	0.38
(ii)	1	0.44	0.93	0.74
	2	0.91	0.32	0.90
	3	0.75	0.91	0.39
(iii)	1	0.85	1.2	0.97
	2	1.1	0.97	1.0
	3	1.0	1.2	0.76

cross-validation. Then the training set is divided into four sub-sets. The first contains sample 1, 5, 9, ..., etc., of each of the three classes. The second sub-set contains samples 2, 6, 10, ..., etc., the third sub-set samples 3, 7, 11, ..., and the fourth sub-set samples 4, 8, 12, ..., of each class.

Then, four separate data analyses are made. In the first analysis sub-set one is made into a test set, resulting in a reduced training set consisting of sub-sets 2, 3 and 4. The data analysis is carried out as usual, developing separate PC models for the three classes. The test set with sub-set 1 is then classified by means of these models. Secondly, sub-set 2 is made into a test set and the reduced training set is made to consist of sub-sets 1, 3 and 4. Now PC models are calculated on the basis of this training set and the "test set" (sub-set 2) is classified using these PC models. The process goes on until each sub-set has been used as a test set once and, in this way, each sample in the training set has been in an "artificial" test set once and once only.

The validated success rate is then calculated from the rate of classification of the samples when they constituted parts of the test sets.

This cross-validation provides an unbiased classification rate provided that the number of really independent samples in each class is larger than the number of sub-sets created. This latter assumption is best checked by looking on tendencies of clustering with each class on θ_1–θ_2 plots. These plots look non-clustered in the present instance. The result of the validation shows that 31 of the 33 samples are correctly classified, 29 of these uniquely. One sample is uniquely misclassified. See the next section for a definition of classification uniquess.

CLASSIFICATION

When a new mould sample is to be classified on the basis of its data vector y^*, this data vector is fitted to each of the class models (now with fixed $\bar{y}$ and β values) using multiple regression, eqn. 3. The resulting residuals e^{*q} have the S.D. s_q^*, defined in eqn. 5 (denoted there by s_k^q).

In order to be classified as belonging to a class q, a sample should have a residual S.D. that does not significantly exceed the "typical" S.D. of the class s_0^q (eqn. 2). If this significance is tested by means of an approximate F-test on a desired level of significance with $(M-A_q)$ and $(M-A_q)(n_q-A_q-1)$ degrees of freedom, we have the following condition for the sample to be classified as belonging to class q:

$$s_0^q \sqrt{F_{crit}} \geqslant s_q^* \tag{7}$$

However, in order to be uniquely classified as class q, the fit of the sample data vector to the other class models must also be significantly worse, *i.e.*, the ratio R must exceed F_{crit} in eqn. 8 with $(M-A_q)$ and $(M-A_q)$ degrees of freedom. Here s_r^* denotes the residual S.D. corresponding to the next best fitting class model:

$$R = (s_r^*/s_q^*)^2 \geqslant F_{crit} \tag{8}$$

From Table II we can see that all samples are closest to their own class. Of the ten samples in class 1 seven are uniquely classified as class 1 on the 95% level, while three (nos. 3, 7 and 10) are closer to class 1 but also rather close to class 3 and therefore not uniquely classified. All fourteen samples in class 2 are uniquely classified (95%) and also seven of the nine samples in class 3 (not nos. 6 and 7). The four test samples are all classified as belonging to none of the classes, *i.e.*, they are outside the confidence intervals for all classes. Evidently the additional humidity in these samples makes them show different pyrolysis behaviour.

It is interesting to compare these results with those obtained when comparing each sample with the average chromatogram of the classes. This corresponds to the use of the $s_k^{(0)}$ values in Table II for the classification. Using these, one sample is closer to a "wrong" class (no. 8 in class 1) and three samples in class 2 (nos. 3, 7 and 12) as close to another class as their "own". Only 15 of the 33 samples are uniquely classified (compared with 28 out of 33 above), showing the substantial loss of information in this traditional analysis.

CONCLUSIONS AND DISCUSSION

The combination of Py–GC and SIMCA pattern recognition (Py–GC–Pr) gives a good classification in the present example of three fungi chosen to illustrate the methodology.

We wish to emphasize that in order to obtain a working method for a specific micro-organism classification, such factors for variability as change of GC column, variation of cultivating medium and drying technique and, probably most important, strain of microorganism, must be incorporated into the training set of each class. This is presently under investigation in our laboratory for common micro fungi. Hence this paper must be seen as an illustration of the possibilities of the methodology, not as a final method paper describing a working classification of *Penicillium* species.

The main result in this paper is, in our view, the much improved separation between the classes when going from the traditional model of reproducibility to the model based on principal components analysis. Although the separation of the classes

is still not 100%, we foresee a further improvement with the use of more GC peaks; one can easily extract 40 reoccurring peaks from the chromatograms we presently obtain on standard packed columns. This is significantly more than the 26 used in the present illustration.

Finally, we wish to comment on our choice of pattern recognition method. The SIMCA method has the advantage of giving direct measures of relevance of the variables. This allows the deletion of "noise" peaks from the data analysis which, in our experience, often significantly improves the classification.

Another important SIMCA feature is that each sample is classified not only according to the closest class but also on the basis that it should be sufficiently close to the class to be a typical class member. This allows the detection of "outliers" both among samples in the test set and in the training set, outliers which might be mutants of an otherwise rare micro-organism not included in the training set. In practice, methods which cannot detect such outliers are, in our view, of little value.

Thirdly, the SIMCA analysis gives a model of each class which often gives interesting insight into more fundamental questions such as the similarities between variables or samples within a class. When quantitative properties of samples are known, say, for example, their sensitivity to heat or disinfectants or their reproduction rate, one can also seek relationships between the position of a sample in a class and the value of its quantitative property. This type of "level 3" pattern recognition[9] has recently been applied in other areas of chemical classification[17,18].

In conclusion, we feel that the Py–GC–Pr combination at least partly removes the problems of apparent lack of reproducibility hitherto complicating the use of Py–GC in the routine classification of complex chemical and biological samples.

ACKNOWLEDGEMENTS

This project was supported by grants from the Swedish Work Environment Fund and the Swedish Natural Science Research Council.

REFERENCES

1 E. Reiner, *Nature (London)*, 206 (1965) 1272.
2 D. T. Burns, R. J. Stretton and S. D. A. K. Jayatilake, *J. Chromatogr.*, 116 (1976) 107.
3 H. Eustache, N. Robin, I. C. Daniel and M. Carrega, *Eur. Polym. J.*, 14 (1978) 239.
4 C. S. Glam, T. E. Goodwin, P. Y. Glam, K. F. Rion and S. G. Smith, *Anal. Chem.*, 49 (1977) 1540.
5 H. L. C. Meuzelaar, P. G. Kistemaker, W. Eshuis and H. W. B. Engel, in H. H. Johnston and S. W. B. Newsom (Editors), *Rapid Methods and Automation in Microbiology*, Learned Information (Europe) Ltd., Oxford, 1976, pp. 225–230.
6 G. Blomquist, E. Johansson, B. Söderström and S. Wold, *J. Chromatogr.*, 173 (1979) 7.
7 S. Wold, *Pattern Recognition*, 8 (1976) 127.
8 S. Wold and M. Sjöström, in B. R. Kowalski (Editor), *Chemometrics: Theory and Application*, ACS Symposium Series 52, American Chemical Society, Washington, D.C., 1977, p. 243.
9 C. Albano, W. Dunn, U. Edlund, E. Johansson, B. Nordén, M. Sjöström and S. Wold, *Anal. Chim. Acta*, 103 (1978) 429.
10 S. Wold, *Technometrics*, 20 (1978) 397.
11 D. L. Duewer, B. R. Kowalski and T. Schatzki, *Anal. Chem.*, 47 (1975) 1973.
12 M. Sjöström and U. Edlund, *J. Magn. Reson.*, 25 (1977) 285.
13 S. Wold, *Proc. First Int. Symp. on Data Analysis and Informatics*, Vol. 2. IRIA, Rocquencourt, France, 1977, p. 683.

14 U. Edlund and Å. Norström, *Org. Magn. Reson.*, 9 (1977) 196.
15 O. Strouf and S. Wold, *Acta Chem. Scand.*, A31 (1977) 391.
16 U. Ulfvarsson and S. Wold, *Scand. J. Work, Environ. Health*, 3 (1977) 183.
17 W. Dunn, S. Wold and Y. Martin, *J. Med. Chem.*, 21 (1978) 922.
18 W. Dunn and S. Wold, *J. Med. Chem.*, 21 (1978) 1001.
19 M. Sjöström and B. R. Kowalski, *Anal. Chim. Acta*, (1979) in press.
20 B. R. Kowalski and C. F. Bender, *J. Amer. Chem. Soc.*, 95 (1973) 696.
21 R. Gnanadesikan, *Methods for Statistical Data Analysis of Multivariate Observations*, Wiley, New York, 1977.
22 H. Wold, in F. N. David (Editor), *Festschrift for J. Neyman*, Wiley, New York, 1966, p. 411.
23 N. A. B. Gray, *Anal. Chem.*, 48 (1976) 2265.
24 D. L. Massart, A. Dijkstra and L. Kaufman, *Evaluation and Optimization of Laboratory Methods and Analytical Procedures*, Elsevier, Amsterdam, 1978.