ELSEVIER

Chemometrics and Intelligent Laboratory Systems 23 (1994) 213–229

Chemometrics and intelligent laboratory systems

Correspondence, principal coordinate, and redundancy analysis used on mixed chemotaxonomical qualitative and quantitative data

Jens C. Frisvad

Department of Biotechnology, Building 221, The Technical University of Denmark, DK-2800 Lyngby, Denmark

(Received 15 July 1993; accepted 24 December 1993)

Abstract

A mixed type data matrix consisting of 11 quantitative carbohydrate variables and 23 binary secondary metabolites data measured in 5–8 isolates of 7 species of *Penicillium* was analyzed using different multivariate statistical methods. This kind of data matrix is common in numerical taxonomy and has formerly been analyzed by consensus methods based on the separate analysis of the quantitative and qualitative data matrix, by using Gower's general similarity coefficient for mixed data or by location models. For the initial data treatment the χ^2, Bray–Curtis and Canberra distance coefficients were useful for cluster analysis and minimum spanning trees (MSTs) combined with principal coordinate analysis (PCO). The multivariate ordination methods hitherto recommended for chemotaxonomic data, principal component analysis (PCA) and its constrained ordination equivalent partial least squares (PLS) analysis (using dummy variables for each species) gave seven quite diffuse clusters with some overlap in two-dimensional ordination plots, while correspondence analysis (CA) gave seven very clear clusters. The results indicate that qualitative data strongly dominate quantitative data and that these qualitative data are best represented in plots by correspondence analysis. However, in physiological studies the quantitative data may be considered the most important, PCA and CA are preferred for the analysis of mixed data. Dummy constrained PLS may be used to select quantitative variables that are species specific rather than related to climatic conditions. In classification studies at the species level it is recommended to use correspondence analysis on mixed chemotaxonomical data. In the latter studies variables based on differentiation, such as the biosynthetic families of secondary metabolites used here, give clear species separations, and can be used for further cladistic analyses.

1. Introduction

In chemotaxonomy data matrices are either polyphenic [1] (from one chemical procedure, i.e., peak areas in chromatography) or a series of quantitative measurements in different scales (using different chemical methods). For many years cluster analysis was the method of choice for evaluating chemotaxonomical data [2], but ordination methods such as principal component analysis have been used extensively for evaluating quantitative data, such as profiles of free fatty acids [1,3]. In principal component analysis (PCA) and cluster analysis the taxa are not known a priori, but in constrained ordination methods such as redundancy analysis, which is equivalent to

SSDI 0169-7439(94)00003-2

Table 1
Penicillium isolates examined

Penicillium echinulatum
1 IBT 3236
2 IBT 3238
3 IBT 3239 = IMI 285508 = FRR 2927 = ATCC 58618
4 IBT 4305 = NRRL 1151 (ex type)
5 IBT 3233 = FRR 1621 (received as *P. crustosum*)
6 IBT 3234 = FRR 1963 (received as *P. verrucosum*)
Penicillium crustosum
7 IBT 3425 = IMI 285510
8 Leistner Sp 458 (received as *P. verrucosum* var. *cyclopium*)
9 LC 75.3045 (received as *P. verrucosum*)
10 IBT 12586 = NRRL 1983
11 NRRL 1983 (few conidia)
12 IBT 12585
Penicillium verrucosum
13 IBT Bpa
14 IMI 293193 = ATCC 58630
15 IBT 5079 = NRRL 3712 (received as *P. viridicatum*)
16 IBT 5010 = IMI 285522 = FRR 2940
17 IBT 6734 = IMI 285523 = FRR 2941
Penicillium aethiopicum
18 CSIR 1375 (received as *P. viridicatum* [a])
19 CCM F-398 (received as *P. crustosum*)
20 IBT ETH 3-8
21 IBT 4701
22 IBT 3353 = Leistner Sp 1448 (received as *P. verrucosum* var. *corymbiferum*)
23 IBT 6569 = CBS 484.84 = IMI 285524 = FRR 2942
Penicillium griseofulvum
24 IMI 285525 = FRR 2943
25 IBT 6868
26 IBT 6867
27 IBT 10559 (few conidia)
28 IBT 6910
29 IBT 10559
30 IBT 6868
31 NRRL A-26933 (received as *P. granulatum*)
32 NRRL 2159A
Penicillium chrysogenum
33 IMI 285517 = FRR 2936 = ATCC 58611
34 IBT JAPA6
35 IBT 6055
36 IBT PGASS2
37 IBT 5233 = NRRL 807
Penicillium dipodomyis
38 IBT 5476 = NRRL A-26825
39 IBT 5315 = NRRL A-26868
40 NRRL A-26881
41 NRRL A-26652
42 IBT 5475 = NRRL A-26737
43 IBT 5318 = NRRL A-26931
44 IBT 5324 = NRRL A-26656

[a] Later reidentified as *P. expansum*.

partial least squares (PLS) analysis, each taxon is known and, in its most simple form, can be given a value of 1 in the **Y** matrix with as many columns as there are taxa if an operational taxonomic unit (OTU) belongs to that taxon and 0 if it is any other taxon (dummy constrained ordination). PLS [3–7] has been recommended by some authors for analysis of two independent sets of chemical data, but has rarely been used for classification or discrimination purposes in chemotaxonomy [3].

Another kind of data matrix can occur in chemotaxonomy: the matrix may be of a mixed nature (e.g., chromatographic data and categorical data based on differentiation). Such data have often been treated by general distance coefficients followed by clustering [8] or by the location model, i.e., based on a homogeneous Gaussian distribution and discriminant analysis [9]. Krzanowski [9] has written an excellent review on the strict statistical treatment of mixed qualitative and quantitative data. In the ecological and chemometrical literature, however, distribution and data structure are often less well defined and more soft modelling methods may be appropriate. In psychometrical and ecological research mixed data are frequent and many authors have recommended methods based on correspondence analysis [10–18]. Correspondence analysis and canonical correspondence analysis, however, [15] have rarely been used in numerical taxonomy and chemotaxonomy and even more seldomly on matrices with mixed quantitative and qualitative data. Ter Braak and Prentice [14] and Ter Braak [15,16] have shown that the computer-intensive method of Gaussian ordination can be approximated by correspondence analysis. Thus the location model, building on a Gaussian distribution, may also be approximated by correspondence analysis when analyzing mixed quantitative and qualitative data.

In a former study a large number of binary secondary metabolite data were analyzed by several statistical techniques. Fuzzy clustering, PCA/principal coordinate analysis (PCO) and PLS gave less sharp species differences than clustering using the Yule coefficient and correspondence analysis (CA) even though all methods were valuable for analyzing the data [19]. However, in chemotaxonomy both quantitative data (free fatty acids, carbohydrates, amino acids, etc.) and qualitative data (secondary metabolites, morphological structures) are often available for a number of strains. The purpose of this study was to find out whether quantitative data (from primary metabolism) or qualitative data (from differentiation) appeared to give most information on species identity and discrimination, especially when they are treated simultaneously. The data treated are based on chemical analyses of several isolates in each of seven species of *Penicillium* that have often been misclassified to see if quantitative and qualitative chemical data would reveal the same seven species.

2. Experimental

A number of *Penicillium* isolates from different sources and geographic regions were selected in each of seven species (Table 1). Several isolates were received under other names, as identified by *Penicillium* experts [20]. Conidium-suspensions were made from Czapek yeast autolysate agar (CYA) [20,21]. All isolates were grown on CYA, malt extract agar (MEA) [21] and yeast extract sucrose (YES) agar [19,21] at 25°C in the dark for 14 or 16 days. For quantitative analysis of carbohydrates and polyols the isolates were grown on YES agar (Difco yeast extract and agar) for 16 days and extracted with 80% methanol for 5 min using a Waring blender. The extract was filtered, freeze dried, dissolved in 1 ml of water and rinsed by small-scale ion exchange column chromatography using the method of Bjerg et al. [22]. The neutral water fraction containing the carbohydrates was free of amino acids and organic acids and analyzed for the quantitative amount of carbohydrates using gas chromatography after silylation. Standards of glycerol, erythritol, arabinitol, inositol, mannitol, sorbitol, fructose, glucose, mannose, sucrose, lactose, and trehalose were analyzed to be able to identify the peaks. After silylation, fructose appears as two peaks and glucose appears as three peaks, so these peak areas were summed. In the case where no peaks could be detected for a certain carbohydrate the detection limit, 1000

area units, was used. The peak areas were not normalized but the logarithmic (log10) values were used for statistical treatment. The secondary metabolite data were used without any transformations.

The contents of each of three plates of CYA, MEA and YES were extracted with chloroform/methanol (2:1) and thereafter with ethylacetate with 1% formic acid and the organic phases were combined after filtering through a phase separation filter. After evaporation of the organic solvents, the residue was taken up in 3 ml of methanol and defatted with petroleum ether [23]. 10 μl of this extract was analyzed by high performance liquid chromatography (HPLC) using diode array detection (DAD) [23,24]. The metabolites were identified by their retention indices as compared to standards and their UV spectrum as measured by DAD. The metabolites identified or characterized were classified into

Table 2
Variables used to characterize seven *Penicillium* species

Variable No.	Variable
Quantitative variables	
1	Unknown a: eluting before glycerol
2	Glycerol
3	Erythritol
4	Arabinitol
5	Unknown b
6	Unknown c
7	Fructose
8	Glucose
9	Mannitol
10	Sucrose
11	Trehalose
Binary variables	
12	Viridicatin family (cyclopeptin, dehydrocyclopeptin, cyclopenin, cyclopenol, viridicatol, viridicatin)
13	Palitantin family (palitantin, occasionally frequentin)
14	Territrems (A, B, C, etc.)
15	Penechins (A, B, C, D, etc.)
16	Terrestric acids (terrestric acid, viridicatic acid)
17	Penitrems (Penitrem A–F)
18	Roquefortines (roquefortine C, D and precursors)
19	Metabolite family 'Q'
20	Anacines (anacine, anacine B, etc.)
21	Verrucolones (verrucolone, verrucolone B, C, etc.)
22	Ochratoxins (A, B, C and precursors)
23	Metabolite family 'I'
24	Griseofulvins (griseofulvin, dechlorogriseofulvin, norlichexanthone, dehydrogriseofulvin, etc.)
25	Tryptoquivalins and tryptoquivalons
26	Viridicatumtoxins
27	Metabolite family 'X'
28	Patulins (patulin, isopatulin, ascladiol, 6-methylsalicylic acid, isoepoxydon, etc.)
29	Cyclopiazonic acids (Cyclopiazonic acid, cyclopiazonic acid imine, bissecodehydrocyclopiazonic acid)
30	Chrysogenines (chrysogenine, 2-pyrovoylaminobenzamide, 2-acetyl-4(3H)-quinazolone)
31	Penicillins (penicillin F, G, etc.)
32	Metabolite family 'Ø'
33	Metabolite family 'D1'
34	Metabolite family 'D2'

chromophore families [24] and finally into biosynthetic families where this was possible using standards [23,25].

Based on the results, three types of data matrices were analyzed: The carbohydrate data matrix (44 objects, Table 1 and 11 variables, Table 2), the combined data matrix (44 objects and 11 quantitative + 23 binary variables, Table 2) and a data matrix consisting of 44 objects and 11 quantitative + 7 binary variables in the **X** matrix. In the latter case these binary variables were representing the biosynthetic family that was unique for a given species, i.e., territrems for *P. echinulatum*, penitrems for *P. crustosum*, verrucolone for *P. verrucosum*, viridicatumtoxins for *P. aethiopicum*, cyclopiazonic acids for *P. griseofulvum*, chrysogenins for *P. chrysogenum* and met D1's for *P. dipodomyis*. For constrained ordination methods a dummy binary matrix was used as the **Y** matrix (i.e. seven binary variables representing species, one if an isolate belongs to that species, zero otherwise).

The data were analyzed using the programs NT-SYS version 1.80 [26] (Exeter Biological Software, Setauket, USA), SYNTAX version 5 [27] (Exeter Biological Software, Setauket, NY, USA), SIRIUS version 2.3 (Pattern Recognition Systems, Ulset, Norway) [28], SIMCA (Umetri, Umeå, Sweden) [29] and CANOCO version 3.12 (Microcomputer Power, Ithaca, NY, USA) [30]. Eigenvalues were compared to the 'broken stick model' [26,31], which gives approximate eigenvalues for data matrices with 'random' data.

3. Results and discussion

3.1. Analysis of the quantitative data

Unweighted pair-group method using arithmetic averages (UPGMA) cluster analysis (using all the similarity, distance and correlation coefficients for interval data in the NT-SYS package) on the quantitative carbohydrate data only showed a partial separation of the 44 fungal isolates (dendrograms not shown). Different ordination methods (principal component analysis (PCA), principal coordinate analysis (PCO)) using the available similarity coefficients in NT-SYS, and correspondence analysis showed that the different species could not be separated using the carbohydrate data. However, an interesting par-

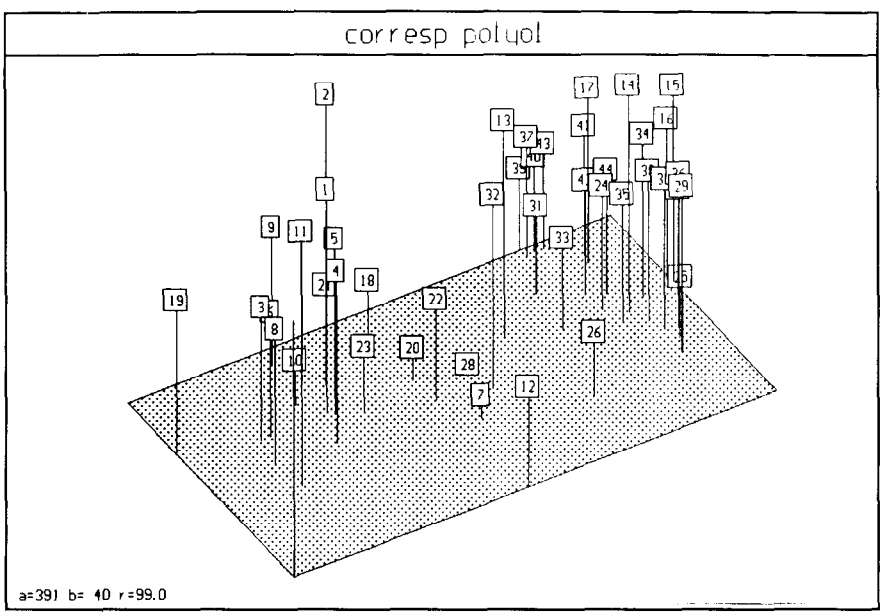

Fig. 1. Correspondence analysis of the carbohydrate data. A partial separation into two physiologically distinct groups is seen.

tial separation into species growing mostly on lipid- and/or protein-rich substrates such as meat or cheese (*P. echinulatum*, *P. crustosum* and *P. aethiopicum* [31], the creatine positive species [32]) and species growing often on carbohydrate rich or low water activity substrates (*P. griseofulvum*,

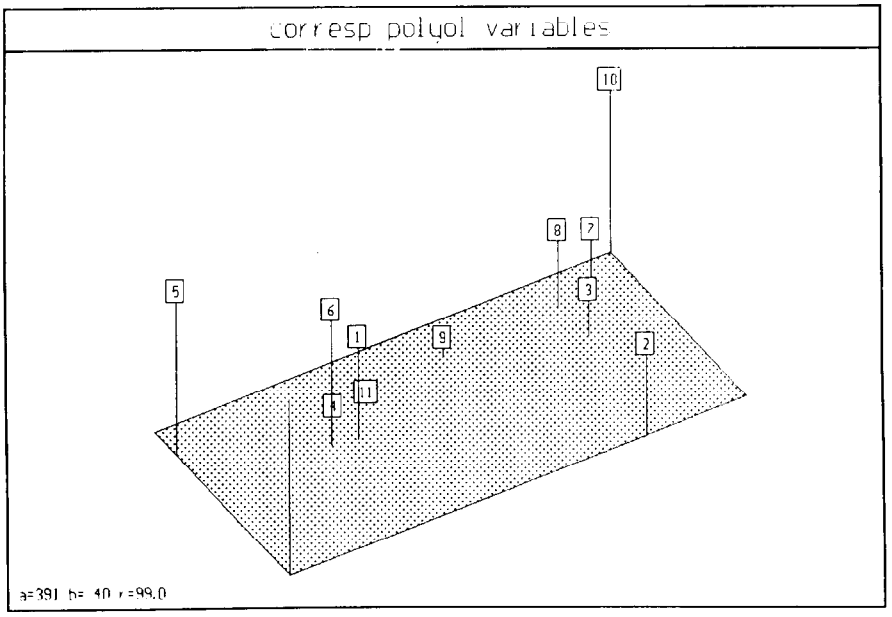

Fig. 2. Correspondence analysis of the carbohydrate data. A variable plot.

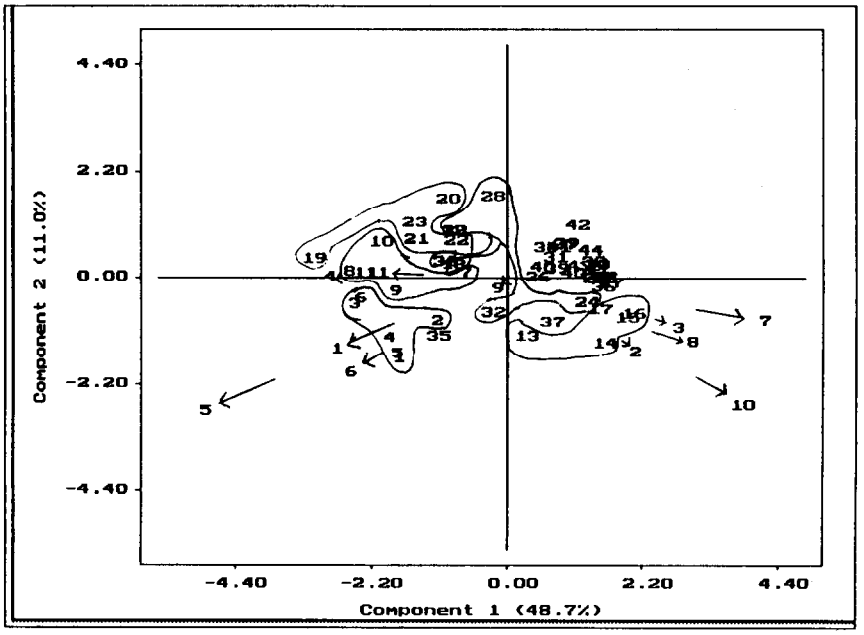

Fig. 3. Canonical PLS biplot of the two first axes explaining 59.7% of the variance in the **X** matrix. For identification of the different isolates and variables, see Tables 1 and 2. Variables are shown by arrows.

P. verrucosum, *P. chrysogenum* and *P. dipodomyis*, growing poor on creatine [32]) was clear from those ordinations (Fig. 1). It is seen from the corresponding variable plot (Fig. 2) that glucose, fructose and sucrose are still present after 16 days in cultures growing at low water activities and that erythritol and glycerol are the principal polyols accumulating in fungi growing on carbohydrate-rich substrates. Trehalose, arabinitol and the unknown compounds (a, b and c) are more prevalent in the fungi growing on lipid- and/or protein-rich substrates. Mannitol has little influence on this physiological classification. Thus the quantitative profile of carbohydrates contains some physiological and only indirectly taxonomical information.

A canonical partial least squares (PLS) biplot (Fig. 3) shows the same type of result as obtained by correspondence analysis, but now the differentiation between species is, of course, a little more pronounced. However, the seven clusters still overlap.

3.2. Analysis of the mixed data matrices

UPGMA cluster analysis of the mixed data matrix showed a clear separation between the seven species for most coefficients. This was most pronounced in coefficients involving subtraction (χ^2, Canberra and a little less pronounced using the Bray–Curtis or Manhattan coefficients [1,26]) and less pronounced using taxonomic or Euclidean distance (data not shown) or correlation coefficients (product-moment, cosine or Morista). The Renkonen coefficient and Penrose's size coefficients gave very 'poor' results. The same coefficients were used for PCO with an overlain minimum spanning tree (MST) (Figs. 4–6). It is seen that PCO using taxonomic distance takes the quantitative variables into account on the first three coordinate axes (Fig. 4), while the 'subtractive' coefficients emphasize the binary data strongly (Figs. 5 and 6). The correspondence analysis is not exactly the same as a PCO of the χ^2 distances, "since in CA the objects are weighted according to their frequencies" [26]. The result of the correspondence analysis of the mixed data matrix is shown in Fig. 7. A very clear result shows that families of secondary metabolites are strongly species-related characters. Not surprisingly a correspondence analysis of the binary characters alone showed that the grouping into species associated with high and low carbohy-

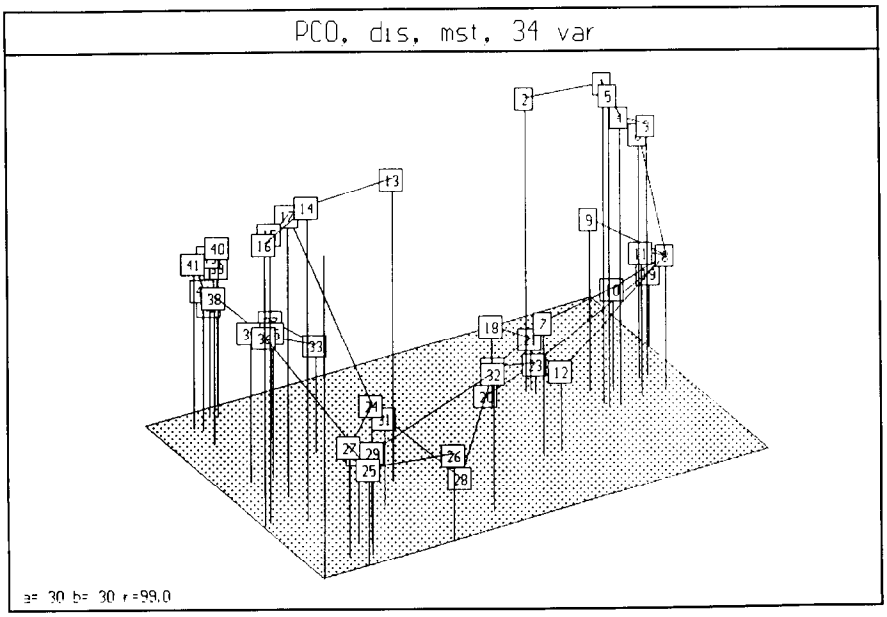

Fig. 4. Principal coordinate analysis of the mixed data matrix overlain by a minimum spanning tree using the taxonomic distance metric (first three coordinate axes).

drate substrates, found in the other analyses, was lost (data not shown), even though the seven species were very clearly separated.

Principal component analysis of the mixed data gave quite diffuse clusters, but all seven species could be separated using the first four coordinate axes (Fig. 8). Canonical PLS analysis gave more clear results. However, as in other methods based on Euclidean distance, it gave relatively more weight to the quantitative variables. In Fig. 9 the

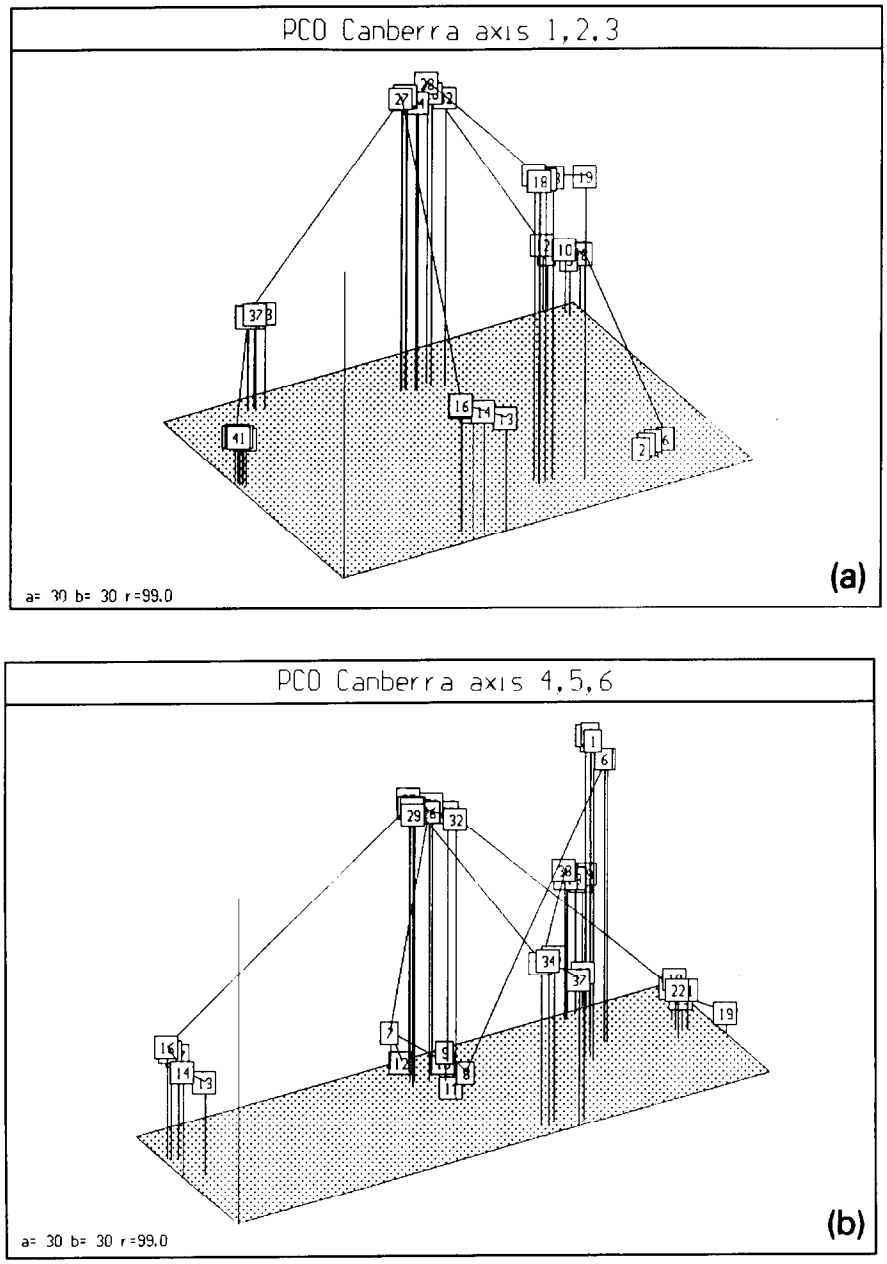

Fig. 5. PCO of the mixed data matrix overlain by a MST using the Canberra metric. Coordinate axes 1–3 (a) and 4–6 (b).

first three PLS axes are shown. Several strains are misplaced, but they are correctly placed along axes 4, 5 and 6 (Fig. 10 for axes 5 and 6).

Using only unique metabolite families (seven binary variables, one for each species) in the **X** matrix together with the quantitative carbohydrate data gave a little less clear separation between the seven taxa than the analysis involving all significant biosynthetic families. Correspondence analysis of that matrix (18 variables in all) placed *P. chrysogenum* close to *P. griseofulvum* and *P. crustosum* close to *P. echinulatum* on the

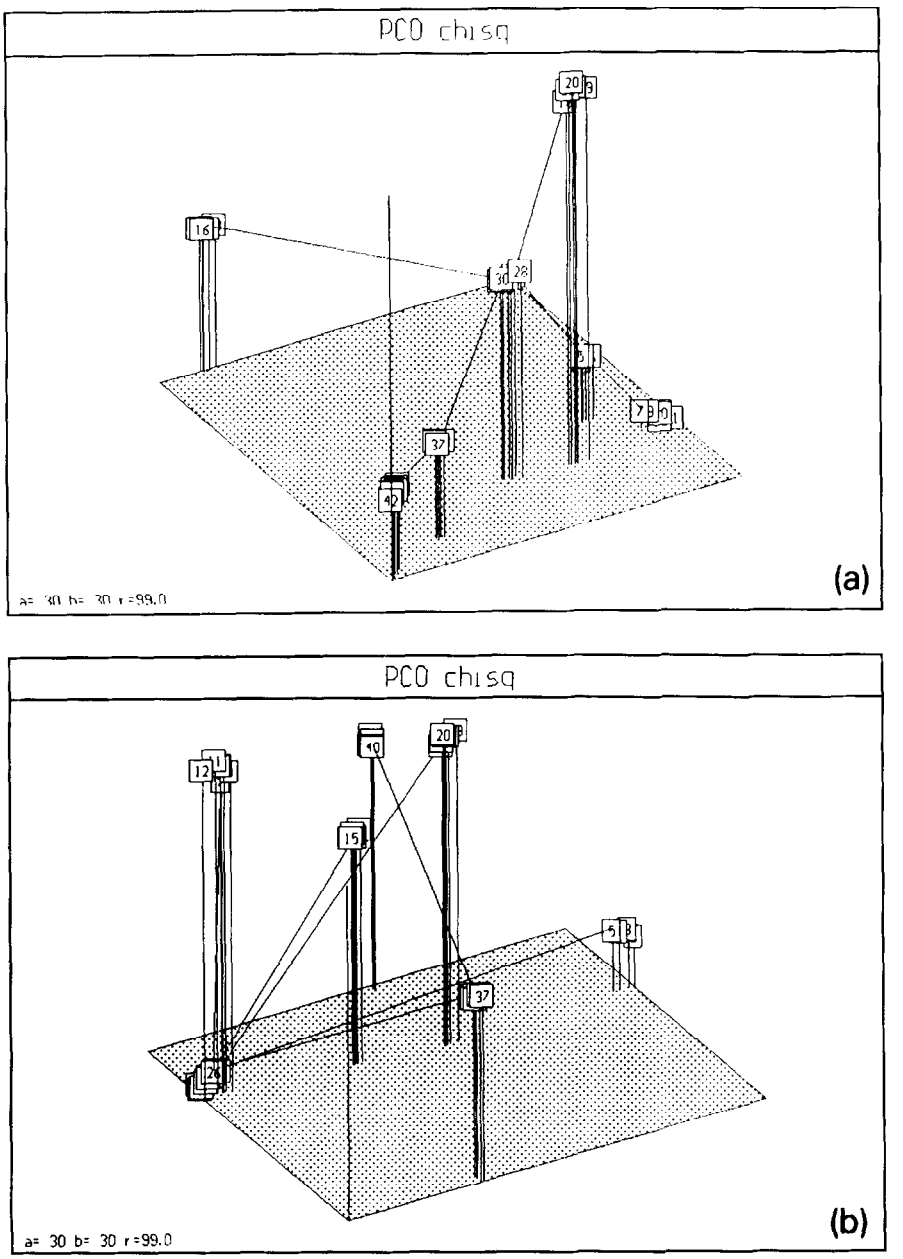

Fig. 6. PCO of the mixed data matrix overlain by a MST using the χ^2 distance. Coordinate axes 1–3 (a) and 4–6 (b).

first three axes (Fig. 11), but the seven taxa were clearly separated on axes 4–6 (Fig. 12). It should be noted that some of the binary characters, the penicillins, metabolite family "Ø", viridicatins, griseofulvins and roquefortines, were produced by two or three of the seven species. Unique secondary metabolites are more difficult to find in a matrix containing many more closely related species, but then unique two or three character combinations can be selected. The situation that

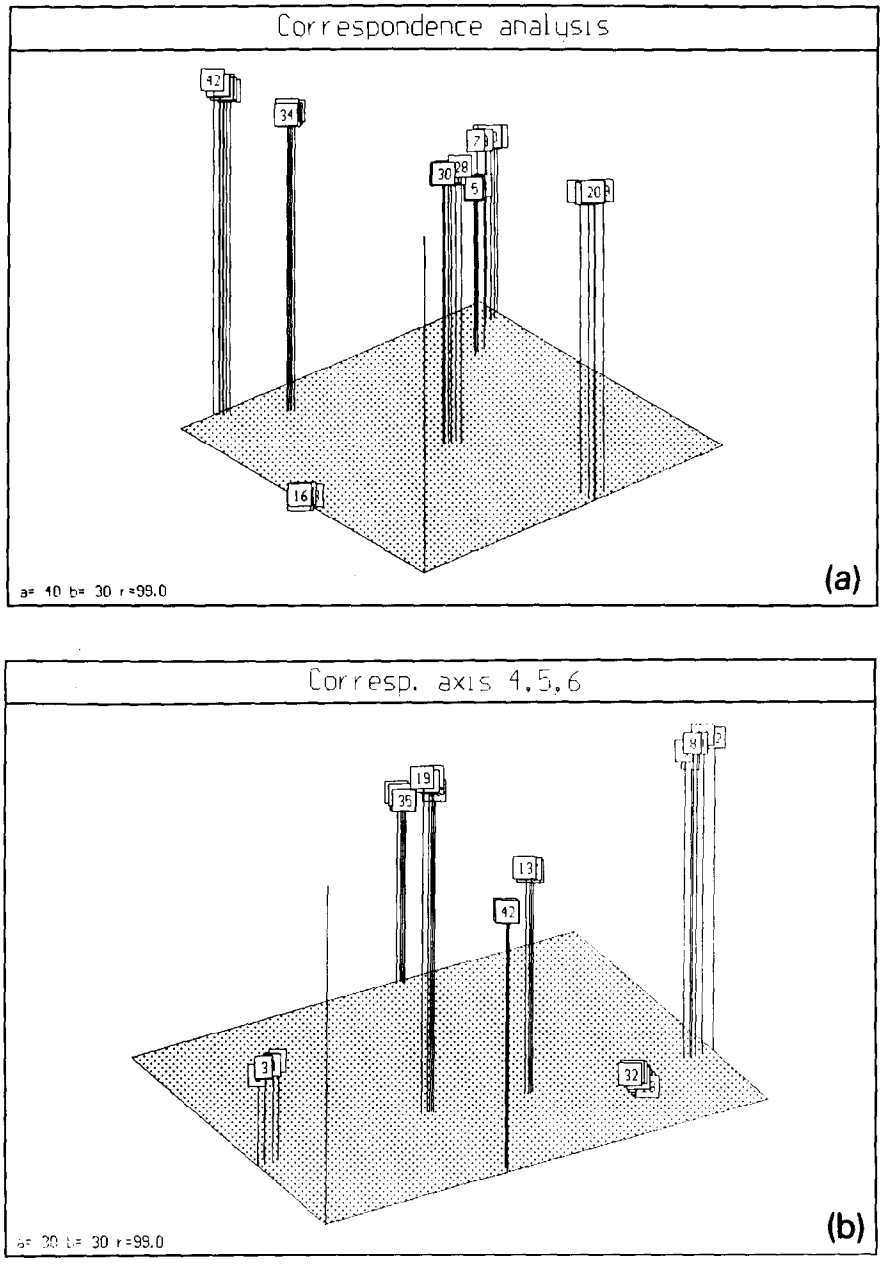

Fig. 7. Correspondence analysis of the mixed data matrix. Coordinate axis 1–3 (a) and 4–6 (b).

unique binary characters are known a priori and are the only binary characters selected is not very realistic, however, but the results (Figs. 11 and 12) do show that very few binary characters influence the final ordination much more than quantitative characters. In contrast binary characters,

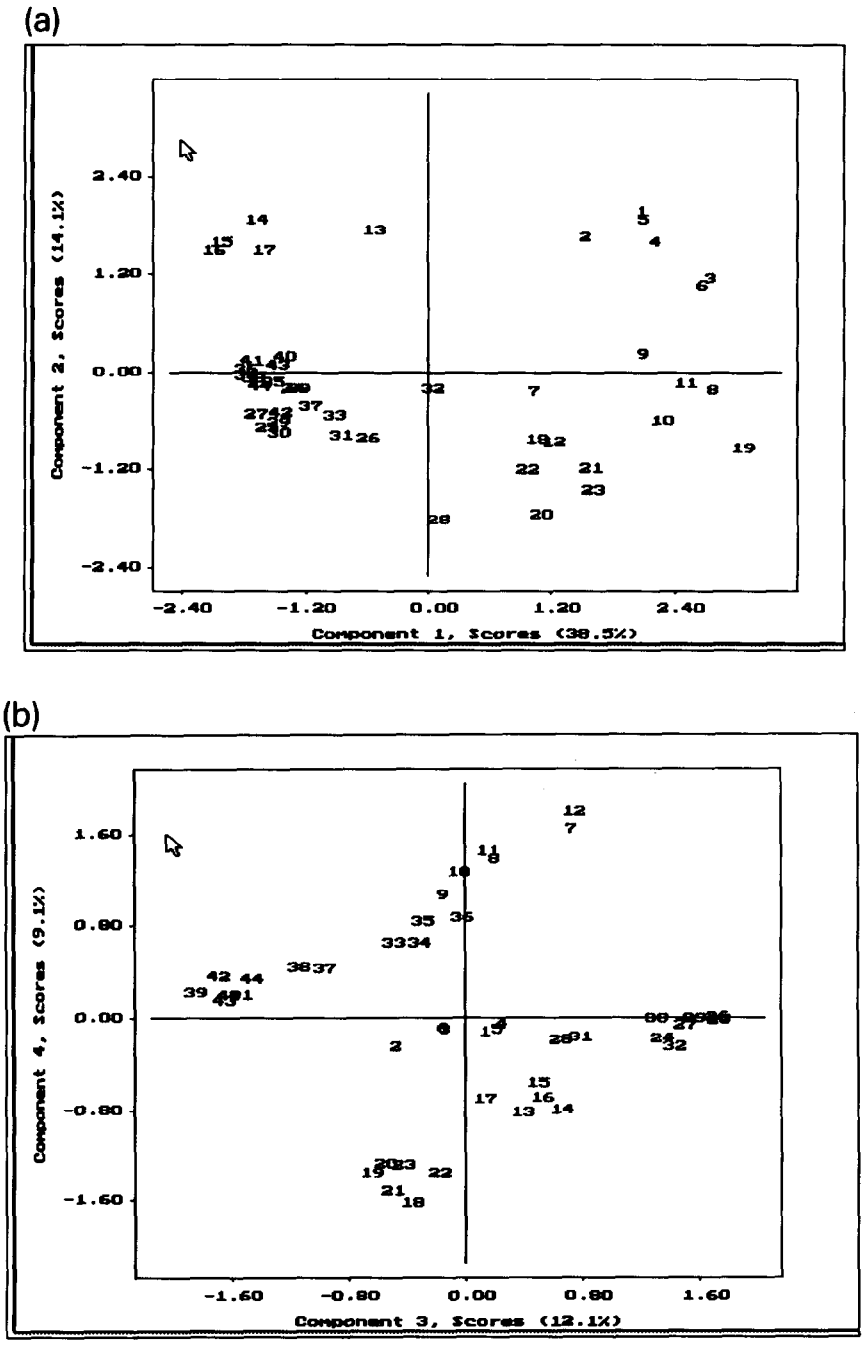

Fig. 8. Principal component analysis of the mixed data matrix. Axes 1 and 2 (a), and axes 3 and 4 (b).

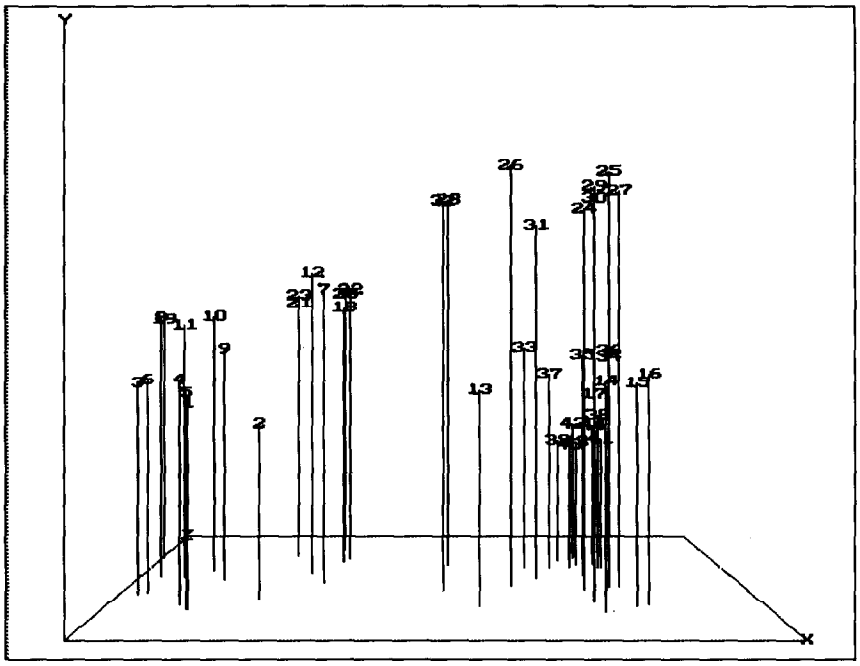

Fig. 9. The first three canonical PLS axes of the mixed data matrix. Note the intergrading isolates of *P. crustosum* and *P. aethiopicum*.

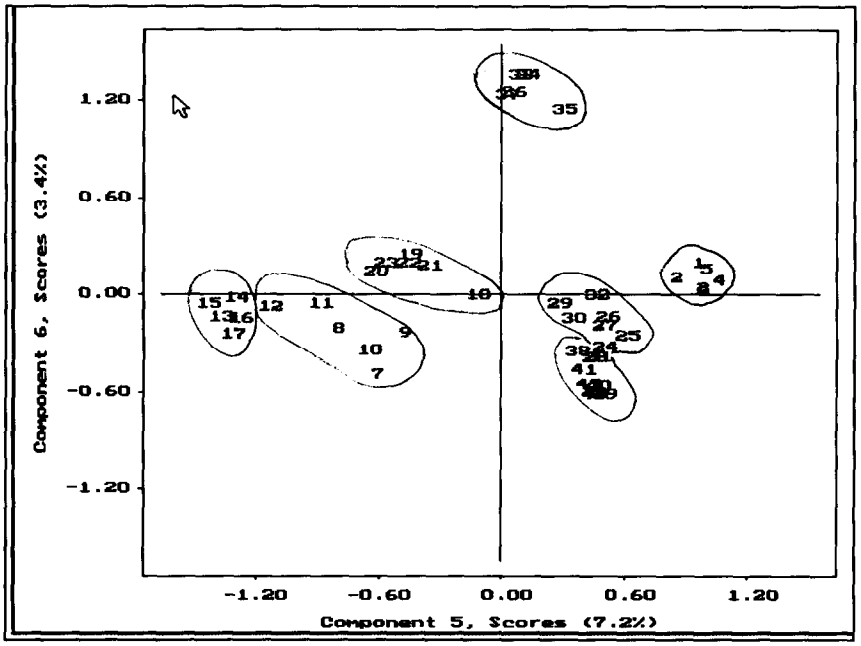

Fig. 10. Canonical PLS axes 5 and 6 of the mixed data matrix. The seven species are more 'correctly' separated on those axes.

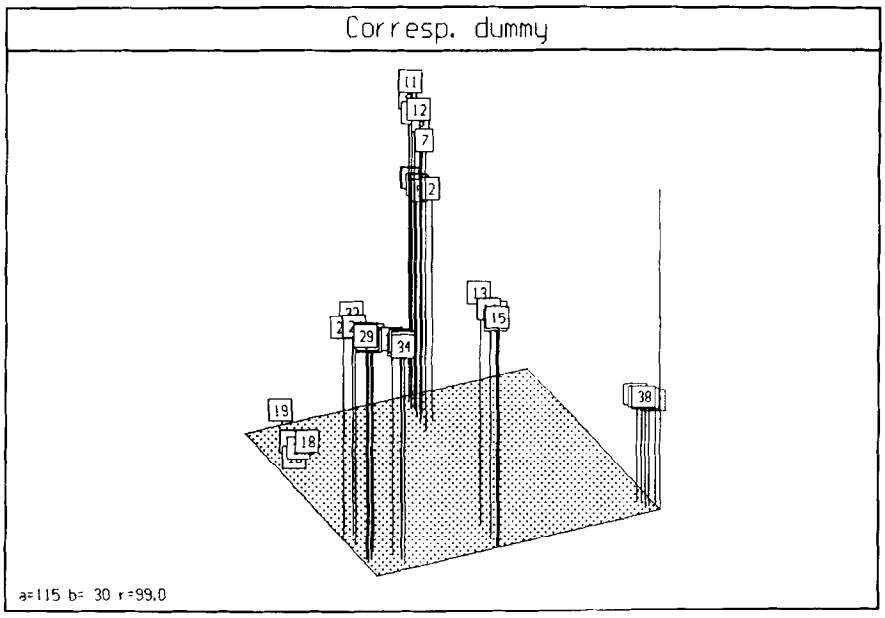

Fig. 11. Correspondence analysis (first three axes) of a data matrix with 11 quantitative variables and 7 binary unique secondary metabolite data.

shown later to be uninformative taxonomically, or quantitative characters that occasionally have the value zero, may give poor results.

Fuzzy clustering, using Euclidean distance, on the raw data (coefficient of fuzziness 2 in the program SIRIUS) also showed that the quantita-

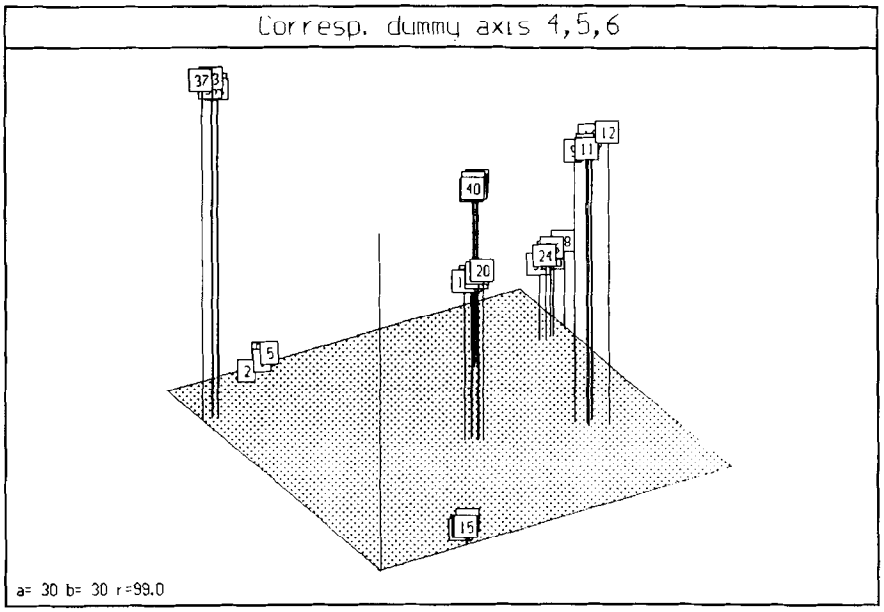

Fig. 12. Correspondence analysis axes 4, 5 and 6 on the same data as in Fig. 11.

Table 3
Fuzzy membership weights (2 decimals, percentage) of the 44 isolates of *Penicillium* based on the raw data and PLS scores for 2 or 7 clusters

Species	Raw data, No. of clusters		PLS scores, No. of clusters	
	2	7	2	7
P. echinulatum				
1	74,26	5,3,3,3,80,4,3	78,22	1,1,1,2,1,1,94
2	68,32	14,9,10,9,36,12,10	79,21	3,3,5,5,3,3,77
3	80,20	7,2,2,2,80,4,2	84,16	1,1,3,4,1,2,87
4	76,24	3,2,2,1,88,2,1	81,19	0,0,0,1,0,0,98
5	74,26	5,3,3,3,81,3,2	78,22	1,1,1,1,1,1,95
6	80,20	7,3,3,2,78,5,3	85,15	1,2,3,4,1,2,87
P. crustosum				
7	61,39	29,13,12,9,13,13,10	69,31	6,8,7,57,6,8,8
8	78,22	83,2,2,2,5,4,2	82,18	1,1,2,89,1,1,3
9	75,25	53,6,7,5,13,10,6	87,13	2,2,3,86,1,2,5
10	74,26	59,6,6,4,10,10,5	82,18	1,1,1,95,1,1,1
11	69,31	41,8,9,7,14,13,7	82,18	1,1,1,94,1,1,2
12	61,39	33,12,12,8,11,14,9	67,33	5,7,8,61,5,7,6
P. verrucosum				
13	43,57	8,11,11,43,9,8,10	40,60	7,7,6,5,61,8,6
14	32,68	4,8,8,65,4,4,7	29,71	1,1,1,0,96,1,1
15	29,71	3,7,8,68,3,4,7	25,75	1,1,0,9,95,1,0
16	27,73	2,4,4,83,2,2,4	24,76	1,2,1,1,93,2,1
17	30,70	3,6,7,71,3,3,7	26,74	1,1,0,0,96,1,0
P. aethiopicum				
18	65,35	11,10,9,7,11,43,9	67,33	3,3,83,3,2,3,3
19	73,27	19,7,7,6,16,37,7	79,21	6,6,51,14,5,6,13
20	63,37	6,5,5,3,4,73,5	63,37	2,2,88,2,1,2,2
21	71,29	9,6,6,4,7,62,6	72,28	0,0,97,1,0,0,1
22	66,34	5,4,4,3,4,77,4	65,35	1,1,92,1,1,1,1
23	70,30	5,3,3,2,3,82,3	71,29	1,1,96,1,0,1,1
P. griseofulvum				
24	25,75	4,60,11,9,4,5,8	19,81	2,92,1,1,2,2,1
25	27,73	4,64,11,7,4,5,7	22,78	2,91,1,1,2,3,1
26	37,63	7,48,13,8,6,9,8	30,70	3,81,3,3,3,5,2
27	23,77	2,78,7,5,2,3,4	18,82	2,91,1,1,2,3,1
28	48,52	12,24,14,8,9,20,12	40,60	11,36,17,10,7,13,7
29	25,75	3,70,9,6,3,4,5	19,81	1,95,1,1,1,2,0
30	23,77	3,71,9,5,2,4,5	17,83	1,95,1,1,1,2,0
31	28,72	6,46,14,9,5,8,12	18,82	5,79,3,2,4,6,2
32	47,53	10,36,14,10,10,12,9	41,58	7,50,9,7,8,11,8
P. chrysogenum				
33	26,74	5,10,56,6,4,6,13	21,78	5,4,3,2,3,82,2
34	18,82	2,6,77,4,2,2,7	15,85	3,2,1,1,2,92,1
35	19,81	3,8,71,5,2,3,9	14,86	1,1,0,0,96,0
36	20,80	3,10,64,7,3,4,10	14,86	4,3,1,1,3,87,1
37	27,73	6,11,40,9,5,8,21	18,82	7,3,2,2,3,81,1
P. dipodomyis				
38	20,80	4,10,22,9,4,6,45	14,86	90,2,1,1,2,4,1
39	27,73	2,4,7,3,2,3,79	23,77	94,1,1,1,1,2,1
40	26,74	3,5,9,5,3,3,72	20,80	96,1,0,0,1,1,0
41	20,80	1,2,5,2,1,1,88	18,82	95,1,1,0,1,2,0
42	27,73	5,10,18,8,5,7,48	22,78	91,2,1,1,1,3,1
43	23,77	1,2,5,2,1,2,87	20,80	98,0,0,0,0,1,0
44	22,78	4,8,15,7,3,5,59	18,82	98,0,0,0,0,0,0

tive data dominated (Table 3). For two clusters, species with high membership weights in group one were those that grow on lipid- and protein-rich substrates (creatine positive fungi) and those in group two are prevalent on carbohydrate-rich substrates such as cereals (creatine negative fungi) [32]. It was examined whether two and up to nine clusters were optimal and two clusters emerged as the optimal number. However, the membership weights in 7 clusters always placed each of the 44 isolates in the correct species. Better results were obtained when applying fuzzy clustering on the PLS scores (first six components), even for the two-cluster separation (Table 3).

In most cases a large proportion of the variance could be explained by the first three component axes and up to 100% by six components (Table 4), and in all cases the eigenvalues were larger than those calculated using the 'broken stick model'. The latter test of significance of the eigenvalues is very gross, however [31]. It was expected that a maximally informative data set would explain close to 100% of the variation on six axes or less, as seven species were analyzed. By comparing the percentage of variance explained by the binary variables alone it is seen that these variables are very strongly emphasised in the analyses that also involve both quantitative and qualitative variables, if the distances are calculated using subtraction (χ^2, Canberra, Bray–Curtis and Manhattan distance). For ordinary distance coefficients (Euclidean, squared Euclidean and taxonomic distance) and correlation coefficients (product–moment correlation, cosine of angle between vectors, Morista and Morista 2) the quantitative variables play a more important role, but the resulting ordinations are still dominated by the qualitative variables. It is obvious that in principal component analysis, in the closely related PCO using taxonomic distance and in PLS, a major part of the information in the quantitative variables is modeled by the first component. For example in the PLS analysis 35% of the variation in **X** is explained by the first component but only 13% of the dummy **Y** matrix, while 94% of **Y** is explained after six PLS components.

Canonical variates analysis (CVA), often used in attempts to discriminate between taxa, is a more strict statistical method based on normal distributions. As it was developed for cases

Table 4
Variance explained (in percentage) on the first six components (cumulative) in different ordinations of the mixed data matrices

	Component					
	1	2	3	4	5	6
Quantitative data set (11 variables)						
PCA	55	69 [a]	78 [a]	86 [a]	90 [a]	93 [a]
PLS, **X**	49	60	67	69 [a]	74 [a]	78 [a]
Y	12	21	28	34 [a]	38 [a]	40 [a]
Whole data set (34 variables)						
PCO, taxonomic distance	39	53	65	74	81	86
PCA, Euclidean distance	39	53	65	74	81	86
PCO, Canberra	30	54	75	90	98	100
PCO, χ^2	22	42	61	77	91	98
CA	22	42	61	77	91	98
PLS, **X**	35	45	58	67	74	78
Y	13	32	46	62	78	94
Binary data set (23 variables)						
CA	22	42	62	79	92	100
Broken stick model [26] [b]	10	18	24	30	35	40

[a] Not significant according to crossvalidation of the whole data set.
[b] In this test the percentage explained can be regarded as significant as long as it is larger than the broken stick model values (Jollife [31]).

"where any one character would not suffice to discriminate between groups with a sufficiently small percentage of misclassification" [33], it is not appropriate in this type of data set where some of the families of secondary metabolites proved to be perfect discriminators. For example penitrems (or the combination of penitrem A and terrestric acid) are only found in *Penicillium crustosum* and these metabolites are not found in any of the other taxa in this study. CVA is more appropriate when taxa are described by variables that are overlapping, taken one at a time [1,33–37]. In larger data sets, however, the combination of penitrem A and terrestric acid has been found, for example in *P. albocoremium* [21], so the full profile of families of secondary metabolites has to be taken into account. Thus in order not to use a priori knowledge, all secondary metabolites found have to be included in the multivariate analyses. Canonical correspondence analysis (CCA) could have been used for this data set, but the results of the correspondence analysis showed that the seven species were already optimally separated (Fig. 7a and b). CCA may be of relevance for finding good discriminators in more complex data sets.

4. Conclusion

Characters rich in information content, such as families of secondary metabolites that are present or not (binary characters) dominate quantitative characters in chemometric analyses of mixed chemical data matrices, especially when using distance coefficients involving subtraction, but also when using Euclidean distance (PCA, PLS). Thus if the binary characters are considered of particular importance, correspondence analysis is recommended. Even if quantitative data are included in the data matrix, CA shows clear species differences. Thus even though Sneath and Sokal [1] and Vogt [3] have called secondary metabolites episemantic molecules, defining them as "simple chemical substances with relatively little information about the organism" [1,3], these are exactly the molecules that yield the most phylogenetic information and clear-cut separations of species [19]. If quantitative and qualitative characters are considered to be of equal importance, however, PCO using Gower's general coefficient or PCA and its constrained form canonical PLS could be used. For data with a known statistical distribution and a larger number of objects than variables (a kind of data matrix that is more rare in chemistry, ecology and taxonomy) more strictly statistical procedures based on discriminant analysis can be used [9]. A simple and straightforward chemometric evaluation of mixed data could apparently be based on correspondence analysis only. For special purpose classifications and for discrimination, the SIMCA method of classification can be used maybe in conjunction with canonical PLS analysis [28,29]. CA, however, has the further advantage that the classification need not be known a priori. Phenetic analysis (numerical taxonomy, the biological equivalent of chemometrics [1,38,39]) has often been criticised for pretending to be objective yet offering a plethora of cluster analysis methods and similarity/distance coefficients [40,41] and ordination methods. However, it is apparently sufficient to use CA for the analysis of most types of data matrices based on differentiation data for practical applications, especially when the matrices are of relatively low rank compared to the number of variables. An extra advantage of CA is that it is computationally very fast.

References

[1] P.H.A. Sneath and R.R. Sokal, *Numerical taxonomy*, Freeman, San Fransisco, CA, 1973.
[2] J.A. Harris and F.A. Bisby, Classification from chemical data, in F.A. Bisby, J.G. Vaughan and C.A. Wright (Editors), *Chemosystematics: Principles and Practice*, Academic Press, London, 1980, pp. 305–327.
[3] N.P. Vogt, Soft modelling and chemosystematics, *Chemometrics and Intelligent Laboratory Systems*, 1 (1987) 213–231.
[4] I.E. Frank and J.H. Friedman, A statistical view of some chemometrics regression tools, *Technometrics*, 35 (1993) 109–135.
[5] S. de Jong and H.A.L. Kiers, Principal covariates regression. Part I. Theory, *Chemometrics and Intelligent Laboratory Systems*, 14 (1992) 155–164.
[6] H. Martens and H. Russwurm (Editors), *Food Research*

and Data Analysis, Applied Science Publishers, London, 1983.

[7] B.R. Kowalski (Editor), *Chemometrics. Mathematics and Statistics in Chemistry*, Reidel, Dordrecht, 1983.

[8] J.C. Gower, A general coefficient of similarity and some of its properties, *Biometrics*, 27 (1971) 857–871.

[9] W.J. Krzanowski, The location model for mixtures of categorical and continous variables, *Journal of Classification*, 10 (1993) 25–49.

[10] A. Gifi, *Nonlinear Multivariate Analysis* (reprinted with corrections), Wiley, Chichester, 1991.

[11] S. Nishisato, *Analysis of Categorical Data: Dual Scaling and its Applications*, University of Toronto Press, Toronto, 1980.

[12] M.J. Greenacre, *Theory and Applications of Correspondence Analysis*, Academic Press, London, 1984.

[13] W.J. Heiser, Joint ordination of species and sites: the unfolding technique, in P. Legendre and L. Legendre (Editors), *Developments in Numerical Ecology*, Springer, Berlin, 1987, pp. 189–221.

[14] C.J.F. ter Braak and I.C. Prentice, A theory of gradient analysis, *Advances in Ecological Research*, 18 (1988) 271–317.

[15] C.J.F. ter Braak, Canonical correspondence analysis: a new eigenvector technique for multivariate direct gradient analysis, *Ecology*, 67 (1986) 1167–1179.

[16] C.J.F. ter Braak, The analysis of vegetation–environment relationships by canonical correspondence analysis, *Vegetatio*, 69 (1987) 69–77.

[17] H.G. Gauch, *Multivariate Analysis in Community Ecology*, Cambridge University press, Cambridge, 1982.

[18] R.H.G. Jongman, C.J.F. ter Braak and O.F.R. van Tongeren, *Data Analysis in Community and Landscape Ecology*, Pudoc, Wageningen, 1987.

[19] J.C. Frisvad, Chemometrics and chemotaxonomy: a comparison of multivariate statistical methods for the evaluation of binary fungal secondary metabolite data, *Chemometrics and Intelligent Laboratory Systems*, 14 (1992) 253–269.

[20] J.C. Frisvad, The connection between the penicillia and aspergilli and mycotoxins with special emphasis on misidentified isolates, *Archives of Environmental Contamination and Toxicology*, 18 (1989) 452–467.

[21] J.C. Frisvad and O. Filtenborg, Terverticillate penicillia: chemotaxonomy and mycotoxin production, *Mycologia*, 81 (1989) 837–861.

[22] B. Bjerg, O. Olsen, K.W. Rasmussen and H. Sørensen, New principles of ion-exchange techniques suitable to sample preparation and group separation of natural products prior to liquid chromatography, *Journal of Liquid Chromatography*, 7 (1984) 691–707.

[23] J.C. Frisvad and U. Thrane, Standardized high-performance liquid chromatography of 182 mycotoxins and other fungal metabolites, based on alkylphenone retention indices and UV–VIS spectra (diode array detection), *Journal of Chromatography*, 404 (1987) 195–214.

[24] J.C. Frisvad, The use of high-performance liquid chromatography and diode array detection in fungal chemotaxonomy based on profiles of secondary metabolites, *Botanical Journal of the Linnean Society*, 99 (1989) 81–95.

[25] J.C. Frisvad and U. Thrane, Liquid column chromatography of mycotoxins, in V. Betina (Editor), *Chromatography of Mycotoxins* (Journal of Chromatography Library, Vol. 54), Elsevier, Amsterdam, 1993, pp. 253–372.

[26] F.J. Rohlf, NTSYS-pc, Numerical taxonomy and multivariate analysis system, version 1.80, Exeter Software, Setauket, 1993.

[27] J. Podani, *SYN-TAX-pc, Computer Programs for Multivariate Data Analysis in Ecology and Systematics, Version 5.0, User's Guide*, Scientia Publishing, Budapest, 1993.

[28] T.V. Karstang and O. Kvalheim, SIRIUS, Version 2.2, Pattern Recognition Systems, Bergen, 1990.

[29] SIMCA-R, Multivariate modelling and analysis, version 4.4, Umetri, Umeå, 1992.

[30] C.J.F. ter Braak, CANOCO, A Fortran program for canonical community ordination by [partial][detrended] [canonical] correspondence analysis and redundancy analysis, version 2.1, Agricultural Mathematics Group, Wageningen, 1988 (update, version 3.10 and 3.12, 1990).

[31] I.T. Jollife, *Principal Component Analysis*, Springer, New York, 1986.

[32] J.C. Frisvad, Modifications on media based on creatine for use in *Penicillium* and *Aspergillus* taxonomy, *Letters in Applied Microbiology*, 16 (1993) 154–157.

[33] R.R. Sokal, Statistical methods in systematics, *Biological Reviews*, 40 (1965) 337–391.

[34] E.S. Gilbert, On discrimination using qualitative variables, *Journal of the American Statistical Association*, 63 (1963) 1399–1412.

[35] S. Feldman, D.F. Klein and G. Honigfeld, A comparison of successive screening and discriminant function techniques in medical taxonomy, *Biometrics*, 25 (1969) 725–734.

[36] T.W. Kurzynski, Generalized distance and discrete variables, *Biometrics*, 26 (1970) 525–534.

[37] K.C. Kim, B.W. Brown, Jr. and E.F. Cook, A quantitative taxonomic study of the *Hoplopleura hesperomydis* complex (Anoplura, Hoplopleuridae), with notes on a posteriori taxonomic characters, *Systematic Zoology*, 15 (1966) 24–45.

[38] J.C. Gower, Numerical techniques as an aid to objectivity, in D.L. Hawksworth (Editor), *Prospects in Systematics*, Clarendon Press, Oxford, 1988, pp. 234–251.

[39] P.H.A. Sneath, The phenetic and cladistic approaches, in D.L. Hawksworth (Editor), *Prospects in Systematics*, Clarendon Press, Oxford, 1988, pp. 252–273.

[40] E. Mayr, Recent historical developments, in D.L. Hawksworth (Editor), *Prospects in Systematics*, Clarendon Press, Oxford, 1988, pp. 31–43.

[41] P.L. Forey, C.J. Humphries, I.J. Kitching, R.W. Scotland, D.J. Siebert and D.M. Williams, *Cladistics. A Practical Course in Systematics* (The Systematics Association Publication No. 10), Clarendon Press, Oxford, 1992.