ELSEVIER

Journal of Pharmaceutical and Biomedical Analysis
28 (2002) 537–548

JOURNAL OF
PHARMACEUTICAL
AND BIOMEDICAL
ANALYSIS

www.elsevier.com/locate/jpba

Characterization of the similarity of chemical compounds using electrospray ionization mass spectrometry and multivariate exploratory techniques

V. Schoonjans [a], N. Taylor [b], B.D. Hudson [b], D.L. Massart [a,*]

[a] *ChemoAC, Pharmaceutical Institute, Vrije Universiteit Brussel, Laarbeeklaan 103, B-1090 Brussels, Belgium*
[b] *Glaxo Wellcome, Medicines Research Centre, Compound Diversity Unit, Stevenage, Herts SGI 2NY, UK*

Received 23 June 2001; received in revised form 3 October 2001; accepted 6 October 2001

Abstract

The applicability and usefulness of electrospray mass spectrometry to the analysis of a small data set of 52 synthetic substances to probe their molecular similarity/diversity has been demonstrated. The first stage was the reduction of the data by Principal Component Analysis (PCA), to visualize the structure of the data. Sequential Projection Pursuit (SPP) was applied to detect outlying objects. Hierarchical cluster analysis was employed to produce a dendrogram, using group-average linkage clustering. Finally, the cluster results of spectral data were compared with that of structural fingerprints and an expert's classification by using the similarity measure of Wallace.

Keywords: Similarity; Mass spectra; Fingerprint; PCA; SPP; Hierarchical upgma-clustering

1. Introduction

Within the pharmaceutical industry, high-throughput screening methods are now routinely used to identify novel active compounds from naturally occurring materials. The ability to quantify the chemical diversity of the newly synthesized samples is here of great significance [1]. Over the years, many different quantitative measures of structural similarity as well as a wide variety of descriptors to attempt to represent chemical structures have been developed [2,3]. Studies have reported that two-dimensional structural fingerprints perform well for diversity studies. A molecular fingerprint consists of a binary bit string where a 'one' (on bit) signals the presence of a certain molecular fragment and a 'zero' (off bit) indicates its absence. [2,4,5].

Multivariate techniques have been frequently used to determine the molecular diversity of thousands of individual compounds with known structure. However, natural product extracts are initially generated as complex mixtures consisting of multiple unknown compounds, which makes these techniques inapplicable. Consequently, the

* Corresponding author. Tel.: +32-2-477-4737; fax: +32-2-477-4735.
E-mail address: fabi@vub.vub.ac.be (D.L. Massart).

0731-7085/02/$ - see front matter
PII: S0731-7085(01)00677-X

different compounds must be represented by other descriptors, e.g. experimental parameters. Many studies and applications use mass spectrometry in combination with other analytical techniques as a sensitive tool in the analysis of drug compounds and metabolites of plant or microorganism origin [6]. We have shown [7] that electron impact mass spectrometry can be used, in combination with other analytical techniques, for assessing similarity/diversity. However, ES-MS is a much softer technique than electron impact mass spectrometry so that there are fewer fragments in the spectra [8]. Therefore, it is not evident that this technique can provide enough information for assessing similarity/diversity of chemical compounds. The aim of this paper is to investigate whether chemometric methods (cluster analysis or principal component analysis (PCA)) can be successfully applied to electrospray mass spectra for characterizing the similarity/diversity of chemical structures.

Table 1
List of chemical compounds

D-maltose
D-glucose
Lactose
D-allose
D-galactose
D-mannose
Cellobiose
Saccharine
Penicilline
Tetracycline
L-aspartic acid
L-asparagine
D-leucine
L-isoleucine
D-fenylalanine
L-tyrosine
Amphetamine
Ephedrine
Dopamine
Serotonine
Melatonine
Fenfluramine
Oxeladin
Lidocaine
Digitoxigenin
Digitoxin
Testosteron
Androsteron
Progesteron
Estradiol
Cholesterol
Prenalterol
Acebutolol
Oxprenolol
Propranolol
Nadolol
Atenolol
Metoprolol
Timolol
Benzylphenol
Menthol
Camphor
Caffein
Pentoxyfyllin
Purine
Lobeline
Amiodarone
Miconazole
Sulfapyridine
Lormetazepam
Flurazepam

2. Data

A small data set of 52 chemical compounds with known structure was selected from the literature in such a way that it consists of some groups of structurally and pharmacologically similar compounds, e.g. β-blockers, steroids, amino acids, sugars and some compounds that were selected at random. They are listed in Table 1. Most compounds were already used in a similar study about assessing similarity/diversity by electron impact mass spectrometry [7].

Electrospray mass spectra for all compounds in the data set were provided by Glaxo Wellcome, UK. Each spectrum was available as a list containing the mass-to-charge values (m/q values) and corresponding intensities of each fragment ion. The ES-mass spectra mostly contain the protonated molecules with relatively little fragmentation in the mass range. A data matrix (52×616) was created from these spectra, where the rows correspond to the 52 compounds and the columns to the 616 m/q ratio. The values in the matrix are the fragment ion intensities and range from 0 (no peak) to 100 (most important peak).

The 2D structural fingerprints for the same substances were obtained using the Daylight Clustering Software.

3. Methods

3.1. Transformation of the data

The relative intensities from the electrospray mass spectra were preprocessed by means of a logarithmic transformation to remove the effect of differences in variance between the variables [9].

3.2. Principal component analysis

PCA is a typical display method that is often applied to reduce the size of the space of the variables whilst preserving most of the variance and, therefore, it is a useful tool for data structure interpretation and visualization [10]. The raw and log transformed electrospray mass spectra were analyzed by means of PCA.

3.3. Sequential projection pursuit

Sequential Projection Pursuit (SPP) is a method that detects outlying observations in the data more easily than PCA [11]. The method is applied on the raw and log transformed mass spectra.

3.4. Hierarchical cluster analysis

A hierarchical clustering method produces a classification in such a way that small groups of very similar objects are included into larger groups of more diverse molecules [12]. The method used here is based on the unweighted pair-group average method, which consists of finding a similarity between two clusters defined as the average of all the similarities belonging to these clusters. The result is represented in a dendrogram [10].

3.5. Comparison of hierarchical classifications

Numerous measures to quantitatively define the similarity between two different clustering of the same set of objects are described in the literature. In this work, the measure of Wallace, s_w (1983), is applied [13]. A more detailed explanation of the methodology is given elsewhere [7].

4. Results and discussion

4.1. Principal component analysis

To obtain a first overview of the structure in the data and to detect major patterns and groups, a PCA was performed on the raw electrospray mass spectra. This PCA resulted in four principal components explaining 32% of the total variance. The first PC was found to describe 11.5% of the variance and the second, third and fourth PC 8.4, 6.5 and 5.6%, respectively. The score plot of PC1 against PC2, PC3 and PC4 and PC2 against PC3 is shown in Fig. 1(a–d), respectively. The corresponding loadings are plotted in Fig. 2(a–c).

In the score plots, V-shaped structures are observed. For instance in Fig. 1(d), all the β-blockers are on one line in the lower right part of the plot, and the sugars appear on another in the upper right part of the plot. The steroids are grouped together in the central region of the plot. The amino acids appear somewhat more dispersed over the central left region. In the score plot of PC1–PC3 (Fig. 1b), one can observe a separation of groups of compounds into three directions. The sugars appear in the upper part of the plot, the amino acids in the right part and the β-blockers in the lower part of the same plot.

An inspection of the scores (Fig. 1) and loadings (Fig. 2) shows that the first PC is related to the total intensity of the fragment ions of the compounds investigated or PC1 equals (loading × intensity) since all loadings are positive. However, fragment ions at low m/q values have higher loadings than fragment ions of high mass. Correspondingly, substances that have high intensity fragment ions of low mass are located in the right part of Fig. 1(a), such as for example compound numbers 13 (D-leucin), 14 (L-isoleucin), 41 (menthol) and 42 (camphor). The second PC discriminates compounds with basepeak and/or very intense peaks at m/q 85 (variable 36), 116 (variable 67), 145 (variable 96), for example compounds number 1 (D-maltose), 3 (lactose), 7 (cellobiose) and 35 (propranolol) from the other compounds. This is also seen in the loading plot of Fig. 2(b) where these variables have a high positive loading. PC3 describes the contrast be-

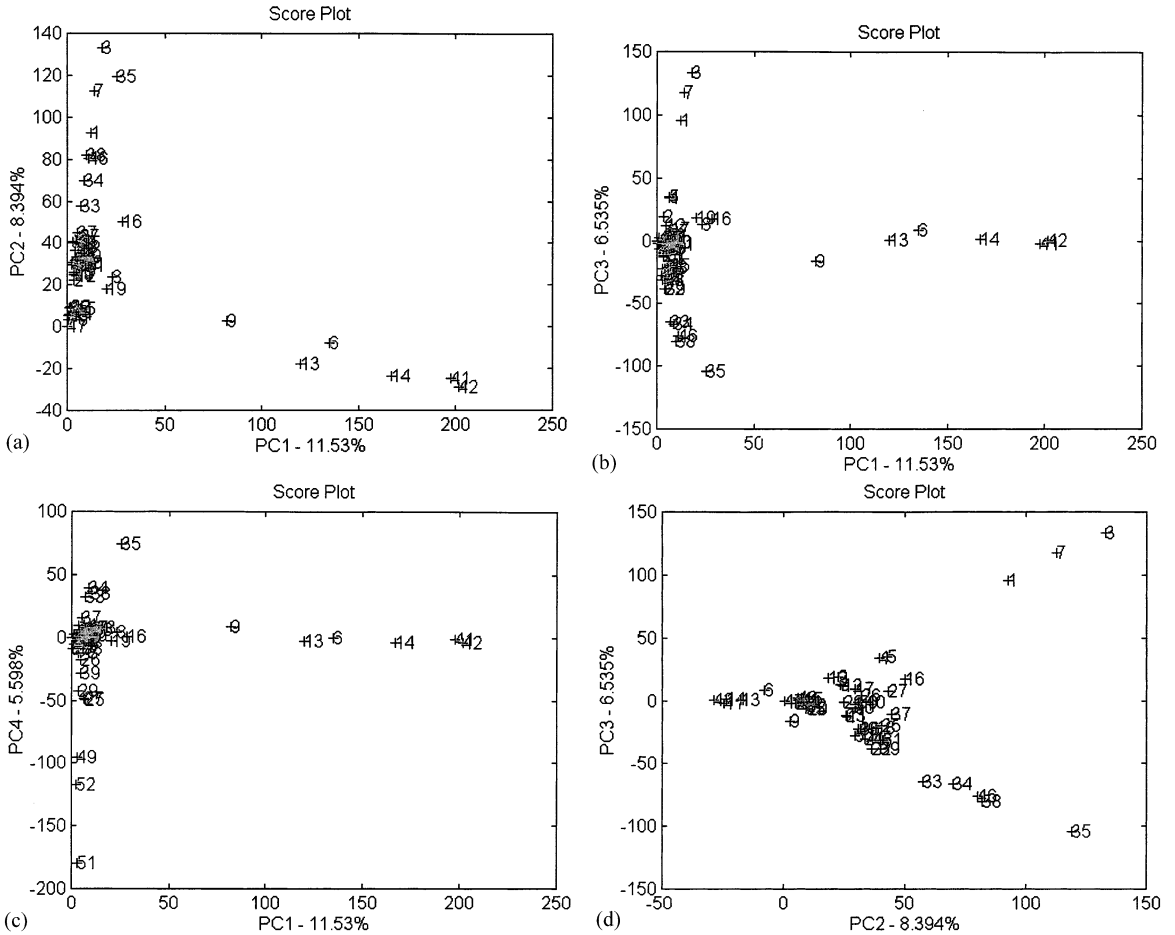

Fig. 1. (a) Score plot from the PCA of the raw electrospray mass spectra, showing PC2 against PC1. For the numbering of the compounds, see Table 1. (b) Score plot from the PCA of the raw electrospray mass spectra, showing PC3 against PC1. Notation as in (a). (c) Score plot from the PCA of the raw electrospray mass spectra, showing PC4 against PC1. The numbering of the compounds is the same as in (b). (d) Score plot from the PCA of the raw electrospray mass spectra, showing PC3 against PC2. The numbering of the compounds is the same as in (c).

tween substances with prominent peak at m/q 85 (variable 36) and very intense peaks at m/q 116 (variable 67). Correspondingly, compounds with basepeak at m/q 85 appear in the upper half of the Fig. 1(b), such as, for instance, compound numbers 1 (D-maltose), 3 (lactose), 7 (cellobiose), whereas compounds with very intense peaks at m/q 116 appear in the lower part of the same figure, for example compound numbers 33 (acebutolol), 34 (oxprenolol), 35 (propranolol), 38 (metoprolol). Fragment peaks of m/q 85 can be attributed to $(C_4H_5O_2)^+$ and are characteristic of sugars, whereas fragment ion peaks of m/q 116 $(C_6H_{14}NO)^+$ are typical of β-blockers. The fourth PC separates compounds that are particularly characterized by basepeak or very intense peaks at very high m/q values, for instance compound numbers 49 (nicardipin), 51 (lormetazepam), 52 (flurazepam) from the rest.

These results demonstrate that ES-mass spectra, in spite of the small amount of fragment ions, indeed provide some valuable information characteristic for a particular molecular structure since clusters of similar compounds are formed in the resulting plots.

The log transformed mass spectral data were also subjected to PCA, which resulted in three principal components that now explain 92.8% (89.7, 1.9 and 1.2%) of the total variance, i.e. much more than for the raw data. Fig. 3(a–c) show the scores of PC2 against PC1, PC3 against PC1 and PC3 against PC2, respectively. The corresponding loadings are plotted in Fig. 4(a–b), respectively.

An examination of the score plot in Fig. 3(c) demonstrates that all amino acids are grouped together in the central left part of the plot. Most sugars appear in the upper left part while the group of β-blockers is situated in the central region of the same figure. The steroids are dispersed over the central and right region of the plot. Also, serotonin and melatonin are located near each other in the upper region. Camphor and menthol are closely clustered in the lower left part and flurazepam and lormetazepam lie near each other in the lower right part of Fig. 3(c).

Looking at the score plots and the loading plots shows that the first PC is related to the degree of total mass spectral fragmentation intensity of the compounds investigated, since compounds that are characterized by many intense fragment ions at low m/q values are situated in the right part of Fig. 3(a). Compounds that show little fragmenta-

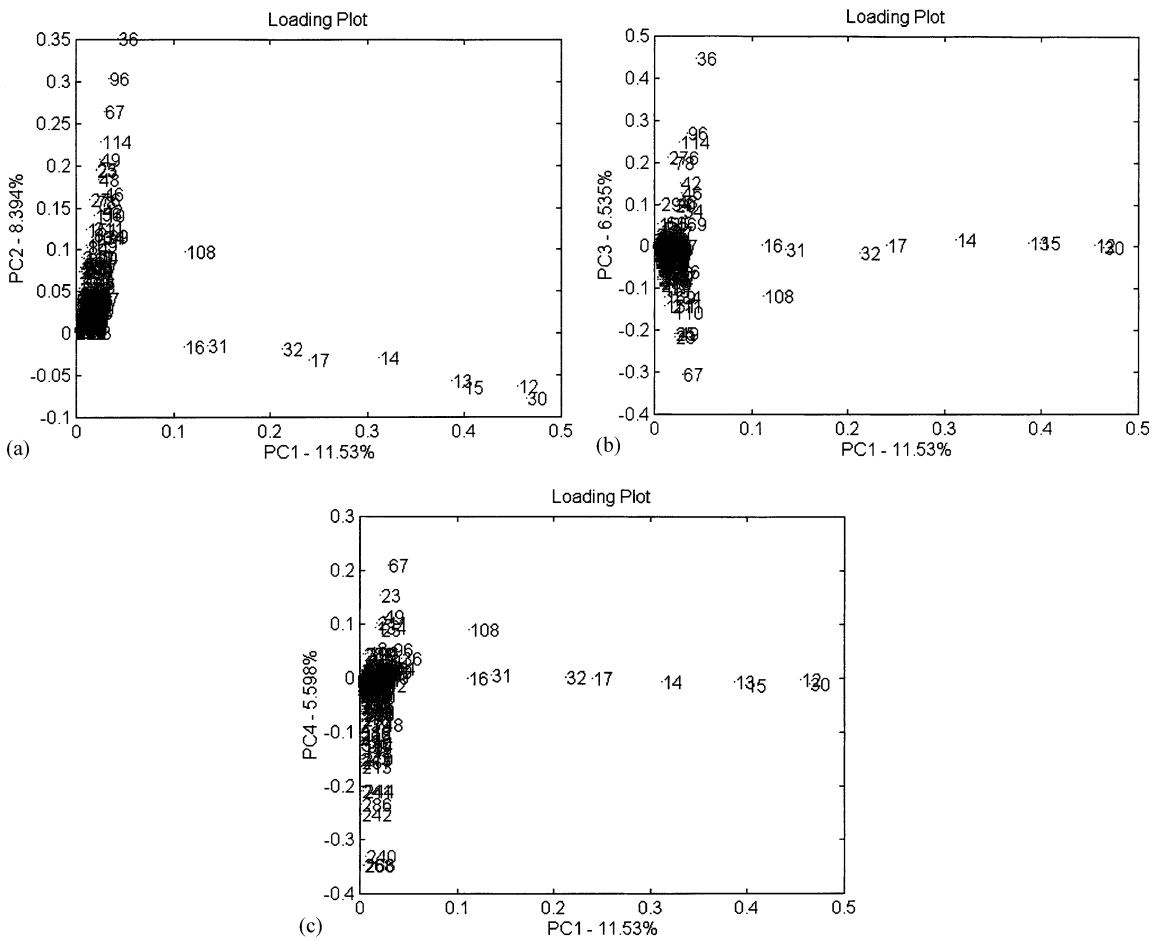

Fig. 2. (a) Loading plot from the PCA of the raw electrospray mass spectra. The second loading vector is plotted versus the first loading vector. (b) Loading plot from the PCA of the raw electrospray mass spectra, with the third loading vector plotted against the first loading vector. (c) Loading plot from the PCA of the raw electrospray mass spectra, with the fourth loading vector plotted against the first loading vector.

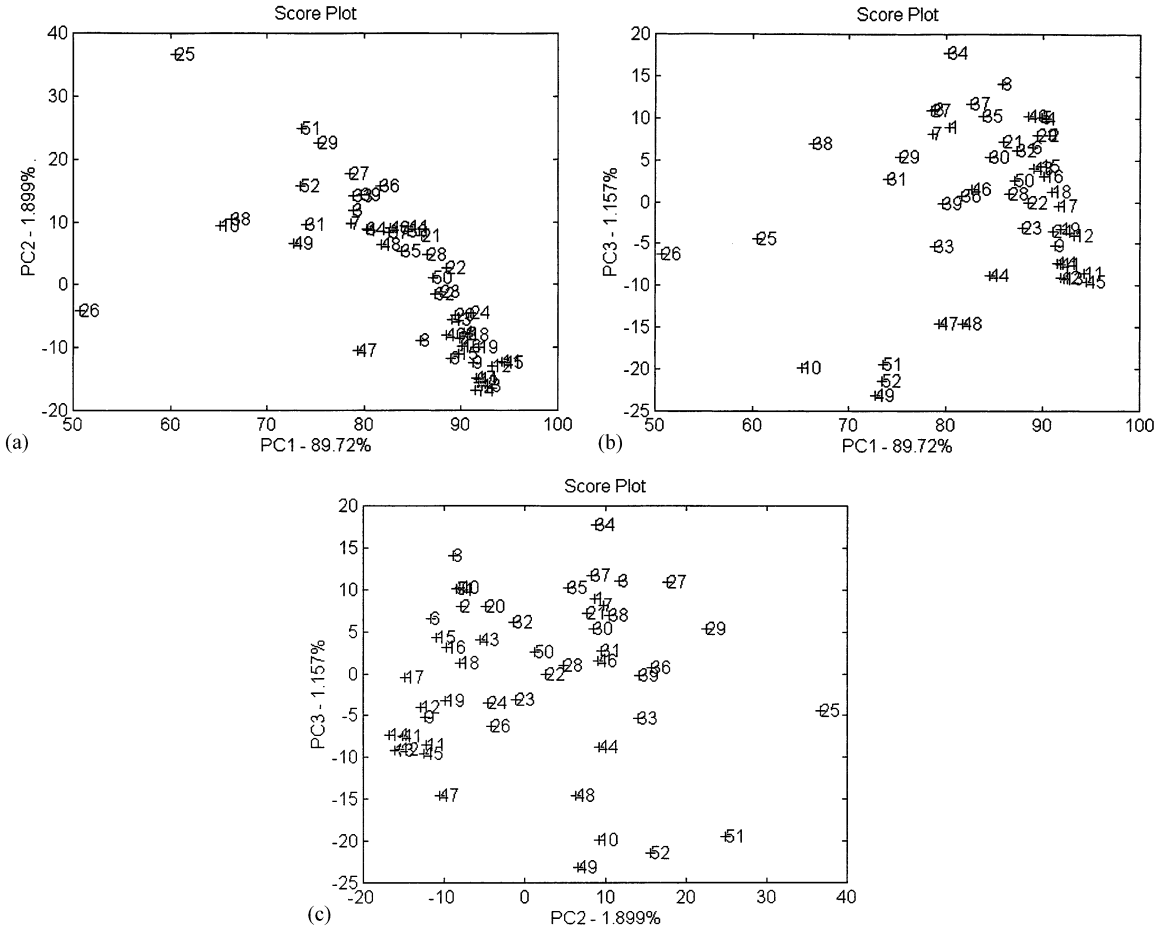

Fig. 3. (a) Score plot from the PCA of the log transformed electrospray mass spectra, with PC2 plotted against PC1. For the numbering of the compounds, see Table 1. (b) Score plot from the PCA of the log transformed electrospray mass spectra, showing PC3 against PC1. Notation as in (a). (c) Score plot from the PCA of the log transformed electrospray mass spectra, showing PC3 against PC2. Notation as in (b).

tion and/or that are especially characterized by fragment ions at higher m/q values are situated in the left part of the same figure. The second PC reflects the difference between compounds primarily characterized by very intense peaks or basepeak at low m/q values (below m/q 120), for example compound numbers 13, 14, 17, 41, 42 and compounds mainly marked by peaks of higher mass (between m/q 150 and 280), for example compound number 25 (digitoxigenin). Along PC3, compounds with basepeak and high intensity peaks at very high m/q values (above m/q 290) are separated from the rest. These are for instance compound numbers 49 (nicardipin), 51 (lormetazepam) and 52 (flurazepam). They are all marked by a high molecular mass, which in an electrospray interface automatically leads to the formation of a number of heavy fragment ions.

Electrospray mass spectra, pretreated by means of a logarithmic transformation, seem to reveal other features as being mainly important for assessing the similarity of chemical structures as raw mass spectra.

4.2. Sequential projection pursuit

The results obtained by applying the method of SPP on the raw electrospray mass spectra are

demonstrated in the score plot of PP1–PP2 (Fig. 5). In the score plot of PP1–PP2, no clusters of structurally similar compounds can be detected. However, it is not the intention to discover groups of similar compounds but to detect outlying objects in the data. Generally, three outliers can be observed, compound numbers 25 (digitoxigenin), 46 (lobelin), and 51 (lormetazepam). The former two compounds can not really be detected as outliers in the resulting PCA-plots, while the latter is clearly outlying along PC4 (Fig. 1(c)). They are all characterized by high intensity fragment peaks of high mass.

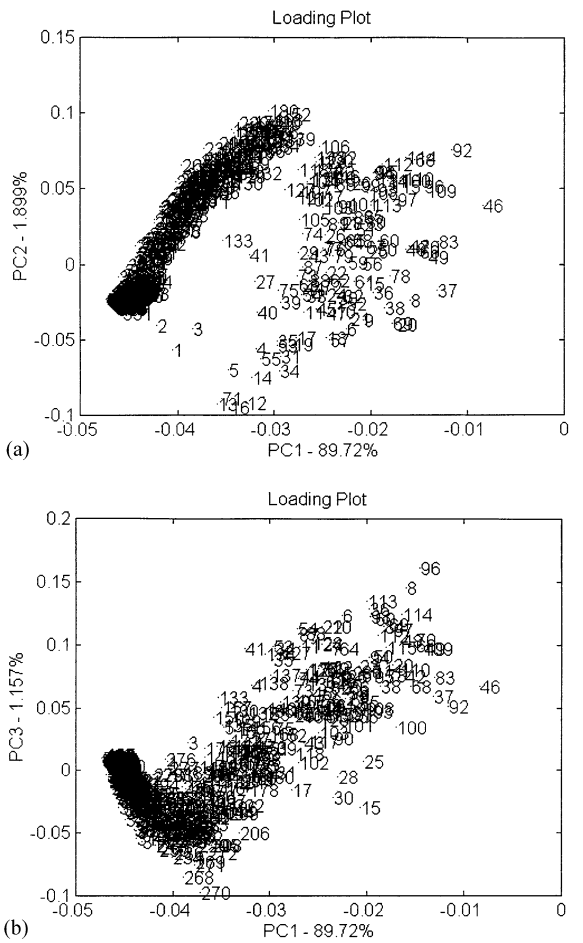

Fig. 4. (a) PCA loading plot of the log transformed electrospray mass spectra, showing the second loading vector against the first loading vector. (b) PCA loading plot of the log transformed electrospray mass spectra, with the third loading vector versus the first loading vector.

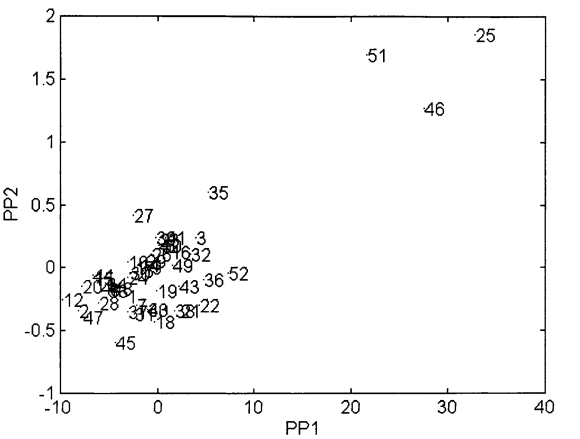

Fig. 5. Score plot from the SPP of the raw electrospray mass spectra, showing PP2 against PP1. For the numbering of the compounds, see Table 1.

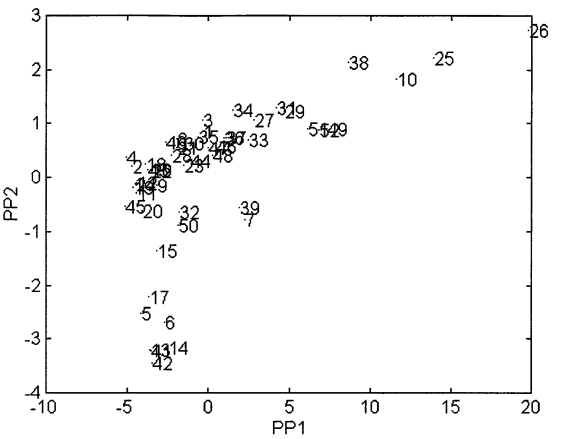

Fig. 6. Score plot from the SPP of the log transformed electrospray mass spectra, showing PP2 against PP1. For the numbering of the compounds, see Table 1.

The log transformed electrospray mass spectra were also subjected to SPP. The results are shown in Fig. 6. One outlier can be detected in the positive direction of PP1, compound number 26 (digitoxin). This object is also outlying along PC1 (Fig. 3(a)) since it is marked by many fragment ions with respect to the other compounds in the data set. A small group of extreme objects can be detected in the negative direction of PP2. It consists of compound numbers 5 (D-galactose), 6 (D-mannose), 13 (D-leucin), 14 (L-isoleucin), 17

(amphetamin), 41 (menthol) and 42 (camphor). This is much more difficult to observe along PC2 (Fig. 3(a)). In fact, these compounds are low molecular mass compounds. They are extreme because they are characterized by very intense peaks or basepeaks at low m/q values.

4.3. Qualitative comparison of hierarchical clusterings

The unweighted pair-group average method was used as the clustering method of choice to cluster the small data set of 52 synthetic substances. The correlation coefficient was used as similarity measure for mass spectral data and the Tanimoto coefficient for Daylight structural fingerprints. The resulting upgma-clusterings, based on raw and log transformed mass spectra, are shown in Figs. 7 and 8, respectively. The upgma-classification of the Daylight fingerprints is given in Fig. 9.

An inspection of the resulting classifications of the raw and log transformed mass spectra shows that in both classifications clearly separated clusters of very similar compounds are formed. For instance, all sugars are grouped together in one cluster.

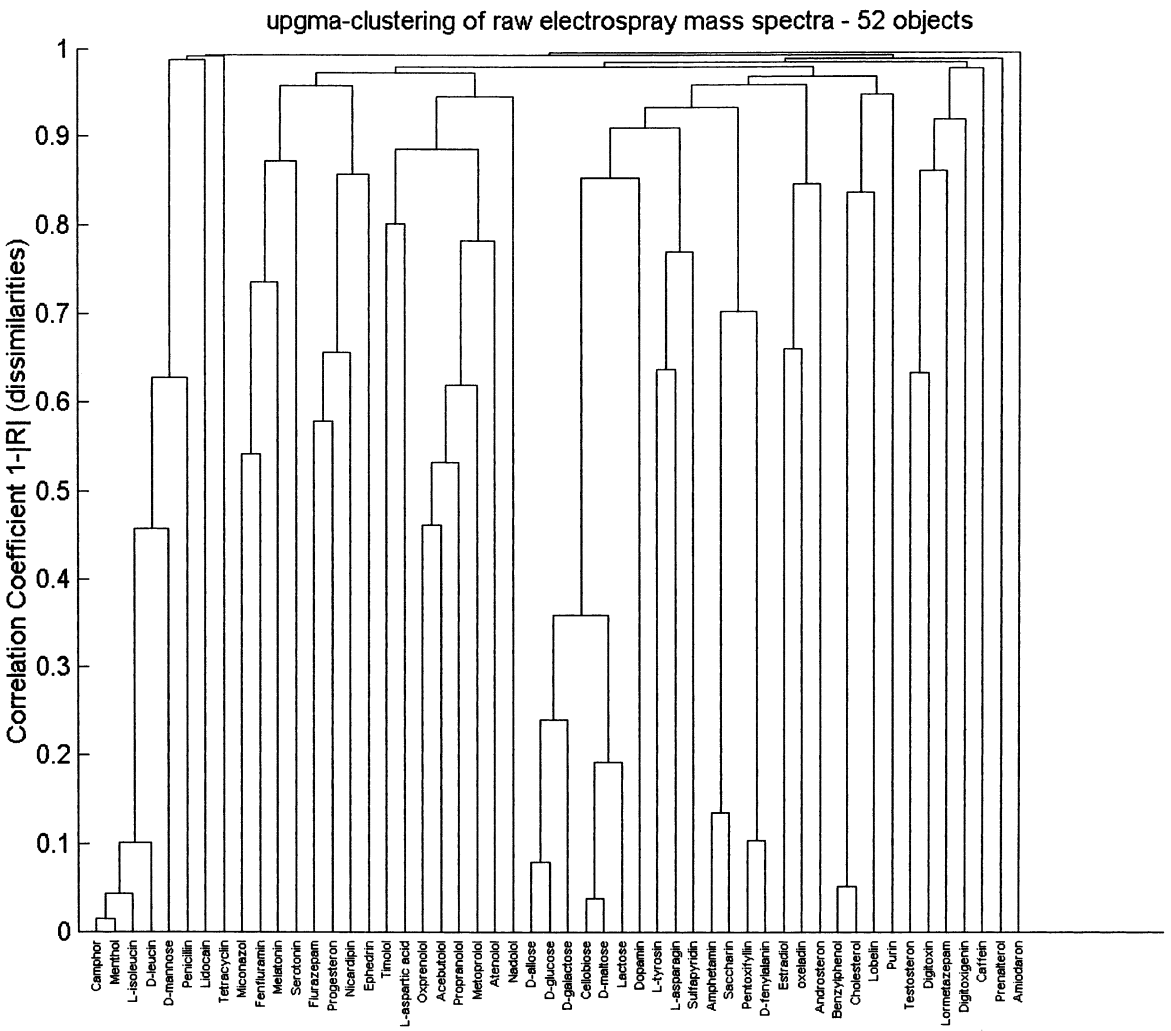

Fig. 7. Hierarchical upgma-clustering of the raw electrospray mass spectra.

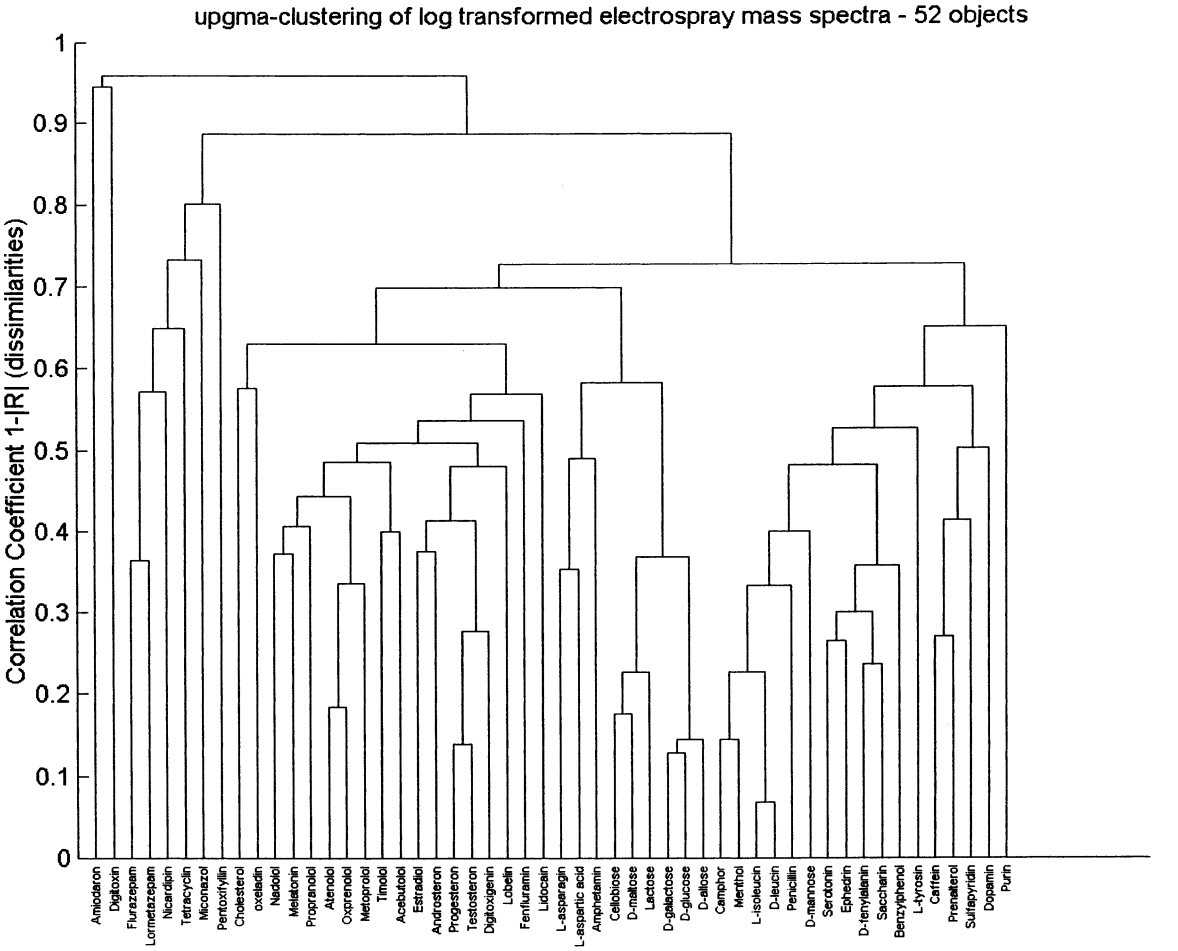

Fig. 8. Hierarchical upgma-clustering of the log transformed electrospray mass spectra.

Also, the group of β-blockers is found as such in one cluster in both classifications. Most corticoids (estradiol, androsteron, progesteron, testosteron, digitoxigenin) are located near each other in the classification of the log transformed mass spectral data, while they appear more dispersed over the tree-structure in the classification of raw spectra. In both classifications, the amino acids are scattered over the tree. Some smaller subgroups, containing similar compounds are also present. An example of this kind is camphor and menthol. Melatonin and serotonin are located near each other in the classification of raw mass spectra and lormetazepam and flurazepam are linked together in the clustering of log transformed spectra.

In the classification of the 2D structural fingerprints, two main clusters are formed. All sugars are linked together in one subcluster of the first main cluster, as well as most corticoids that are also contained in one smaller cluster. Also, most amino acids (L-asparagin, L-aspartic acid, L-isoleucin, D-leucin) are located near each other. The second main cluster is more heterogeneous but consists of various small groups of very similar compounds. For instance, the group of β-blockers is found as such in one subgroup. Another example is given by melatonin and serotonin, lormetazepam and flurazepam, tyrosin and fenylalanin.

Consequently, it appears that the classification, based on mass spectral characteristics, is qualitatively as good as the classification based on 2D structural fingerprints so that it seems that one can use such analytical properties for similarity/diversity assessments.

4.4. Quantitative comparison of hierarchical clusterings

The measure of Wallace is used to obtain a more quantitative measure of the similarity between two different clusterings of a same set of compounds. A quantitative comparison was performed between the different upgma-clusterings mutually and with an expert's classification of the same set of compounds, according to known structure and pharmacological activity. The expert's classification is shown in Fig. 10. However, due to the diversity of compounds in the data set, other classifications might be proposed by others.

The results of the quantitative comparison of the different upgma-clusterings with the expert's classification (Table 2) show that the classification based on 2D structural fingerprints compares best with the expert's classification. However, the classification based on mass spectral characteristics compares almost as well with the expert's classification as the clustering of the 2D structural fingerprints. Only a slight difference exists be-

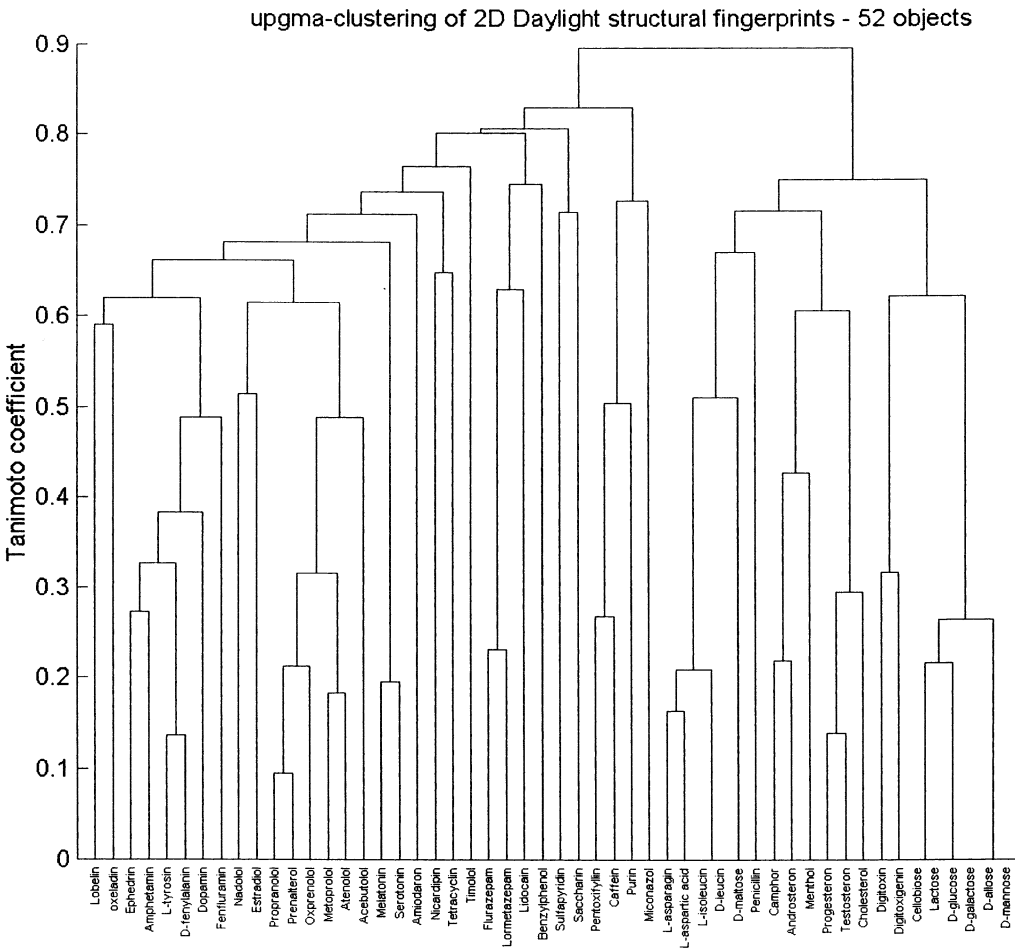

Fig. 9. Hierarchical upgma-clustering of the 2D Daylight structural fingerprints.

1:	serotonin	6.	nadolol
	melatonin		propranolol
	pentoxifyllin		timolol
	caffein		oxprenolol
	H-purin		metoprolol
			atenolol
			acebutolol
2.	Saccharin		prenalterol
	Sulfapyridin		lidocain
	Penicillin		oxeladin
			amiodaron
3.	Miconazole		lormetazepam
	Nicardipin		flurazepam
	lobelin		
4.	tetracyclin		
	estradiol		
	progesteron		
	testosteron		
	androsteron		
	cholesterol		
	digitoxigenin		
	digitoxin		
	maltose		
	glucose		
	lactose		
	allose		
	galactose		
	mannose		
	cellobiose		
	benzylphenol		
	menthol		
	camphor		
5.	leucin		
	isoleucin		
	aspartic acid		
	asparagine		
	tyrosin		
	phenylalanin		
	amphetamin		
	fenfluramin		
	dopamin		
	ephedrin		

Fig. 10. Expert's classification of the set of 52 substances.

tween the clusterings of the raw and log transformed mass spectra, with the latter producing somewhat better results.

The results for the comparison of the different upgma-clusterings between them are reported in Table 3. Comparing the two largest clusters of each classification, both clusterings based on raw and log transformed mass spectra are almost identical. When comparing the four or six largest clusters of the respective clusterings, the classification of the log transformed mass spectra seems to compare better with the classification of the 2D structural fingerprints than the clustering of raw spectral features.

In conclusion, the classification of the electrospray mass spectra seems to perform well as compared with the classification of the structural fingerprints.

5. Conclusion

This study demonstrates the use of multivariate exploratory techniques to investigate whether experimental electrospray-mass spectra can be applied for similarity/diversity assignments. For this, hierarchical upgma-clusterings of the 2D structural fingerprints or the mass spectra of 52 synthetic substances were qualitatively and quantitatively compared between them and with an expert's classification of the same set of compounds. It is found that a good classification of the compounds is established using experimental mass spectra instead of the exact structure. Moreover, a logarithmic transformation pretreatment of the spectral data gives rise to even better cluster solutions.

In conclusion, electrospray mass spectra, in spite of the limited fragmentation, are found to provide a valuable source of information on the classification of compounds and, therefore, in

Table 2
(1) Comparison with two largest clusters of the respective clusterings; (2) comparison with four largest clusters of the respective clusterings; (3) comparison with the six largest clusters of the respective clusterings

	Expert's/raw mass spectra	Expert's/log mass spectra	Expert's/daylight fingerprint
1	0.4589	0.4396	0.4938
2	0.3696	0.4172	0.5456
3	0.3599	0.4173	0.4953

Table 3
(1) Comparison of two largest clusters of the respective clusterings; (2) comparison of four largest clusters of the respective clusterings; (3) comparison of six largest clusters of the respective clusterings

	Raw mass spectra/log mass spectra	Raw mass spectra/daylight fingerprint	Log mass spectra/daylight fingerprint
1	0.9798	0.7004	0.6935
2	0.7397	0.5270	0.5506
3	0.7096	0.4472	0.5247

combination with other spectroscopic techniques (IR, UV) and chromatographic separations, they probably can be employed to establish the similarity/diversity of compounds.

References

[1] M.A. Strege, Hydrophilic interaction chromatography-electrospray mass spectrometry analysis of polar compounds for natural product drug discovery, Anal. Chem. 70 (1998) 2439–2445.
[2] D.M. Bayada, H. Hamersma, V.J. van Geerestein, Molecular diversity and representativity in chemical databases, J. Chem. Inf. Comput. Sci. 39 (1999) 1–10.
[3] R. Bursi, T. Dao, T. van Wijk, M. De Gooyer, E. Kellenbach, P. Verwer, Comparative Spectra Analysis (CoSA): spectra as three-dimensional molecular descriptors for the prediction of biological activities, J. Chem. Inf. Comput. Sci. 39 (1999) 861–867.
[4] R.D. Brown, Y.C. Martin, The information content of 2D and 3D structural descriptors relevant to ligand–receptor binding, J. Chem. Inf. Comput. Sci. 37 (1997) 1–9.
[5] H. Matter, Selecting optimally diverse compounds from structure databases: a validation study of two-dimensional and three-dimensional molecular descriptors, J. Med. Chem. 40 (1997) 1219–1229.
[6] J. Smedsgaard, J.C. Frisvad, Using direct electrospray mass spectrometry in taxonomy and secondary metabolite profiling of crude fungal extracts, J. Microbiol. Methods 25 (1996) 5–17.
[7] V. Schoonjans, F. Questier, A.P. Borosy, B. Walczak, D.L. Massart, B.D. Hudson, Use of mass spectrometry for assessing similarity/diversity of natural products with unknown chemical structures, J. Pharm. Biomed. Anal. 21 (2000) 1197–1214.
[8] R.D. Smith, J.A. Loo, C.G. Edmonds, C.J. Barinaga, H.R. Udseth, New developments in Biochemical Mass Spectrometry: electrospray ionization, Anal. Chem. 62 (1990) 882–899.
[9] D.L. Massart, B.G.M. Vandeginste, L.M.C. Buydens, S. De Jong, P.J. Lewi, J. Smeyers-Verbeke, Data Handling in Science and Technology: Handbook of Chemometrics and Qualimetrics: Part A–B, Elsevier, Amsterdam, 1997.
[10] D.L. Massart, L. Kaufman, The Interpretation of Analytical Chemical Data by the Use of Cluster Analysis, Wiley, New York, 1983.
[11] Q. Guo, W. Wu, F. Questier, D.L. Massart, C. Boucon, S. De Jong, Sequential Projection Pursuit using genetic algorithms for data mining of analytical data, Anal. Chem. 72 (2000) 2846–2855.
[12] L. Kaufman, P.J. Rousseeuw, Finding Groups in Data: an Introduction to Cluster Analysis, Wiley, New York, 1990.
[13] W. Vogt, D. Nagel, H. Sator, Cluster Analysis in Clinical Chemistry: a Model, Wiley, New York, 1987.