ELSEVIER

Journal of Pharmaceutical and Biomedical Analysis
26 (2001) 225–239

JOURNAL OF
PHARMACEUTICAL
AND BIOMEDICAL
ANALYSIS

www.elsevier.com/locate/jpba

Combining spectroscopic data (MS, IR): exploratory chemometric analysis for characterising similarity/diversity of chemical structures

V. Schoonjans, D.L. Massart *

ChemoAc, Pharmaceutical Institute, Vrije Universiteit Brussel, Laarbeeklaan 103, B-1090 Brussels, Belgium

Received 19 October 2000; received in revised form 30 January 2001; accepted 5 February 2001

Abstract

Combined infrared–mass spectra (IR–MS) have been used to examine a small data set of synthetic substances in order to elucidate whether a combination of spectral descriptors yield better classification and similarity predictions than their corresponding individual spectral descriptors. To eliminate differences in variation, a logarithmic transformation or log double-centering pretreatment was necessary. Principal component analysis (PCA) was applied to observe clusters of similar compounds. Hierarchical upgma-cluster analysis was also used for data classification.

Keywords: Mass spectra; Spectral features; Similarity; Upgma-clustering; Principal component analysis; Sequential projection pursuit

1. Introduction

In the lead generation phase of drug discovery, the concept of molecular diversity has become an increasingly important tool. However, the process of comparing substances and quantitatively assessing similarity starts with the choice of the appropriate descriptors. A wide variety of descriptors have been developed for diversity studies and among them, two-dimensional structural fingerprints are one of the most popular since they encode a great deal of information. A molecular fingerprint identifies several different patterns of the structure and describes those patterns in a bit string, based on the absence or presence of a set of two to seven atom patterns [1–4]. A great number of similarity measures based on different molecular aspects are described in the literature, and all of them try to quantify the comparison between the structure of the elements of a set of compounds [5].

Multivariate chemometric techniques (e.g. principal component analysis (PCA), cluster analysis) have been used as a method to calculate the molecular diversity of large collections of individual compounds with known structure. However, traditional sources of therapeutic agents include natural products and microbiological broths

* Corresponding author. Tel.: +32-2-4774737; fax: +32-2-4774735.
E-mail address: fabi@vub.vub.ac.be (D.L. Massart).

0731-7085/01/$ - see front matter
PII: S0731-7085(01)00427-7

wherefrom the structures are not known so that the current techniques become inapplicable and the knowledge of diversity severely reduced. Consequently, the different molecules have to be converted into other descriptors, e.g. experimental parameters. Both mass spectrometry (MS) and infrared spectroscopy (IR) are capable of providing structure-related information and multivariate methods based on PCA and cluster analysis have shown to be useful in either spectral data space to separate groups of structural similar compounds [6,7]. An interesting question that arises from the analysis based upon single spectral descriptors is whether a combination of them can lead to an improved classification and a better predictive power of chemical diversity. To answer this question, both spectral descriptors were combined into one matrix and subsequently employed in a PCA and a hierarchical cluster analysis.

2. Data

2.1. Spectral features

There is an important difference between data resulting from IR-spectroscopy and those from mass spectrometry, experimental mass spectra are obtained as a list of peak positions at nominal mass units and corresponding intensities. On the other hand, IR data are supplied as a series of observed intensities at regular intervals, i.e. IR-peaks are not always situated at the same wavenumber. However, use can be made of spectral features, calculated for predefined wavenumber intervals for IR peak position, to solve this problem of chemical shifts. Spectral features are vectors that characterise as much as possible the spectrum of a compound. Feature $\mathrm{INT}(\upsilon_1, \upsilon_2)$ is the intensity of a spectral absorption and is calculated for each interval by using the following formula

$$\mathrm{INT}(\upsilon_1, \upsilon_2) = \begin{cases} A_{\max} \\ 0 \end{cases} \quad (1)$$

with $A_{\max}$ being the maximum absorption in this predetermined interval [8].

2.2. Spectra

The small data set, used in this study, contains 61 synthetic substances with known chemical structure, listed in Table 1. They were already used in an analogue study about assessing similarity/diversity by mass spectrometry [6] and Fourier transform-infrared (FT-IR) spectroscopy [7]. The data set was chosen in this way that it consists of a relatively high number of structurally and pharmacologically similar compounds, e.g. the β-blockers, some smaller groups of similar substances (steroids, amino-acids) and some compounds that were randomly selected.

All the FT-IR-spectra were recorded with the use of a Perkin–Elmer FT-IR Spectrum 100 Spectrometer. In order to obtain very good spectra, 16 scans for each sample were recorded in the range 4000–400 cm^{-1} with a 4 cm^{-1} resolution at a sampling interval of 1 cm^{-1}. Before starting a measurement a background spectrum was recorded. The full-curve IR-spectra were converted to ASCII format and truncated at 3700 cm^{-1}. In this range, from 3700 to 400 cm^{-1}, all the compounds present absorption peaks. The absorbance values were normalised to the range 0–1. As described by Robb and Munk [9], the studied range was divided into 245 intervals with the widths continuously increasing with growing wavenumber and features were calculated by applying Eq. (1) to each interval. A data matrix (61×245) was created where the rows correspond to the 61 compounds and the columns to the 245 wavenumber intervals. The values in the matrix are the INT spectral features at each predefined interval.

Electron impact (EI) mass spectra for the same set of substances were obtained from the NIST/EPA/NIH MS Database for PC and were represented as peak tables containing positions (m/q-ratio) and intensities of all fragment ions. A data matrix (61×390) whose rows are the 61 substances and whose columns are the 390 m/q-values was created. The values in the matrix are the fragment ion intensities normalised for total mass equal to 100.

Both matrixes were then combined into one big matrix (61×635) where the first 245 columns are

Table 1
List of synthetic substances

1	Maltose
2	Glucose
3	Saccharin
4	Penicillin
5	Tetracyclin
6	L-aspartic acid
7	L-asparagin
8	D-leucin
9	L-isoleucin
10	DL-phenylalanin
11	L-tyrosin
12	Amphetamin
13	Ephedrin
14	Dopamin
15	Serotonin
16	Histamin
17	Melatonin
18	Mexiletin
19	Fenfluramin
20	Oxeladin
21	Procain
22	Lidocain
23	Digitoxigenin
24	Digitoxin
25	Testosteron
26	Androsteron
27	Progesteron
28	Estradiol
29	Cholesterol
30	Terbutalin
31	Acebutolol
32	Pindolol
33	Oxprenolol
34	Sotalol
35	Propranolol
36	Nadolol
37	Atenolol
38	Alprenolol
39	Metoprolol
40	Betaxolol
41	Prenalterol
42	4-Benzylphenol
43	Menthol
44	Camphor
45	Guanidin
46	Caffein
47	Pentoxifyllin
48	H-purin
49	Lysergide
50	Strychnin
51	Codein
52	Heroin
53	Morphin
54	Cocain
55	Nicotin
56	Lobelin
57	Amiodaron
58	Miconazole
59	Nicardipine
60	Sulfapyridin
61	Lormetazepam

the IR spectral descriptors and the following 390 columns the MS descriptors.

Two-dimensional structural fingerprints for the same substances were obtained using the Daylight clustering software.

3. Exploratory analysis of the data

3.1. Importance of transformations before exploratory analysis

If common scaling methods are used, the clustering results are either dominated by the MS descriptors or the IR spectral features. However, an improvement is realised by applying a log transform to the combined spectral data matrix. This data transformation is often applied in multivariate data analysis to eliminate the differences in variation between the variables. Pre-processing by log double-centering consists of first taking the logarithm and then centering the data both by rows and by colums. This transform eliminates the size effect when it is present so that only the contrasts (ratios) between the variables will be expressed [10].

3.2. Principal component analysis

PCA is often applied to data sets as a tool for data set size reduction and to uncover the structure of the data [11]. The combined spectral data matrix, after a logarithmic transformation and a log double-centering pretreatment, was analysed.

3.3. Sequential projection pursuit

Sequential projection pursuit (SPP) is a method that can be used to find inhomogeneities (outliers) in high dimensional data more easily than PCA [12]. The method was applied on both log transformed and log double-centered data.

3.4. Cluster analysis

Hierarchical clustering is the process of subdividing a group of compounds into clusters of compounds that exhibit a high degree of both

intracluster similarity and intercluster dissimilarity [15]. The output can be represented as a dendrogram [13]. The unweighted pair-group average linkage (upgma) method with the correlation coefficient as similarity measure was used to cluster the small data set of 61 synthetic substances, based on their combined IR–MS. This clustering method is probably one of the most commonly used hierarchical clustering method and the similarity between the clusters is calculated as the average of the similarities between their objects [14].

3.5. Comparison of clusterings

There are a number of ways to measure quantitatively the similarity between two different clusterings of the same finite set of objects. The one used in this study is the measure of Wallace, s_w (1983). A description of the methodology as applied to mass spectrometry is given in [6].

4. Results and discussion

4.1. Principal component analysis of combined spectral data

To obtain an overview of the dominating patterns and major trends, a PCA was first performed on the log transformed spectral descriptor matrix. The first four principal components explain 89.1% of the total variance, 84.8% is explained by PC1, 2.2% by PC2, 1.3% by PC3 and 0.8% by PC4. Fig. 1a shows the score plot of PC1 against PC2. The third and fourth components are visualised in the same way in Fig. 1b and c, respectively. The corresponding loadings are plotted in Fig. 2a–d.

In the score plot of Fig. 1a, all steroids are clustered in the upper left part, most amino-acids are positioned in the lower part of the plot and the group of β-blockers appears in the central region of the same plot.

The score plots and loading plots (Figs. 1 and 2) show that the first PC is probably an indicator of general size, in this case a variable describing how high the total MS fragmentation and IR-absorption of the compounds is. However, IR spectral features have higher loadings and therefore seem to be more important for the first dimension than MS descriptors. This can be seen in the loading plot of Fig. 2a. Correspondingly, compounds with low scores in Fig. 1a are primarily characterised by their mass spectrum, while compounds with high scores are mainly marked by their IR absorption. PC2 reflects the difference between compounds primarily characterised by extensive MS fragmentation, for example compound numbers 23–29 and substances mainly characterised by a strong IR-absorption in the C–H stretching region (2800–2300 cm^{-1}), such as, for instance compound number 6. The MS fragmentation pattern of the compounds mainly determines the third PC. Objects with negative scores show peaks at m/q 30, m/q 42, m/q 64 and m/q 79. This can also be seen in the loading plot of Fig. 2c. Fragment ions with m/q 30 ($^+NH_2{=}CH_2$) and m/q 42 ($CH_2{=}C{=}O^+$) arise from cleavage in the N-alkyl chains of aliphatic amines and amides. For this reason, chemical structures that contain aliphatic nitrogen appear in the lowest region of Fig. 1b. Peaks at m/q 64 and m/q 79 can be associated with, respectively ($C_5H_4^+$) and ($C_6H_7^+$) and originate from the breakdown of aromatic ethers and aromatic amides. The fourth PC only explains about 0.8% of the total variance. The interpretation of this component is difficult since many variables seem to be important. This is seen in the loading plot of Fig. 2d. Correspondingly, PC4 primarily discriminates substances with characteristic peaks at m/q 44 (variable 274), m/q 57 (variable 287), m/q 71 (variable 301), m/q 72 (variable 302) and m/q 86 (variable 316) and/or very intense peaks of high mass from the rest. These fragment ions arise from the breakdown of aliphatic amines. The peaks with m/q 72 ($C_4H_{10}N^+$) and m/q 86 ($C_5H_{12}N^+$) arise from α-cleavage next to the N-atom, with the loss of the largest alkyl fragment. Fragment ions of m/q 86 then break up to form ions of m/q 71 ($C_4H_9N^+$) and m/q 57 ($^+C_4H_9$). The resulting ion of m/q 72 breaks down further, giving rise to $(CH_3CH=NH_2)^+$, m/q 44. However, fragment ions of m/q 44 can also be attributed to $(O=C=NH_2)^+$, $(CH_2=CHOH)^+$.

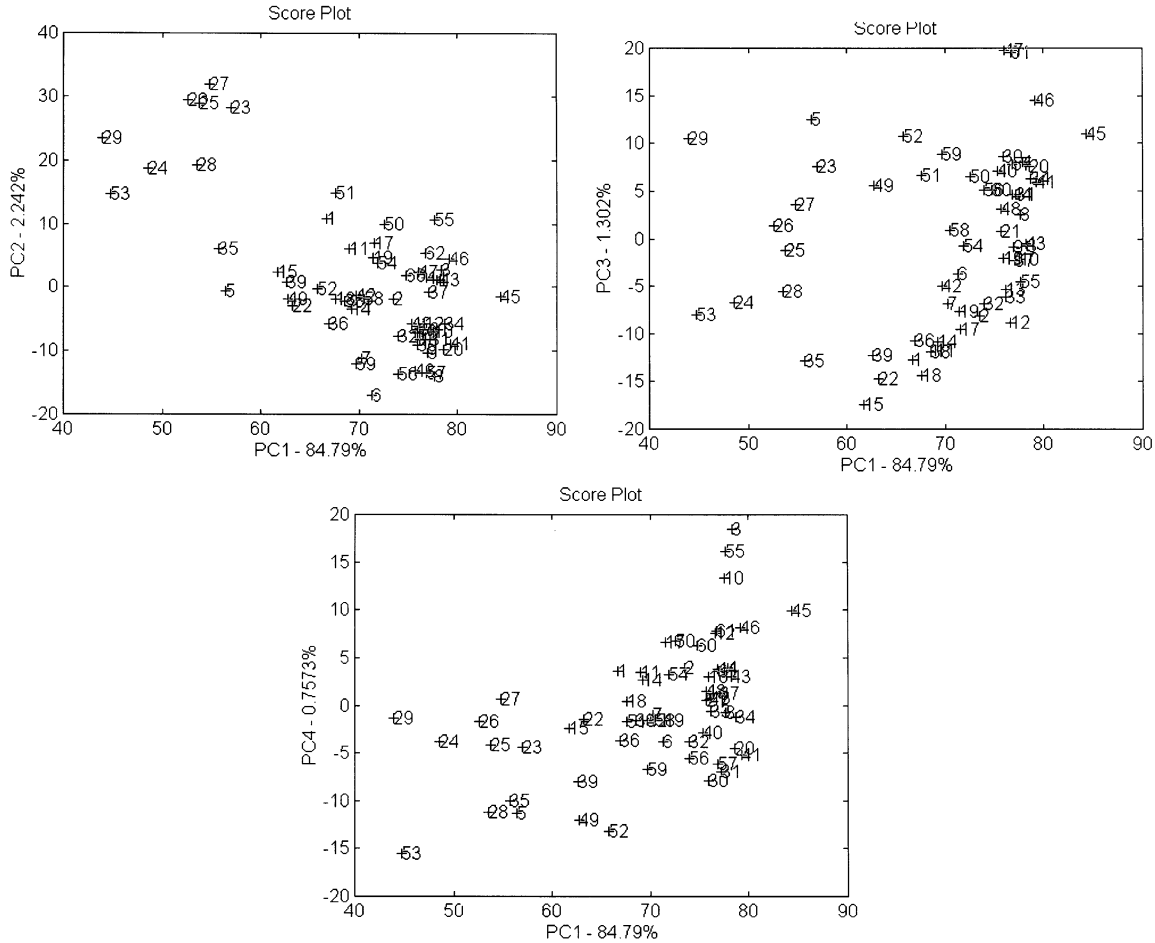

Fig. 1. (a) Score plot from the PCA of the log transformed combined spectra, showing PC2 against PC1. For the numbering of the compounds, see Table 1. (b) Score plot from the PCA of the log transformed combined spectra, showing PC3 against PC1. Notation as in a. (c) Score plot from the PCA of the log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

Furthermore, infrared absorption in the O-H and C-H stretching region is also somewhat determining the fourth PC.

The results demonstrate that a number of classes of compounds are formed on the basis of similarity in structure. The IR spectral properties as well as the MS features of the considered compounds play a role in the resulting PCA-plots. In conclusion, combined IR–MS can be used in applications such as similarity searching and clustering since they are capable of providing a lot of information about the structure of compounds.

The combined spectral data matrix, after a log double-centering pretreatment, was also subjected to PCA, which resulted in four principal components that describe 33.4% of the total variance. The first PC describes 16.7% of the variance and the second, third and fourth PC 8.0, 4.7 and 4.0%, respectively. Fig. 3a–c show the scores of PC2 against PC1, PC3 against PC1 and PC4 against PC1, respectively, and Fig. 4a–d the respective loadings.

The set of compounds is split up in three groups along PC1: most amino-acids are situated in the left part of Fig. 3a, the steroids are grouped in the extreme right part and the β-blockers appear in the central region of the plot.

An inspection of the scores (Fig. 3) and loadings (Fig. 4) shows that the first PC is no longer an indicator of general size. After a log double-centering of the combined spectral data matrix, the first PC describes the same information as the second PC for log transformed combined spectral data. The same holds for the second and third PC, that explain the same features as described by PC3 and PC4, respectively, for the log transformed combined spectral data. PC4 describes the contrast between compounds with fragment ion peaks at m/q 57 (variable 287), m/q 72 (variable 302) and m/q 86 (variable 316), which are characteristic of aliphatic amines and a strong IR absorption in the O-H and C-H stretching region from compounds with nitrogen heterocycles in their chemical structures that give peaks at m/q 92 ($C_7H_8^+$), m/q 78 ($C_6H_6^+$) and m/q 65 ($C_5H_5^+$).

After performing a log double-centering pretreatment, the size effect of PC1 disappears. Furthermore, the results show the same characteristic features for assessing the similarity of the compounds as log transformed combined spectra. However, since some valuable information is contained in the first PC, a logarithmic transformation is preferred to a log double-centering.

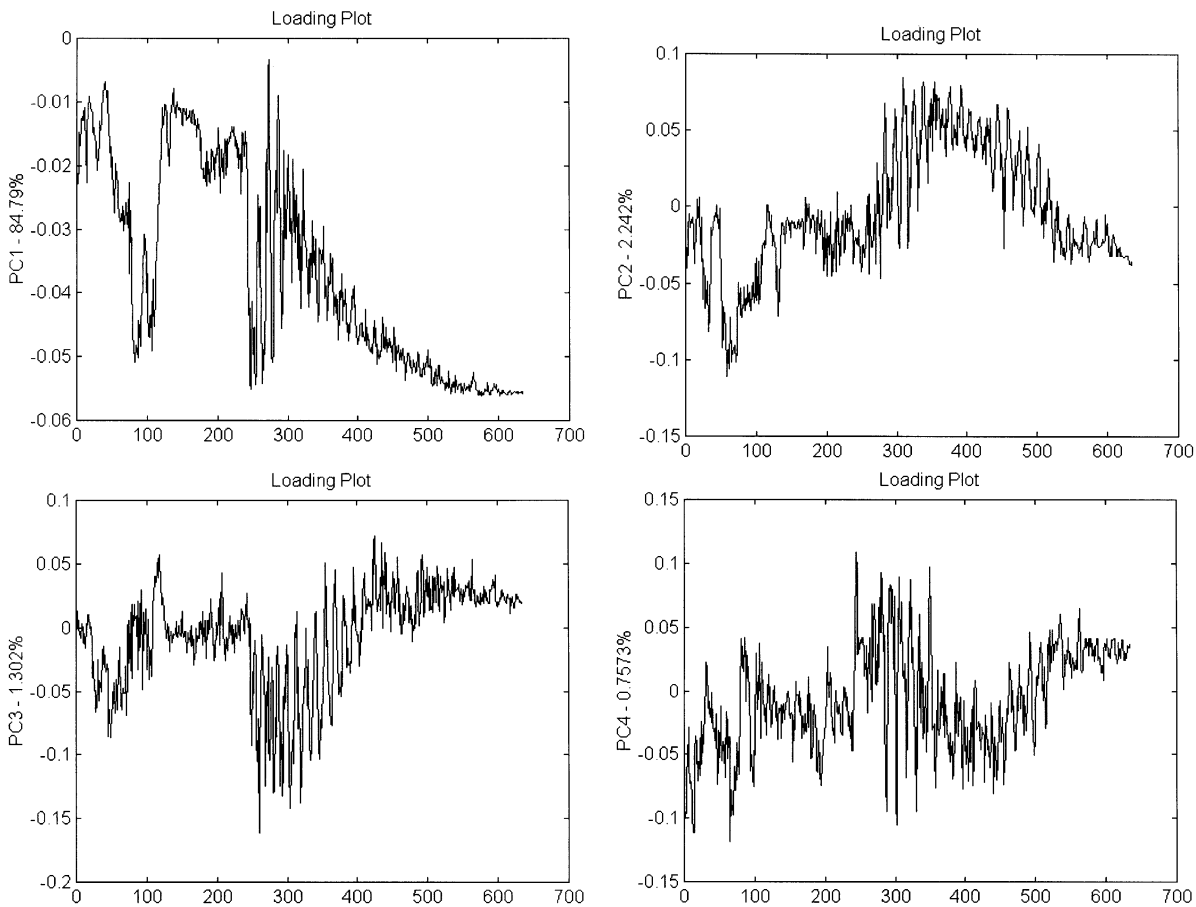

Fig. 2. (a) Loading plot from the PCA of the log transformed combined spectra. The first loading vector is plotted vs. IR and MS descriptors. (b) Loading plot from the PCA of the log transformed combined spectra, with the second loading vector plotted against IR and MS descriptors. (c) Loading plot from the PCA of the log transformed combined spectra, with the third loading vector plotted against IR and MS descriptors. (d) Loading plot from the PCA of the log transformed combined spectra, with the fourth loading vector plotted against IR and MS descriptors.

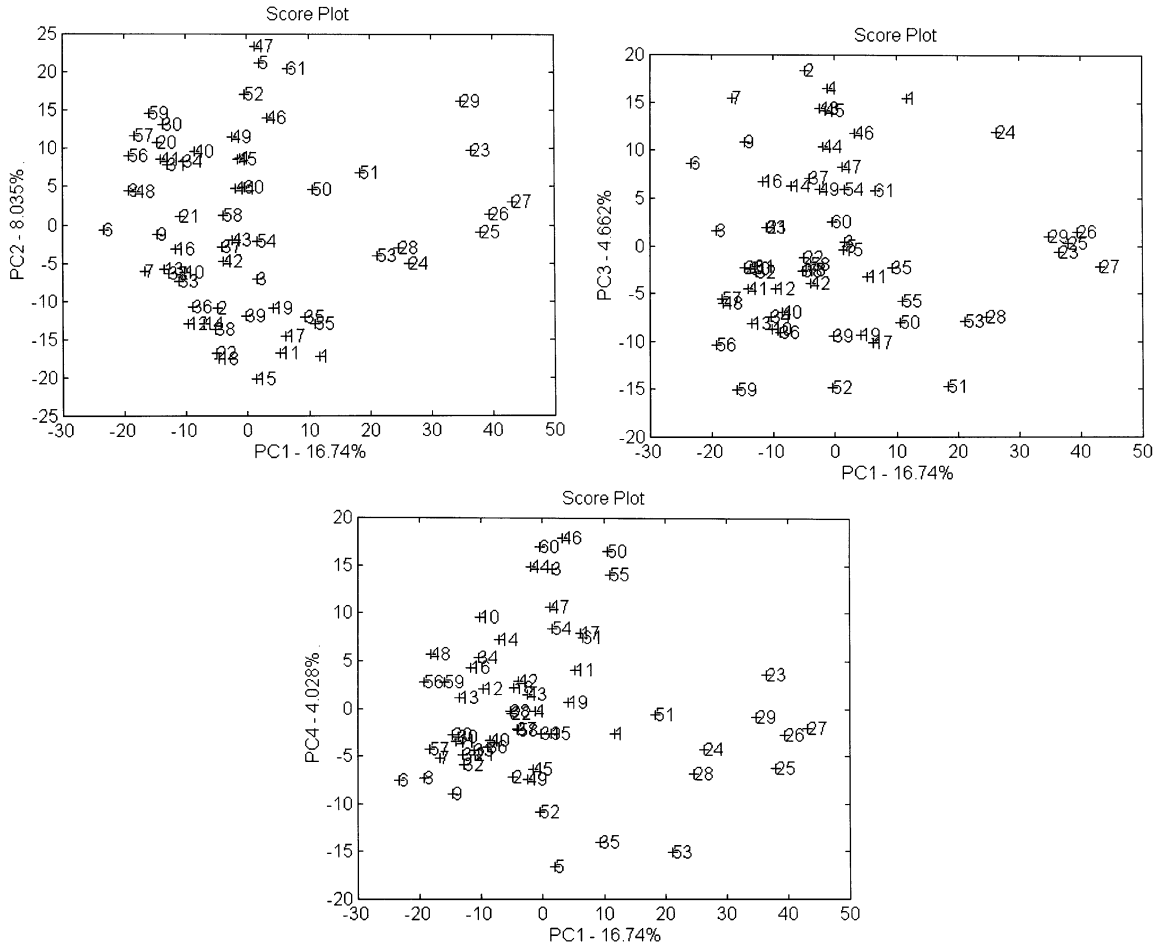

Fig. 3. (a) Score plot from the PCA of the combined spectra after a log double-centering, with PC2 plotted against PC1. For the numbering of the compounds, see Table 1. (b) Score plot from the PCA of the combined spectra after a log-double-centering, showing PC3 against PC1. Notation as in a. (c) Score plot from the PCA of the combined spectra after a log double-centering, showing PC4 against PC1. Notation as in b.

4.2. Sequential projection pursuit of combined spectral data

The results obtained by SPP on the log transformed combined spectral descriptors are shown in the score plot of PP1–PP2 (Fig. 5). SPP clearly shows a separation of groups of compounds. The group of steroids is separated from the amino-acids and the β-blockers in PP1. Along PP2, the group of β-blockers is found as such in the lower part of the plot, while both the group of amino-acids and the group of steroids have fallen apart. One can also observe a layered structure of two elongated clusters and the upper cluster contains substances that absorb little IR radiation, due to many features set to zero. This problem can be solved by using transmissions instead of absorptions. Therefore, it seems that SPP can be used to detect artefacts in the data that might influence the results in PCA. In general, SPP is applied to detect inhomogeneities in the data instead of finding groups of similar compounds. One can find two outliers on PP2, compound numbers 25 (testosteron), 27 (progesteron). With PCA, these outlying objects in the data are much more difficult to distinguish on PC2 (Fig. 1a). Both

substances are characterised by extensive MS fragmentation compared with the other compounds in the data set.

The combined spectral data matrix, after a log double-centering was also subjected to SPP. In the score plot of PP1–PP2 (Fig. 6), one can detect two clusters along PP2. As with log transformed data, most compounds that show little IR absorption are located in the upper right of the plot. In the lower right part, the group of β-blockers[3] can be found entirely but, compared with PCA, one can detect no clustering due to chemical similarity. One outlier can be identified in the negative direction of PP1, compound number 5 (tetracyclin). This object also has a high negative score on PC4 (Fig. 3c) and is characterised by very intense peaks at low m/q-values.

4.3. Hierarchical cluster analysis of combined spectral data

4.3.1. Qualitative comparison of upgma-clusterings

The technique used for performing a hierarchical cluster analysis is based on the unweighted pair-group average (upgma) method, with the correlation coefficient as similarity measure for spectral data and the Tanimoto coefficient for 2-D

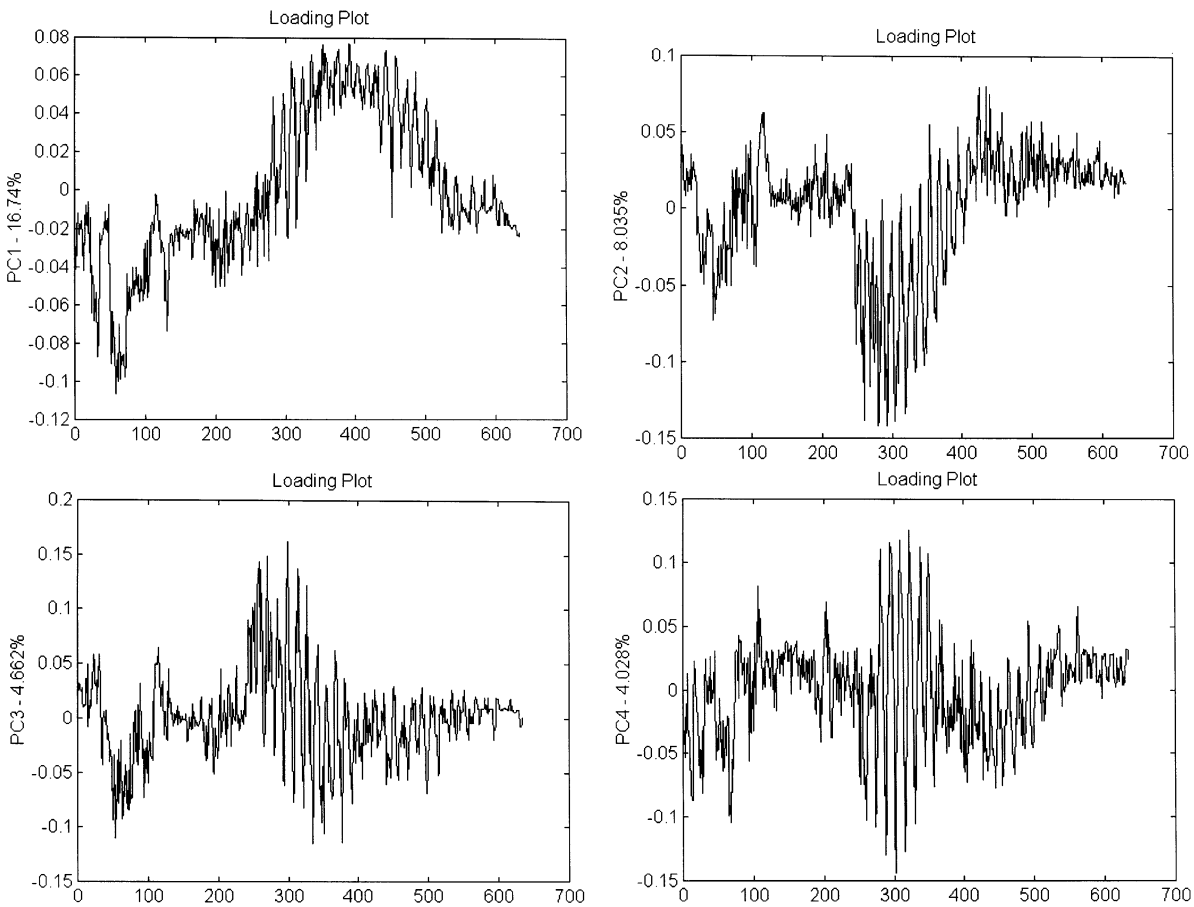

Fig. 4. (a) PCA loading plot of the combined spectra after a log double-centering, showing the first loading vector against spectral descriptors. (b) PCA loading plot of the combined spectra after a log double-centering, with the second loading vector plotted vs. spectral descriptors. (c) PCA loading plot of the combined spectra after a log double-centering, with the third loading vector plotted against spectral descriptors. (d) PCA loading plot of the combined spectra after a log double-centering, with the fourth loading vector plotted against spectral descriptors.

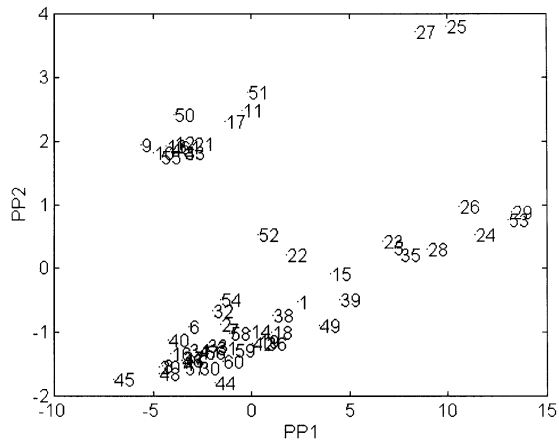

Fig. 5. Score plot from the SPP of the log transformed combined spectra, showing PP2 against PP1. For the numbering of the compounds, see Table 1.

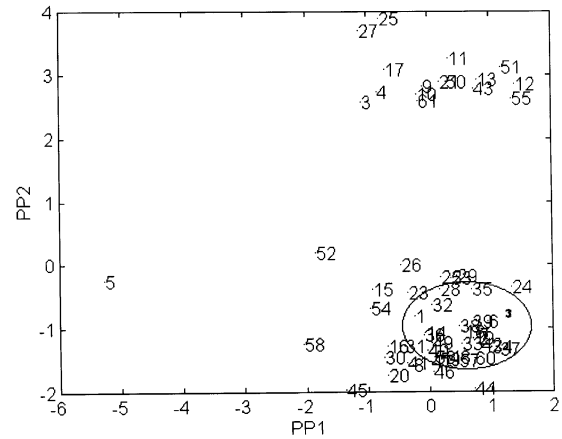

Fig. 6. Score plot from the SPP of the combined spectra after a log double-centering, showing PP2 against PP1. For the numbering of the compounds, see Table 1.

structural fingerprints. The resulting hierarchical upgma-clustering, based on Daylight structural fingerprints is shown in Fig. 7 and the clusterings, based on combined spectral data, after a logarithmic transformation and a log double-centering in Figs. 8 and 9, respectively.

In both classifications, based on combined spectral data (Figs. 8 and 9), the group of steroids can be found entirely in one cluster. Also, most amino-acids (L-isoleucin, D-leucin, L-asparagin, L-aspartic acid) are grouped together in one smaller cluster, as well as both sugars, maltose and glucose that are located near each other in the respective clusterings. In the classification of log transformed combined spectra, heroin and codein are linked together. The β-blockers appear more dispersed over the tree-structure in both clusterings.

The classification of the Daylight structural fingerprints (Fig. 7) clearly shows clusters of similar compounds, for example, the group of β-blockers is found entirely in one cluster. Also, most steroids are contained in one cluster, as well as most amino-acids. Another example is given by the alkaloids, codein, morphin and heroin, as well as melatonin and serotonin, camphor and men-

Table 2
(1) Comparison with four largest clusters of the respective clusterings; (2) comparison with six largest clusters of the respective clusterings, based on combined spectra and with five largest clusters of the clustering, based Daylight fingerprints

Expert's/log combined spectra	Expert's/log dbl centered combined spectra	Expert's/Daylight fingerprints	
1	0.3908	0.4101	0.4197
2	0.3733	0.3402	0.4390

Table 3
(1) Comparison of four largest clusters of the respective clusterings; (2) comparison of six largest clusters of the clusterings, based on combined spectra between them and with five largest clusters of the clustering, based on Daylight structural fingerprints

Log spectra/log dbl centered spectra	Log spectra/Daylight fingerprint	Log dbl centered data/Daylight fingerprint	
1	0.5798	0.6862	0.6174
3	0.4102	0.5548	0.4163

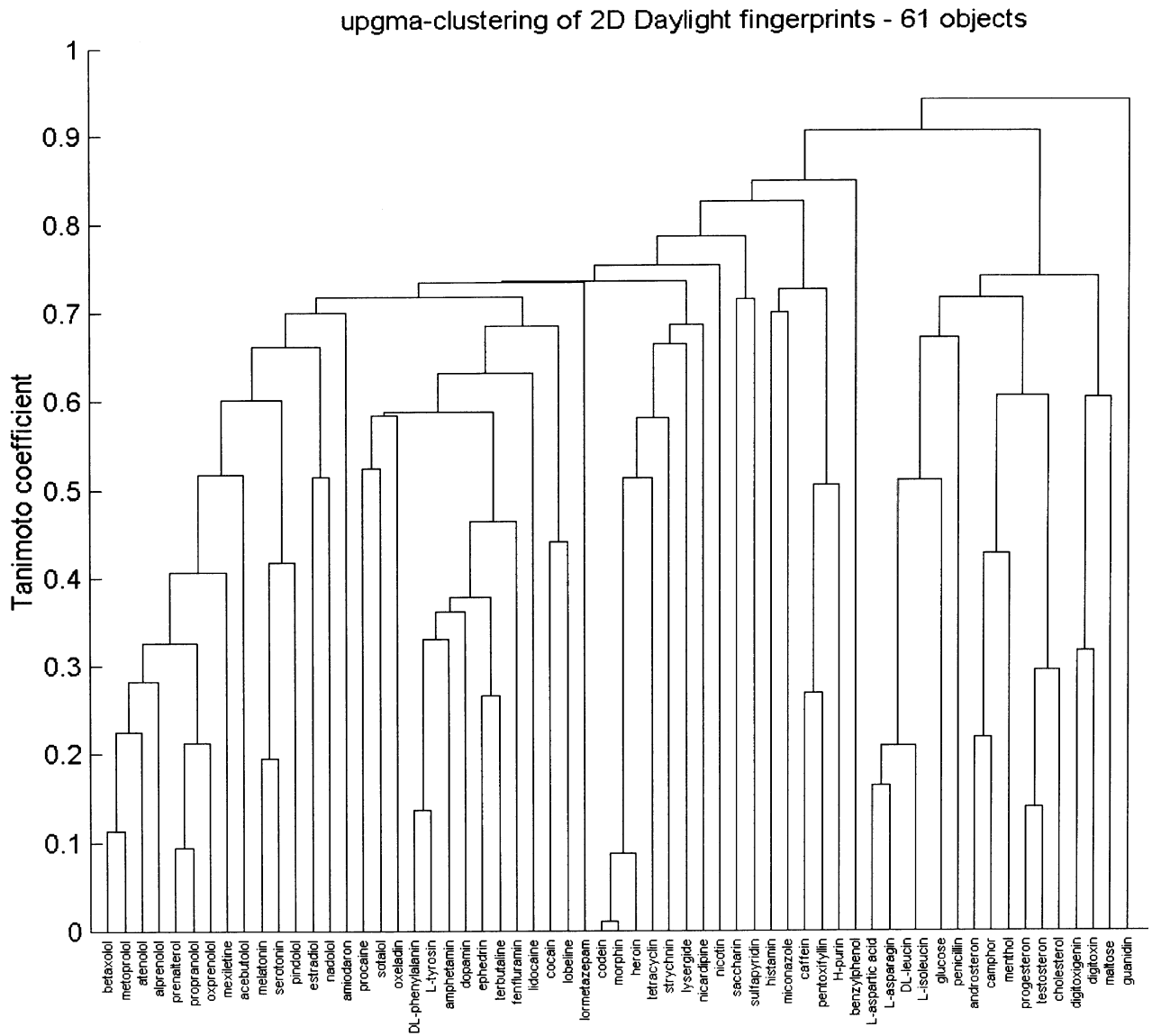

Fig. 7. Hierarchical upgma-clustering of the 2-D Daylight structural fingerprints.

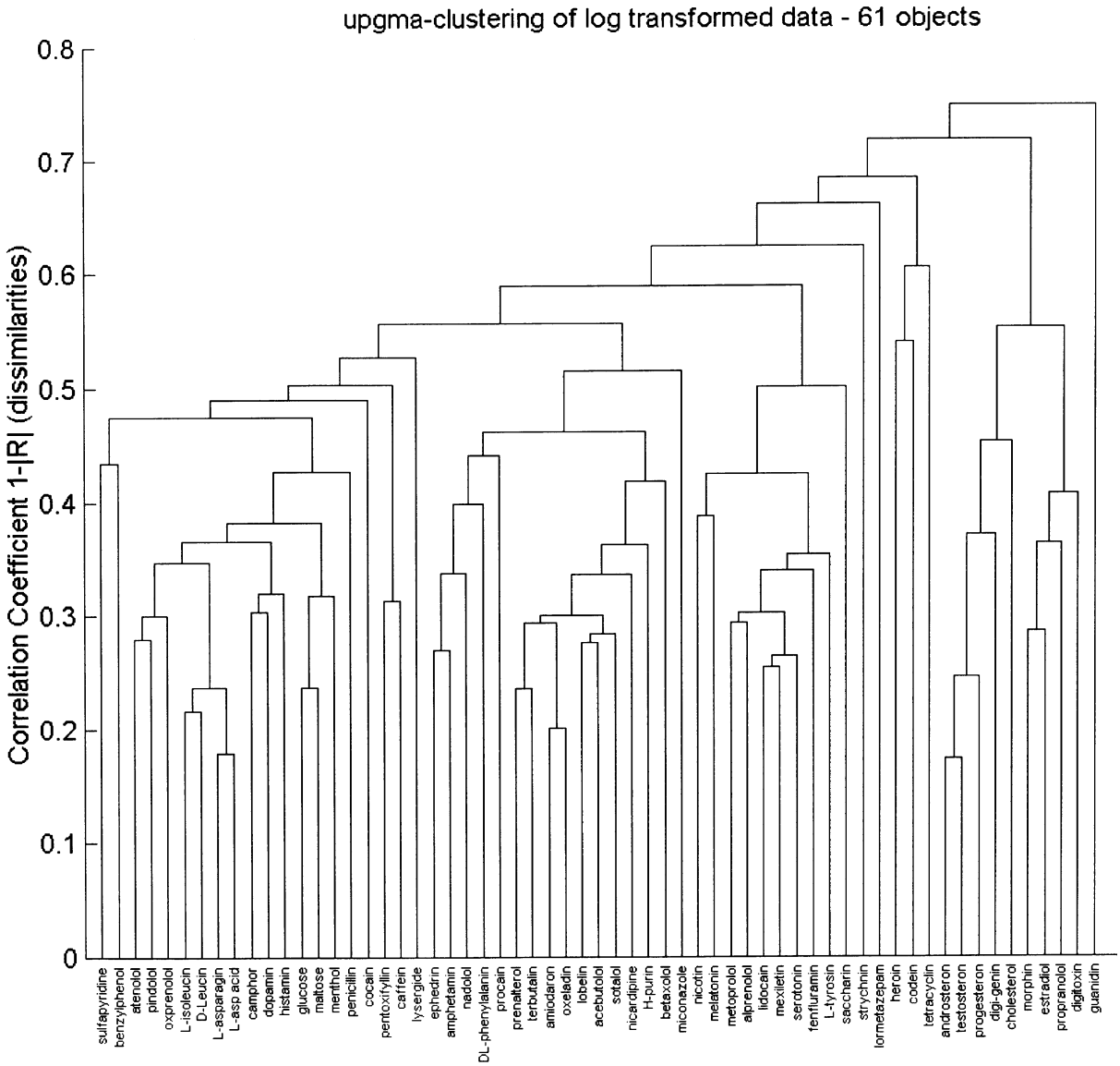

Fig. 8. Hierarchical upgma-clustering of the log transformed combined spectra.

thol and the purine derivatives, caffein, pentoxifyllin and purin that are linked together in the tree-structure. At first sight, the clustering with combined mass-infrared spectra looks to give clusters of similar compounds, so that it seems that combined spectroscopies indeed can provide characteristic information about the structure of a compound for assessing similarity/diveristy.

4.3.2. Quantitative comparison of upgma-clusterings

In addition to a visual qualitative comparison, we also used a quantitative measure of similarity, i.e. the measure of Wallace, for comparing two clusterings of the same set of compounds. A comparative study was carried out between the upgma-clusterings, based on Daylight structural

fingerprints or combined IR–MS. Furthermore, the different clustering results were compared with an expert's classification of the same set of compounds, based on our own evaluation according to known structure and pharmacological activity. However, other classifications might be proposed by others due to the wide variety of compounds in the data set. The expert's classification is shown in Fig. 10.

In Table 2, the results from the quantitative comparison of the different upgma-clusterings with the expert's classification are presented. All three upgma-classifications compare almost as well with the expert's classification. However, the clustering of the Daylight structural fingerprints compares most with the expert's classification, while no real distinction can be made between the clustering of the log transformed

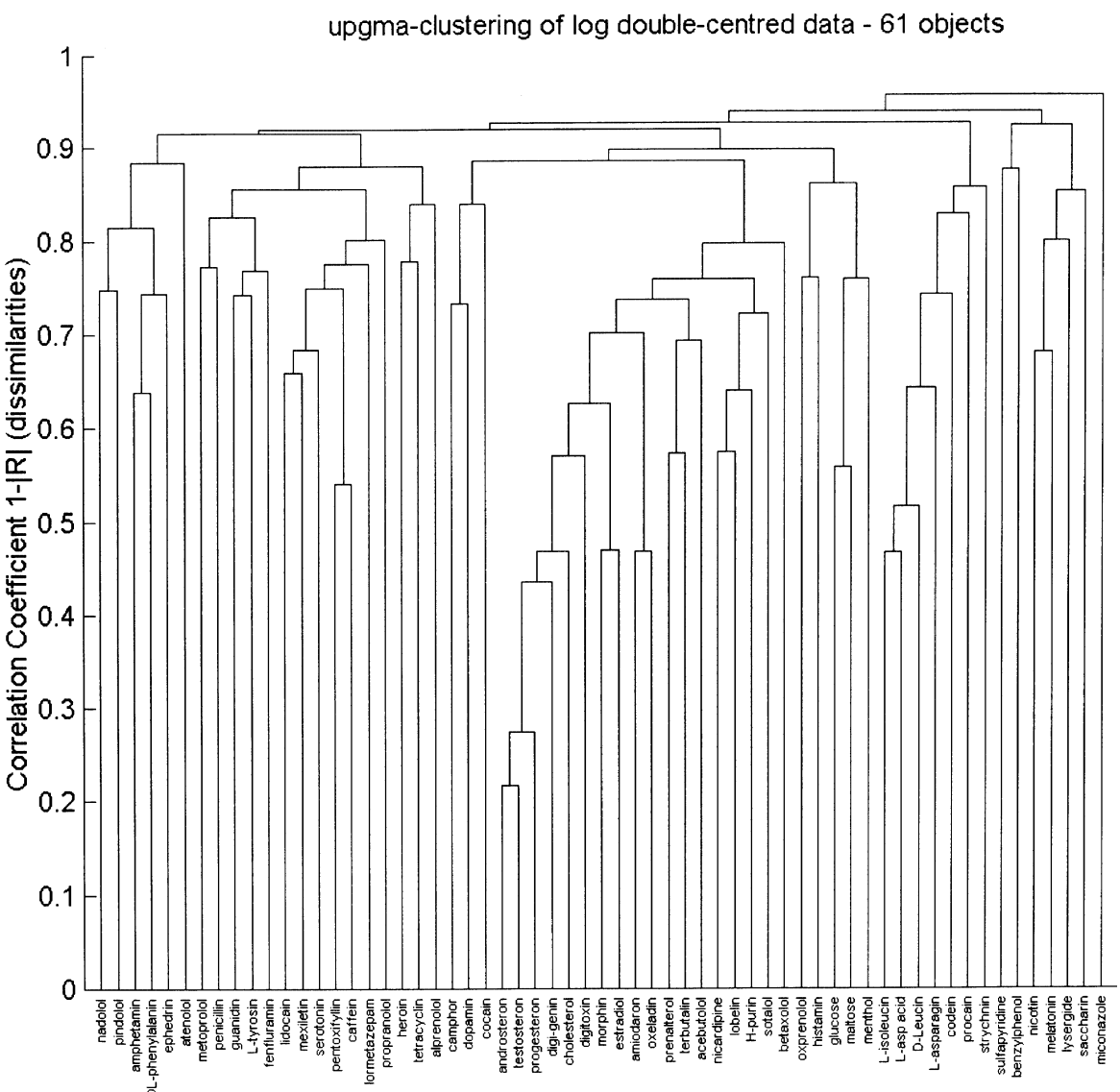

Fig. 9. Hierarchical upgma-clustering of the combined spectra after a log double-centering.

Table 4
Comparison of the different hierarchical classifications with the expert's classification; comparison with four largest clusters of the respective clusterings

Expert's/MS	Expert's/norm MS	Expert's/raw IR features	Expert's/log IR features	Expert's/log combined spectra	Expert's/log dbl centred combined spectra
0.3638	0.4125	0.4458	0.4343	0.3908	0.4101

Table 5
Comparison of the different hierarchical classifications, based on spectral characteristics, with the hierarchical classification based on 2-D structural Daylight fingerprints[a]

MS/fingerprints	Norm MS/fingerprints	Raw IR features/fingerprints	Log IR features/fingerprints	Log combined spectra/fingerprints	Log dbl centred combined spectra/fingerprints	
1	0.5040	0.4966				
2			0.5699	0.5946	0.6862	0.6174

[a] (1) Comparison of three largest clusters of the respective clusterings; (2) comparison of four largest clusters of the respective clusterings.

1:	guanidin	
	histamin	
	serotonin	
	melatonin	
	pentoxifyllin	
	caffein	
	nicotin	
	H-purin	
2:	saccharin	
	sulfapyridin	
	penicillin	
3:	miconazole	
	nicardipin	
	lobelin	
	cocain	
4:	codein	digitoxin
	morphin	benzylphenol
	heroin	maltose
	lysergide	glucose
	strychnin	menthol
	tetracyclin	camphor
	estradiol	
	progesteron	
	testosteron	
	androsteron	
	cholesterol	
	digitoxigenin	
6:	terbutalin	
	nadolol	
	propranolol	
	pindolol	
	oxprenolol	
	alprenolol	
	metoprolol	
	betaxolol	
	atenolol	
	acebutolol	
	sotalol	
	prenalterol	
	lidocain	
	mexiletin	
	oxeladin	
	procain	
	amiodaron	
	lormetazepam	
5:	leucin	
	isoleucin	
	aspartic acid	
	asparagine	
	tyrosin	
	phenylalanin	
	amphetamin	
	fenfluramin	
	dopamin	
	ephedrin	

Fig. 10. Expert's classification of the set of 61 substances.

and the log double-centered combined spectral data.

The numerical results for the comparison of the hierarchical upgma-clusterings of combined spectra and Daylight fingerprints between them are reported in Table 3. Comparing the four largest clusters of each classification, both clusterings, based on log transformed and log double-centered combined spectra are quite similar to the classification, based on Daylight fingerprints. However, the clustering of the log transformed combined spectra compares somewhat more with the clustering of the Daylight fingerprints than the clustering of the log double-centered combined spectra. The same conclusion can be made when comparing the six largest clusters of the clusterings, based on combined spectra with the five largest clusters of the clustering, based on Daylight structural fingerprints.

These results show that a classification, based on combined IR–MS characteristics is equally good as that obtained with 2-D Daylight structural fingerprints. Therefore, it seems that one can use combined spectral features instead of the structure of the compounds for characterising their similarity/diversity. However, a logarithmic transformation gives rise to somewhat better cluster solutions as compared with a log double-centering pretreatment.

The results for combined spectral data were subsequently compared with those obtained for MS [6] and IR spectra [7]. They are tabulated in Tables 4 and 5. It is found that the resulting upgma-classification of combined MS–IR spectral properties is about as similar to the expert's clas-sification as the Ward's classification of the MS data, but it appears less similar as compared with the upgma-clustering of the IR spectral features. Furthermore, no major distinction can be observed between both clusterings based on a single spectroscopic technique and the clustering of the combined spectral data when comparing them with the hierarchical classification of the Daylight structural fingerprints. Therefore, it can be concluded that a combination of both spectroscopic methods is not able to provide more information about the structure of compounds than one single spectroscopic technique. Moreover, it seems, at least for the small data set of 61 chemical compounds, that IR spectroscopy can predict the similarity/diversity of chemical compounds somewhat better than mass spectrometry.

5. Conclusion

This paper approaches exploratory chemometric techniques to elucidate whether combined analytical spectroscopies, i.e. mass spectrometry and IR spectroscopy, enable us to deduce the similarity/diversity of compounds better than one single spectroscopic technique. For this, hierarchical upgma-clusterings based on combined IR–MS data or Daylight structural fingerprints were compared mutually and with an expert's classification of the same set of substances.

This application of clustering techniques has demonstrated that combined IR–MS are capable of providing a lot of structural relevant information since the resulting classification compares well with the classification of 2-D structural fingerprints and with the expert's classification. However, a logarithmic transformation or a log double-centering

pretreatment is necessary to obtain good clustering results, determined by both spectroscopic techniques.

In conclusion, it seems that a combination of two complementary spectra for the same compound, i.e. MS and IR spectroscopy, is not necessarily more powerful for similarity/diversity assignments than just one spectroscopic technique.

Acknowledgements

We are grateful to Professor P. Van Boxlaer (Faculty of Pharmacy, University of Gent) for placing the NIST/EPA/NIH Mass Spectral Database for PC (USA) at our disposal. Also, we thank F. Parmentier, Head of the Department of Bromatology, Instituut voor Hygiene en Epidemiologie, for placing their FT-IR spectrometer at our disposal and C. Jacques for collaboration in the IR-measurements.

References

[1] M. Hassan, J.P. Bielawski, J.C. Hempel, M. Waldman, Optimization and visualization of molecular diversity of combinatorial libraries, Mol. Divers. 2 (1996) 64–74.
[2] Ajay, W. Patrick Walters, M.A. Murcko, Can we learn to distinguish between 'drug-like' and 'nondrug-like' molecules, J. Med. Chem. 41 (1998) 3314–3324.
[3] D.H. Drewry, S. Stanley Young, Tutorial: approaches to the design of combinatorial libraries, Chemomet. Intelligent Lab. Syst. 48 (1999) 1–20.
[4] 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.
[5] D. Robert, R. Carbó-Dorca, A formal comparison between molecular quantum similarity measures and indices, J. Chem. Inf. Comput. Sci. 38 (1998) 469–475.
[6] 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.
[7] V. Schoonjans, F. Questier, Q. Guo, Y. Van der Heyden, D.L. Massart, Assessing molecular similarity/diversity of chemical structures by FT-IR spectroscopy, J. Pharm. Biomed. Anal., in press.
[8] P.N. Penchev, G.N. Andreev, K. Varmuza, Automatic classification of infrared spectra using a set of improved expert-based features, Anal. Chim. Acta 388 (1999) 145–159.
[9] E.W. Robb, M.E. Munk, A neural network approach to infrared spectrum interpretation, Mikrochim. Acta 1 (1990) 131–155.
[10] 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.
[11] D.S. Frankel, Pattern recognition of Fourier transform infrared spectra of organic compounds, Anal. Chem. 56 (1984) 1011–1014.
[12] 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, J. Anal. Chem. 72 (2000) 2846–2855.
[13] P.M. Dean (Ed.), Molecular Similarity in Drug Design, Blackie Academic, London, 1995.
[14] D.L. Massart, L. Kaufman, The Interpretation of Analytical Chemical Data by the Use of Cluster Analysis, Wiley, New York, 1983.
[15] K. Baumann, J.T. Clerc, Computer-assisted IR spectra prediction-linked similarity searches for structures and spectra, Anal. Chim. Acta 348 (1997) 327–343.