ELSEVIER

Journal of Pharmaceutical and Biomedical Analysis
24 (2001) 613–627

JOURNAL OF
PHARMACEUTICAL
AND BIOMEDICAL
ANALYSIS

www.elsevier.com/locate/jpba

# Assessing molecular similarity/diversity of chemical structures by FT-IR spectroscopy

V. Schoonjans *, F. Questier, Q. Guo, Y. Van der Heyden, D.L. Massart

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

Received 19 June 2000; received in revised form 29 August 2000; accepted 7 September 2000

**Abstract**

FT-IR spectra have been investigated for their ability to distinguish compounds which are chemically diverse and to produce clusters of compounds which makes sense chemically. Principal component analysis (PCA) was applied to the analysis of a small database of FT-IR spectra. The effect of the data pretreatment step of log transformation on spectral data pattern was also visualized by using PCA plots. The method of sequential projection pursuit (SPP) was applied to detect inhomogeneities in the data. Finally, cluster analysis of these spectra, depending on unweighted pair-group average linkage, was carried out. © 2001 Elsevier Science B.V. All rights reserved.

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

## 1. Introduction

The field of chemical diversity has become fashionable in drug discovery research with the development of high-throughput screening and combinatorial chemistry. A major step in the lead generation phase is the ability to quantify the chemical similarity between compounds. Although there is no general agreement on how to quantify chemical diversity, much of the early work on similarity searching was concerned with making a number of attempts to asses it [1–4]. Numerous distances and coefficients have been proposed for diversity studies. Among them, the Tanimoto coefficient has demonstrated its value in measuring intermolecular similarity for binary fingerprints and the Euclidean distance for molecular vectors. For the assessment of spectral resemblance, correlation coefficients are often used, because they provide simple and obvious solutions [5,24]. However, caution must be exercised in predicting the degree of structural similarity between different substances. The smaller the distance measure between the objects, the more valid the prediction of 'similarity' [6]. The choice of the clustering method when classifying molecular components, which are characterized by a given set of fragment-based fingerprints or structural physicochemical properties, has been shown to be crucial for the effectiveness of separation of similar from diverse compounds [2]. Various methods,

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


PII: S0731-7085(00)00437-4

that can be used to cluster a data set, have been described in the literature but for several real data sets, studies have indicated that hierarchical agglomerative procedures perform best at separating active from inactive molecules [8–10,21].

Multivariate exploratory methods have been successfully applied in pharmaceutical drug discovery research for diversity studies on large numbers of compounds with known chemical structure. However, if samples are complex mixtures containing different substances with unknown chemical structures, as in the majority of natural product collections, these techniques are inapplicable and so the knowledge of the diversity of these samples is severely reduced. Consequently, each molecule must be represented by other descriptors, e.g. experimental parameters which must be easy to measure due to the large number of compounds. Of the various forms of spectroscopy from which the organic chemist derives structural information, mid infrared spectroscopy presents the greatest challenge for identification and structure evaluation of compounds and is thus a likely candidate [11]. Many chemometrical approaches (e.g. PCA) have been shown to be useful in mass spectrometry to reveal spectra-structure relationships and to discriminate between classes of chemical compounds with different substructures [28]. Mid infrared spectroscopy also produces multivariate data. Therefore, the use of multivariate exploratory methods should, as in mass spectrometry, show promise for the classification and interpretation of infrared spectral data, following the assumptions that structurally similar compounds should have similar spectra [7,12,13].

It is the aim of this preliminary study to apply chemometric techniques as principal component analysis (PCA) and cluster analysis to a small data set of 61 synthetic substances in order to elucidate whether FT-IR spectroscopic data could be used for characterizing similarity/diversity of chemical compounds.

Table 1
List of synthetic substances

| | |
|---|---|
| 1 maltose | 31 acebutolol |
| 2 glucose | 32 pindolol |
| 3 saccharin | 33 oxprenolol |
| 4 penicillin | 34 sotalol |
| 5 tetracyclin | 35 propranolol |
| 6 L-aspartic acid | 36 nadolol |
| 7 L- asparagin | 37 atenolol |
| 8 D-leucin | 38 alprenolol |
| 9 L-isoleucin | 39 metoprolol |
| 10 DL-phenylalanin | 40 betaxolol |
| 11 L-tyrosin | 41 prenalterol |
| 12 amphetamin | 42 4-benzylphenol |
| 13 ephedrin | 43 menthol |
| 14 dopamin | 44 camphor |
| 15 serotonin | 45 guanidin |
| 16 histamin | 46 caffeine |
| 17 melatonin | 47 pentoxifyllin |
| 18 mexiletin | 48 H-purin |
| 19 fenfluramin | 49 lysergide |
| 20 oxeladin | 50 strychnin |
| 21 procain | 51 codein |
| 22 lidocain | 52 heroin |
| 23 digitoxigenin | 53 morphin |
| 24 digitoxin | 54 cocaine |
| 25 testosteron | 55 nicotine |
| 26 androsteron | 56 lobelin |
| 27 progesteron | 57 amiodaron |
| 28 estradiol | 58 miconazole |
| 29 cholesterol | 59 nicardipine |
| 30 terbutalin | 60 sulfapyridin |
| | 61 lormetazepam |

## 2. Theory

### 2.1. Daylight structural fingerprints

Daylight hashed fingerprints are one of the most commonly used two-dimensional molecular descriptors. They encode molecules in terms of chemical substructures and therefore consist of a large amount of relevant structural information. The molecule is broken down into short, connected paths of atoms and those chemical patterns are described with bits strung together in a binary bit string. Each bit that is set to one (1) describes the occurrence of a certain fragment, whereas a zero indicates its absence [25–27].

### 2.2. Spectral features

An important problem with multivariate data resulting from infrared spectroscopy is that IR

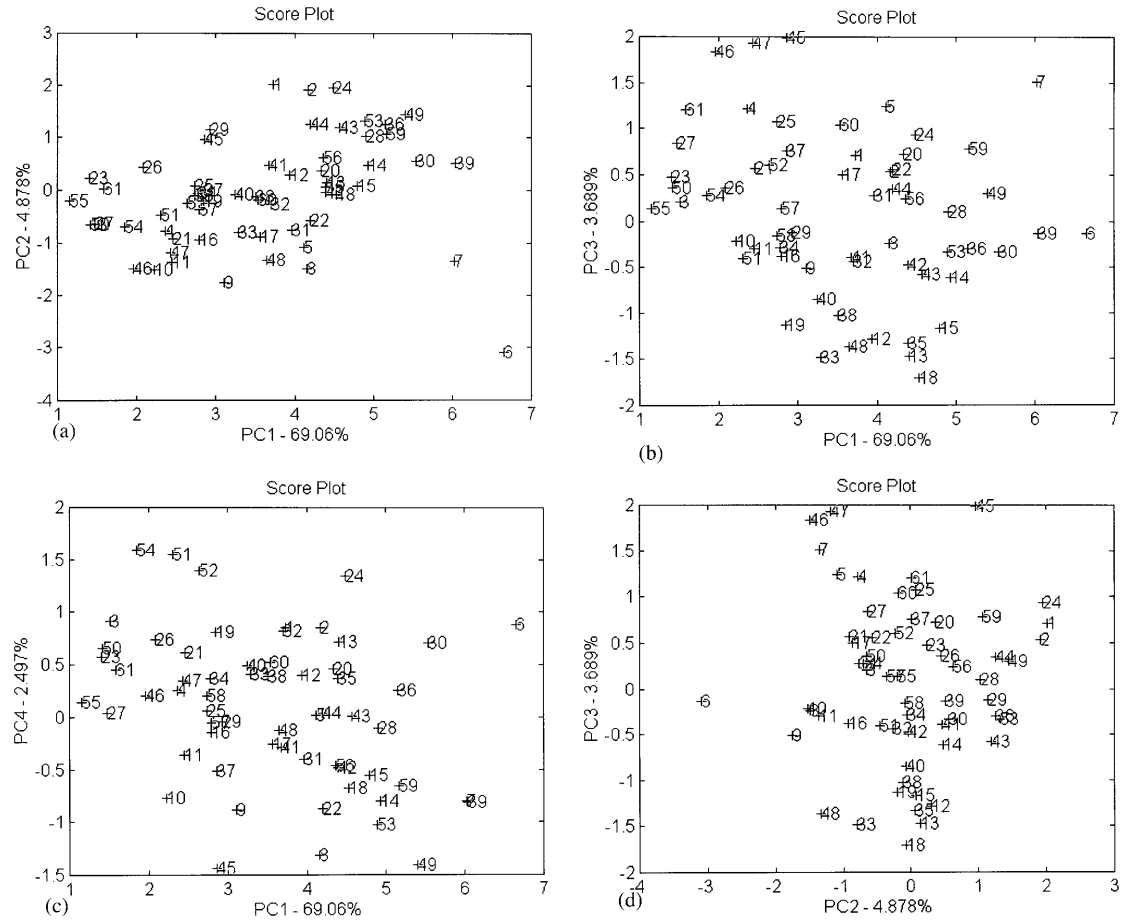


Fig. 1. (a) Score plot from the principal component analysis (PCA) of the raw infrared spectral features, showing PC2 against PC1. For the numbering of the compounds, see Table 1. (b) Score plot from the PCA of the raw infrared spectral features, showing PC3 against PC1. Notation as in (a). (c) Score plot from the PCA of the raw infrared spectral features, showing PC4 against PC1. The numbering of the compounds is the same as in (b). (d) Score plot from the PCA of the raw infrared spectral features, showing PC3 against PC2. The numbering of the compounds is the same as in (c).

peaks caused by the same substructure are not always situated at the same single real number on the wavenumber scale. This fuzziness of IR peak position must be taken into account when applying exploratory methods. One possibility is to specify wavenumber intervals for IR band positions and to calculate spectral features for each predefined interval [13–15]. The selection of appropriate wavenumber intervals is crucial in feature generation.

Spectral features are vectors that represent as much as possible the essential information contained in the infrared spectrum of a compound. Feature $INT(\upsilon_1, \upsilon_2)$ is the intensity of a spectral absorption. If the difference between the maximum and minimum absorption in an interval is $< 0.005$, i.e. if there is no increasing infrared absorption, then the input unit corresponding to that interval is given a zero value. If there is an increasing absorption in a frequency interval, then an input value between 0 and 1 in proportion to the strength of the increase was given according to:

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

with $A_{\max}$being the maximum absorption in this predetermined interval [16].

### 2.3. Data

The study was carried out by using a small data set of 61 synthetic substances. The structure of all substances is known. They are listed in Table 1 and were used in an analogue study about assessing similarity/diversity by mass spectrometry [28]. The choice of the small data set is justified by their differences in structure and pharmacological activity: a relatively large class of highly similar compounds, e.g. the β-blockers, some smaller more vague groups of structurally similar substances and furthermore, some compounds picked at random.

Fourier transform infrared spectra (FT-IR) were recorded with the use of a Perkin Elmer FT-IR Spectrum 1000 spectrometer operating at 4 $cm^{-1}$ resolution and measured from 4000 to 400 $cm^{-1}$ at a sampling interval of 1 $cm^{-1}$. Before starting a measurement a background spectrum

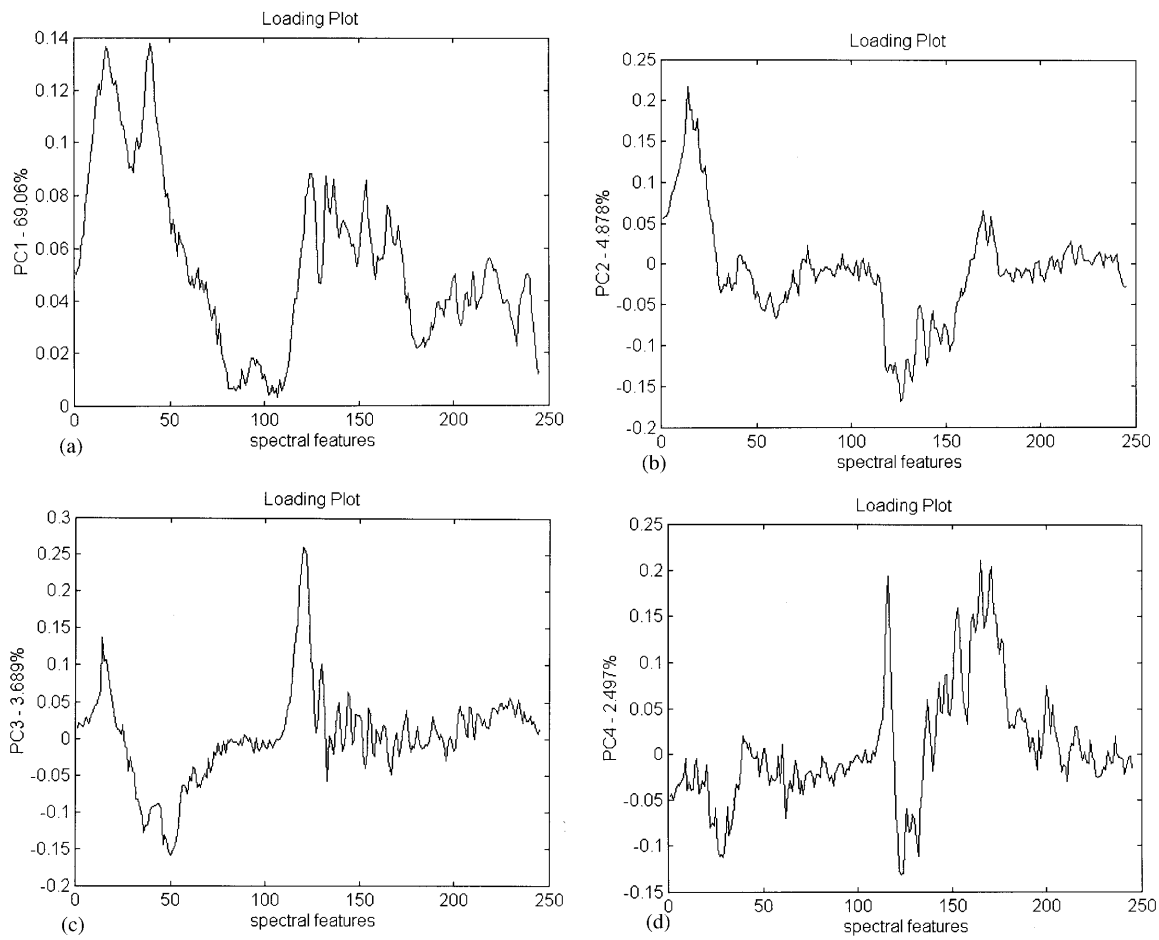


Fig. 2. (a) Loading plot from the principal component analysis (PCA) of the raw infrared spectral features. The first loading vector is plotted vs. infrared spectral features. (b) Loading plot from the PCA of the raw infrared spectral features, with the second loading vector plotted against infrared spectral features. (c) Loading plot from the PCA of the raw infrared spectral features, with the third loading vector plotted against infrared spectral features. (d) Loading plot from the PCA of the raw infrared spectral features, with the fourth loading vector plotted against infrared spectral features.

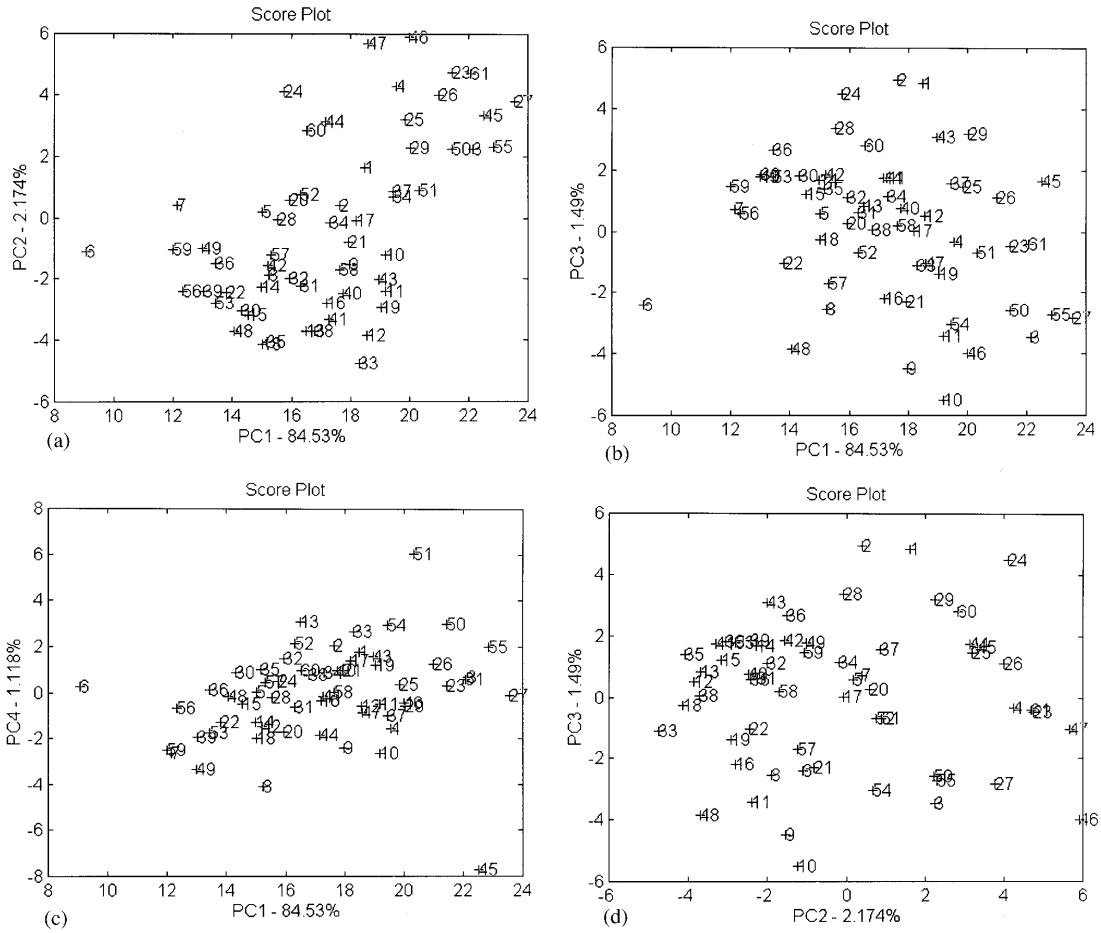


Fig. 3. (a) Score plot from the principal component analysis (PCA) of the log transformed infrared spectral features, with PC2 plotted against PC1. For the numbering of the compounds, see Table 1. (b) Score plot from the PCA of the log transformed infrared spectral features, showing PC3 against PC1. Notation as in (a). (c) Score plot from the PCA of the log transformed infrared spectral features, showing PC4 against PC1. Notation as in (b). (d) Score plot from the PCA of the log transformed infrared spectral features, showing PC3 against PC2. Notation as in (c).

was recorded and each spectrum was automatically averaged over 16 scans. Solids were examined as dispersions in compressed KBr discs, liquids as films on NaCl plates. The full-curve IR-spectra were then converted from hexadecimal to ASCII format and were subsequently truncated at 3700 $cm^{-1}$. The absorbance values were normalized to the range 0–1. For the generation of spectral features, the latter spectra, starting at 3700 $cm^{-1}$ and ending at about 400 $cm^{-1}$, were divided into 245 intervals with the widths continuously increasing with growing wavenumber as described by Robb and Munk [15]. Features have been calculated by applying Eq. (1) to each interval. A data matrix whose rows are the 61 samples and whose columns are the 245 variables (wavenumber intervals) was built. The elements of this matrix are the INT features in each of the predefined intervals for one of the substances.

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

The data matrixes were then imported into Matlab version 4.2 (The Mathworks, v4.2c.1) to perform PCA and sequential projection pursuit (SPP) so that exploratory data analysis could be conducted.

### 2.4. Chemometric analysis of the FT-IR spectral data

#### 2.4.1. Pre-processing of the FT-IR spectral data

Prior to the actual data analysis the original data is often transformed to eliminate the differences in variable dimensions [10]. An often applied transform is the log transform. This transform has the advantage that differences in variation are minimized so that variables will have equal importance in the analysis [17,18,22].

#### 2.4.2. PCA

PCA was used, as the multivariate method, to reduce the size of the space of the variables and visually represent a clustering of the substances [7,19]. Infrared spectral features with and without a log transform pretreatment were analyzed.

#### 2.4.3. SPP

SPP is a method that reveals more easily information about inhomogeneities in the data than PCA [23]. The method is applied on the raw and log transformed infrared spectral features.

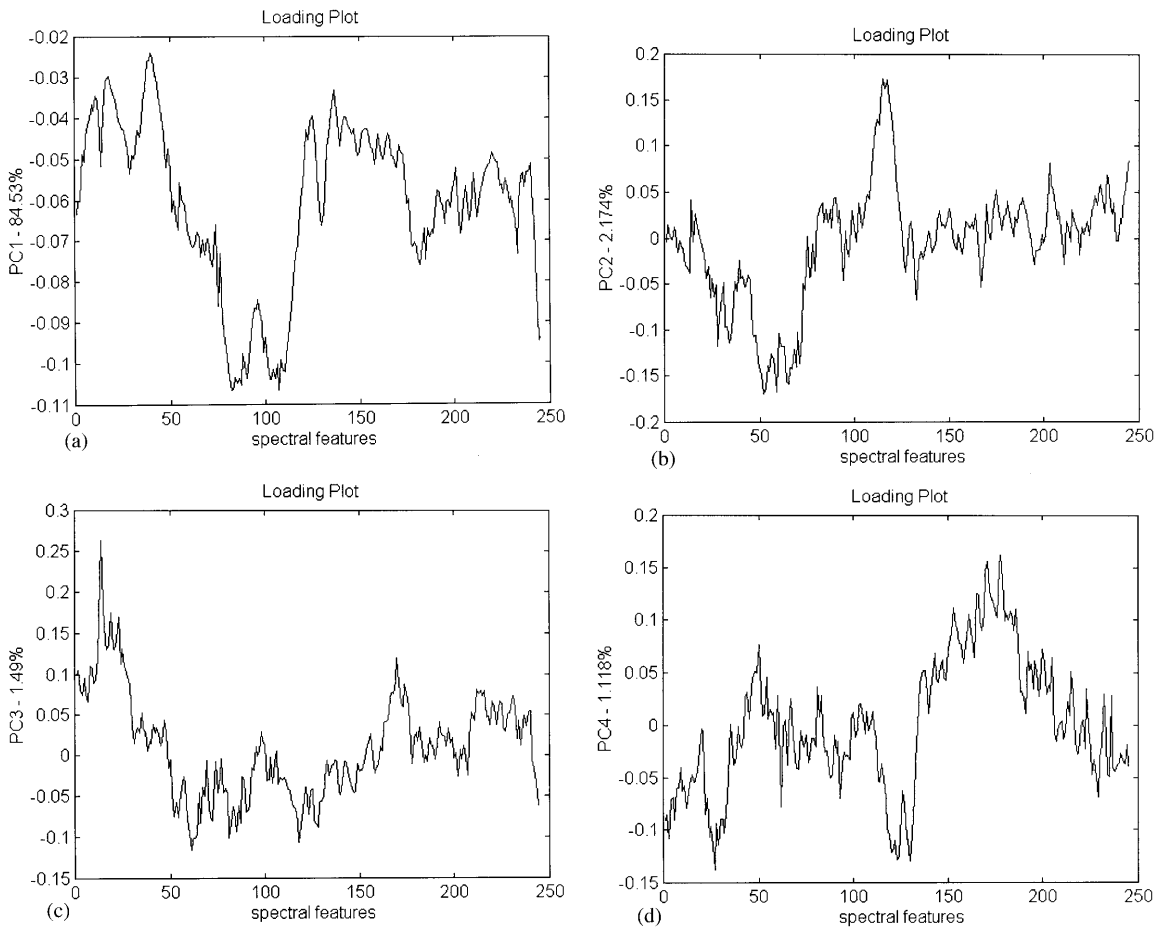


Fig. 4. (a) Principal component analysis (PCA) loading plot of the log transformed infrared spectral features, showing the first loading vector against spectral features. (b) PCA loading plot of the log transformed infrared spectral features, with the second loading vector versus spectral features. (c) PCA loading plot of the log transformed infrared spectral features, with the third loading vector plotted against spectral features. (d) PCA loading plot of the log transformed infrared spectral features, with the fourth loading vector plotted against spectral features.

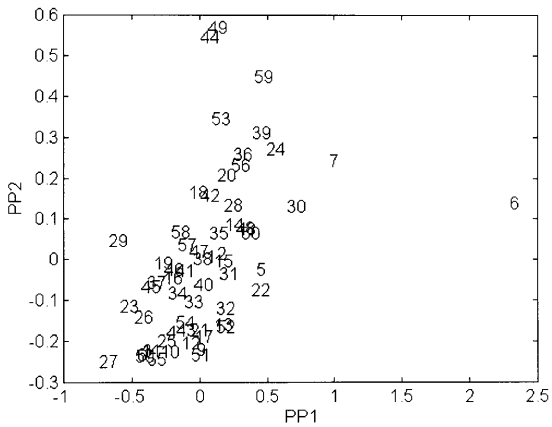


Fig. 5. Score plot from the sequential projection pursuit (SPP) of the raw infrared spectral features, showing PP2 against PP1. For the numbering of the compounds, see Table 1.

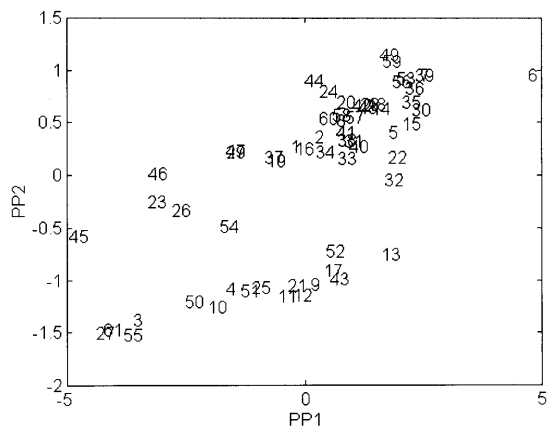


Fig. 6. Score plot from the sequential projection pursuit (SPP) of the log transformed infrared spectral features, showing PP2 against PP1. For the numbering of the compounds, see Table 1.

### 2.4.4. Cluster analysis

Clustering is the process of subdividing a set of entities into subsets in which the members are similar to each other, but different from members of other subsets [3]. A hierarchical cluster analysis of samples was performed using the 245 infrared spectral features. The correlation coefficient was used as a similarity measurement and the unweighted pair-group average method as an amalgamation rule. The results are displayed graphically as a dendrogram [10,20].

### 2.5. Comparison of classifications

Many methods for quantitatively defining the similarity between two different clusterings of the same set of objects have been proposed. In this study Wallace's measure $s_w$ (1983) was used for comparing two hierarchies of the same finite set of objects. The measure is based on a $(k \times l)$ contingency table for two different clusterings $H$ ($k$ groups) and $G$ ($l$ groups) of a same set $S$ of $n$ objects.

| | | Partition $G$ | | | | | |
|---|---|---|---|---|---|---|---|
| | Group | $g_1$ | $g_2$ | ··· | ··· | $g_k$ | Sums |
| Partition $H$ | $h_1$ | $n_{11}$ | $n_{12}$ | ··· | ··· | $n_{1k}$ | $n_{1.}$ |
| | $h_2$ | $n_{21}$ | $n_{22}$ | ··· | ··· | $n_{2k}$ | $n_{2.}$ |
| | ··· | ··· | ··· | ··· | ··· | ··· | ··· |
| | ··· | ··· | ··· | ··· | ··· | ··· | ··· |
| | $h_l$ | $n_{l1}$ | $n_{l2}$ | ··· | ··· | $n_{lk}$ | $n_{l.}$ |
| | Sums | $n_{.1}$ | $n_{.2}$ | ··· | ··· | $n_{.k}$ | $n$ |

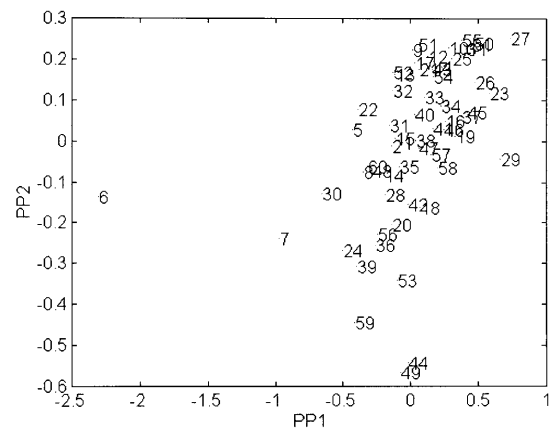


Fig. 7. Score plot from the sequential projection pursuit (SPP) of the log transformed infrared spectral features (transmissions), showing PP2 against PP1. For the numbering of the compounds, see Table 1.

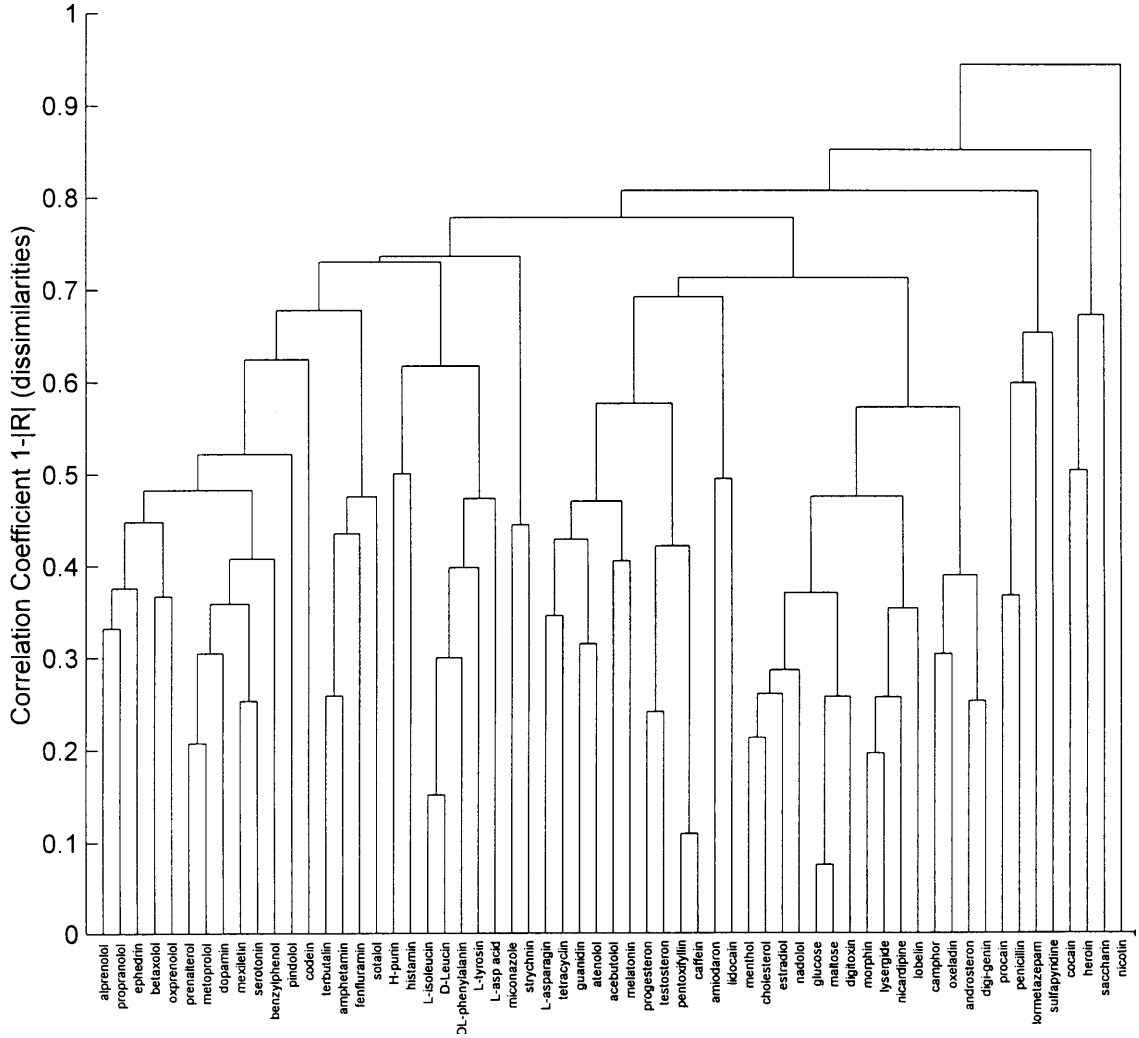


Fig. 8. Hierarchical upgma-clustering of the raw infrared spectral features.

$$s_w(G, H) = \frac{\sum_{i=1}^{l}\sum_{j=1}^{k}\binom{n_{ij}}{2}}{\sum_{i=1}^{l}\binom{n_{i.}}{2}}$$

$$s_w(H, G) = \frac{\sum_{i=1}^{l}\sum_{j=1}^{k}\binom{n_{ij}}{2}}{\sum_{j=1}^{k}\binom{n_{.j}}{2}}$$

This similarity measure gives the probability that two objects that are picked at random are placed in the same group in $G$ and in the same group in $H$. However, the measure is not symmetric, so it should be used only in cases where one partition can be considered to be the correct one.

The measure proposed by Fowlkes and Mallows (1983) is an alternative which avoids this asymmetry.

$$s_{FM}(G, H) = \sqrt{s_w(G, H)s_w(H, G)}$$

This measure, like Wallace's measure, ranges

from 0 when there is no similarity at all to 1 when the two partitions are identical.

## 3. Results and discussion

### 3.1. PCA

#### 3.1.1. PCA of the raw infrared spectral features

To obtain an overview of the dominating patterns and major trends in the data set, a PCA was first performed on the raw spectral features. The first four principal components (PCs) explained 80.1% of the total variance, of which 69.1% explained by PC1, 4.9% by PC2, 3.7% by PC3 and 2.5% by PC4. To visualize the trends of the data, the scores for samples and the loadings for variables were represented in the space of the four PCs obtained from PCA. The score plot of PC1 against PC2, PC3 and PC4 and PC2 against PC3 is shown in Fig. 1a–d, respectively. The corresponding loadings are plotted in Fig. 2(a–d).

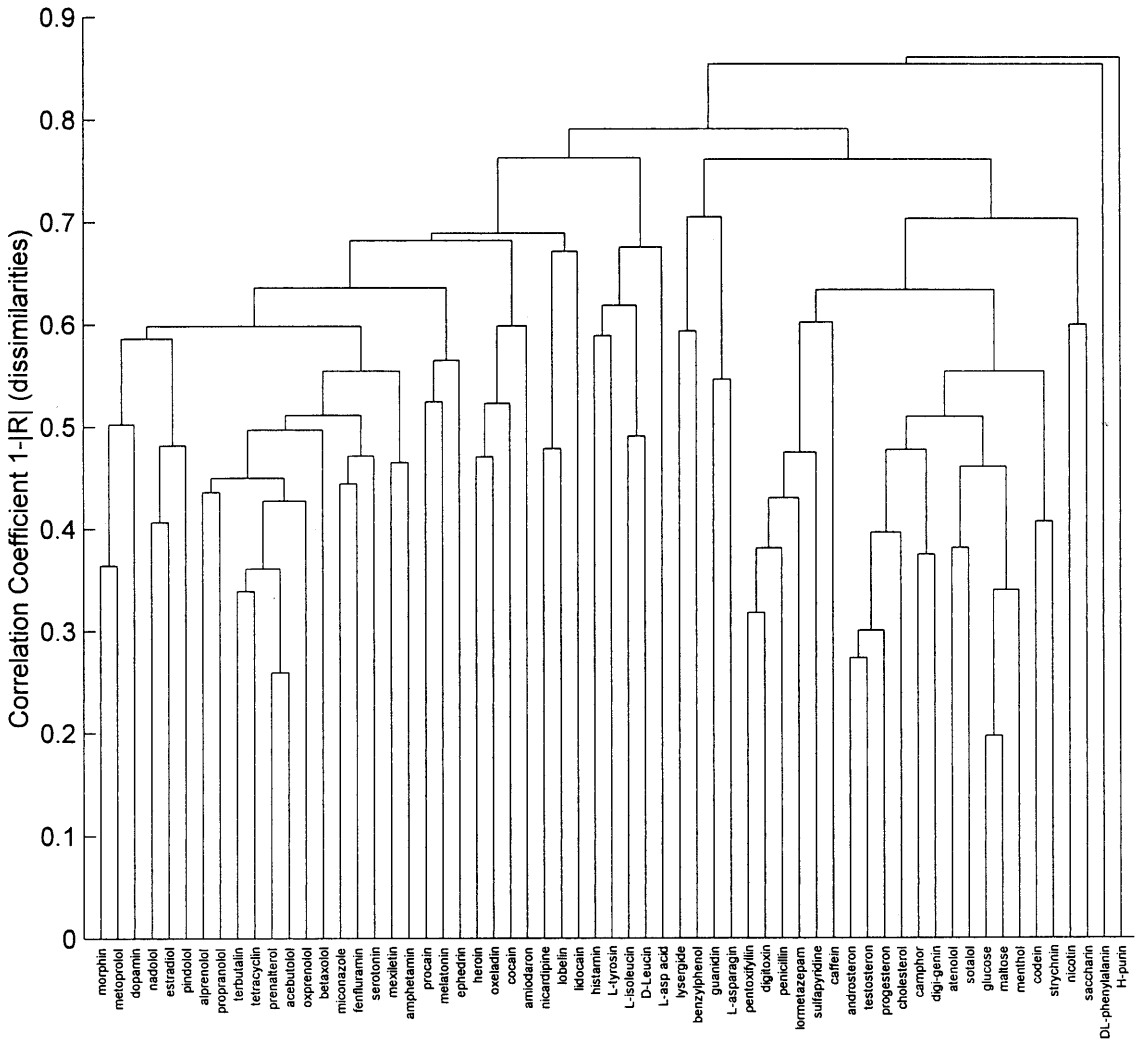


Fig. 9. Hierarchical upgma-clustering of the log transformed infrared spectral features.

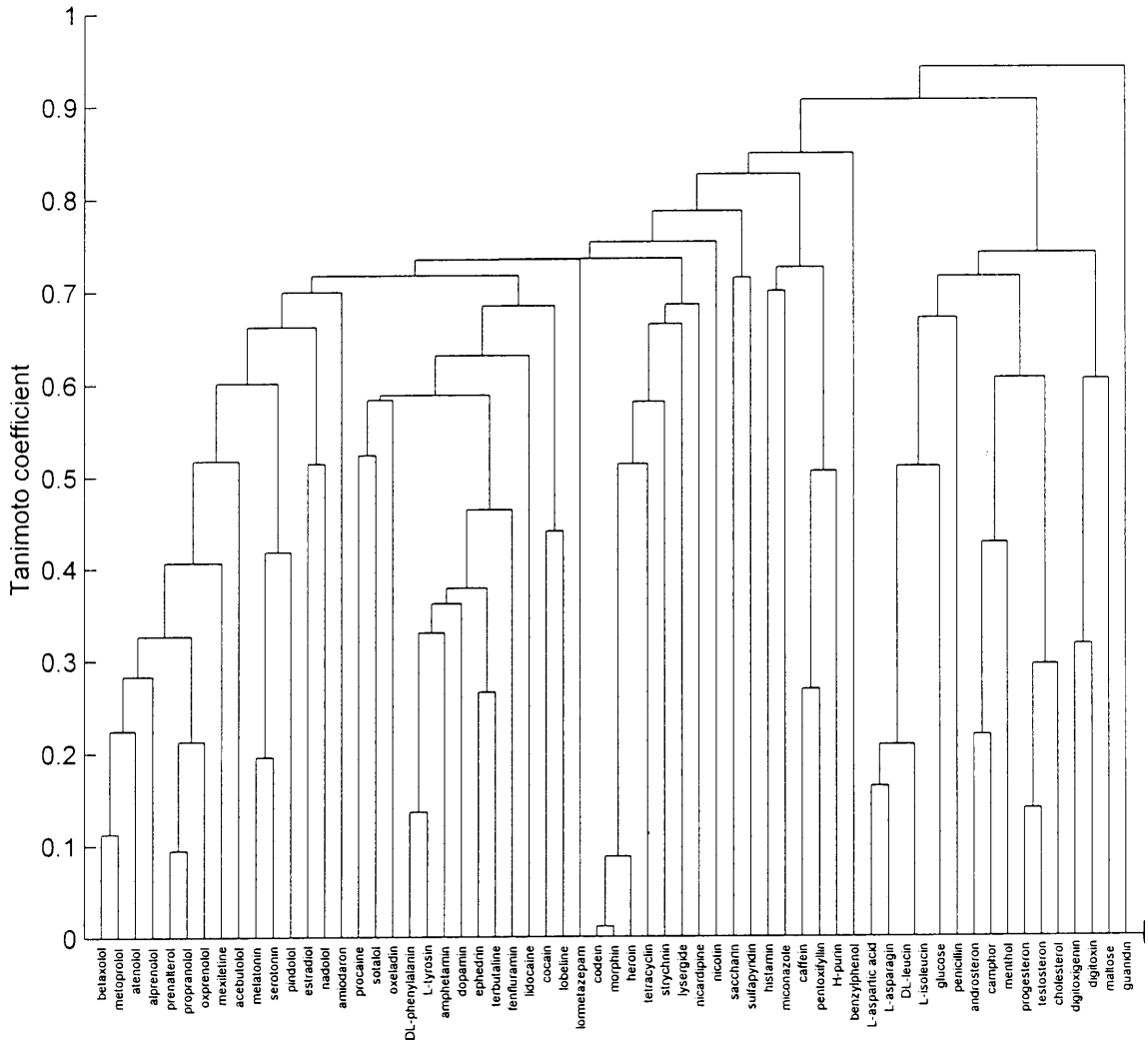


Fig. 10. Hierarchical upgma-clustering of the 2D Daylight structural fingerprints.

The best results are obtained in the score plot of PC2–PC3 (Fig. 1d). One can detect a clustering due to chemical similarity, most pronounced for the amino-acids (1) that are grouped together in the left part of the plot, the steroids (2) that are situated in the upper right region and the β-blockers (3) that appear in the lower part of the same plot. Also, maltose and glucose, as well as lidocain and procain are closely clustered in the central region.

The largest variation (Figs. 1 and 2) within the data set of 61 synthetic substances is explained along PC1 by the overall IR-absorption of the compounds investigated since all loadings are positive. The more IR peaks of high intensity a spectrum has, the higher the overall absorption of the substance. Going from left to right in Fig. 1a, the order of total absorption intensity goes from small to high. PC2 reflects the difference between those substances that are mainly characterized by features 14–20, corresponding to the wavelength region 3450–3300 $cm^{-1}$, which can be attributed to the O–H stretching vibrations of an alcohol and substances that have a characteristic N–H

bending absorption, occurring in the region 1630–1500 $cm^{-1}$ (feature 124–132). This is seen in the loading plot of Fig. 2b where features 14–20 are at the top and features 124–134 at the bottom of the plot. Compounds with a sharp absorption around 1650 $cm^{-1}$, due to the C=O stretching mode of an amide or the C=N stretching of imines are situated in the positive direction of PC3, while compounds characterized by intense C–H stretching bands that occur below 3000 $cm^{-1}$ lie in the negative direction. Some variables that we have already encountered seem also to be important for the fourth PC. A very intense absorption at around 1700 $cm^{-1}$, indicative of the C=O stretching mode of carbonyls, and between 1100 and 1000 $cm^{-1}$, due to C–O stretching vibrations of alcohols and esters is characteristic for compounds situated in the positive direction of PC4, like for example compound no. 51, 52, 54. Substances characterized by an intense band occurring at around 1600 $cm^{-1}$, due to N–H bending vibrations and above 3000 $cm^{-1}$, indicative of N–H stretching absorptions, as for example compound no. 45, 49 lie in the negative direction of PC4.

1: guanidin
histamin
serotonin
melatonin
pentoxifyllin
caffein
nicotin
H-purin

2: saccharin
sulfapyridin
penicillin

3: miconazole
nicardipin
lobelin
cocain

4: codein
morphin
heroin
lysergide
strychnin
tetracyclin
estradiol
progesteron
testosteron
androsteron
cholesterol
digitoxigenin
digitoxin
benzylphenol
maltose
glucose
menthol
camphor

5: leucin
isoleucin
aspartic acid
asparagine
tyrosin
phenylalanin
amphetamin
fenfluramin
dopamin
ephedrin

6: terbutalin
nadolol
propranolol
pindolol
oxprenolol
alprenolol
metoprolol
betaxolol
atenolol
acebutolol
sotalol
prenalterol
lidocain
mexiletin
oxeladin
procain
amiodaron
lormetazepam

Fig. 11. Expert's classification of the set of 61 substances.

PCA-mapping shows that raw mid-infrared spectra indeed contain at least some information about structure since clusters of similar objects are formed in spectral data space.

### *3.1.2. PCA of the log transformed infrared spectral features*

A PCA-analysis was also performed on the log transformed infrared spectral features and the resulting PCA-plots, in which the first four PCs accounted for 89.3% of the total variation in the data, are shown in Fig. 3a–d. The first PC described 84.5% of the variance and the second, third and fourth PC 2.2, 1.5 and 1.1%, respectively. Fig. 3a–d show the score plot of PC2 against PC1, PC3 against PC1, PC4 against PC1 and PC3 against PC2 respectively. The corresponding loadings are plotted in Fig. 4a–d, respectively.

The best discriminant plot is shown in Fig. 3d, in which three well-separated groups appear, one of them containing all steroid cases and the other two containing the β-blockers and amino-acid samples. Also, both sugars, maltose and glucose, lie closely clustered in the top region.

Looking at the score plots and loading plots (Figs. 3 and 4) shows that PC1 again detects the overall IR-absorption as being the main feature in all the spectra investigated, since substances with very little IR absorption appear in the right part of Fig. 3a, such as, for instance compound no. 27, 45, 55 while substances that strongly absorb are in

Table 2
1: comparison with four largest clusters of the respective clusterings; 2: comparison with six largest clusters of the clusterings, based on log transformed IR spectral features and with five largest clusters of the clustering, based on raw features and Daylight fingerprints

| | Expert's/raw INT features | Expert's/log INT features | Expert's/Daylight fingerprints |
|---|---|---|---|
| 1. | 0.4458 | 0.4343 | 0.4197 |
| 2. | 0.3918 | 0.4160 | 0.4390 |

Table 3
1: comparison of four largest clusters of the respective clusterings; 2: comparison of five largest clusters of the clustering, based on raw features with six largest clusters of clustering based on log transformed features; 3: comparison of five largest clusters of clustering based on Daylight fingerprints with six largest clusters of the resp.clusterings, based on INT features

| | Raw INT features/log INT features | Raw INT features/Daylight fingerprints | log INT features/Daylight fingerprints |
|---|---|---|---|
| 1. | 0.6030 | 0.5699 | 0.5946 |
| 2. | 0.3921 | | |
| 3. | | 0.4306 | 0.6310 |

the left part, for example compound no. 6. The second PC differentiates between substances with characteristic broad absorption bands in the range 3150–2500 $cm^{-1}$, indicative of the C–H stretching mode and compounds that have a strong absorption between 1820 and 1660 $cm^{-1}$, due to C=O stretching vibrations. In PC3, the compounds that show a strong OH-stretching absorption are separated from the rest. This is also seen in Fig. 4c where the variables related to this wavenumber interval (features 13–15) have a high positive loading. The fourth PC explains almost the same information as described by PC4 for the raw infrared spectral features.

PCA-analysis of IR-spectral features after a logarithmic transformation pretreatment shows the same characteristic features for assessing similarity as raw IR spectral features and therefore this transform does not seem necessary for good results.

### 3.2. SPP

#### 3.2.1. SPP on raw infrared spectral features

The SPP-results obtained on the raw infrared spectral features are shown in the score plot of PP1–PP2 (Fig. 5). Compared with PCA, SPP fails to separate the different groups of similar compounds. However, SPP is not meant to find groups in the data, but to detect inhomogeneities. In Fig. 5, one outlier is found namely compound no. 6 (L-aspartic acid) in the positive direction of PP1. With PCA, this outlying object in the data can also be distinguished on PC2, although not so clearly (Fig. 1a). This substance may be regarded as an inhomogeneity in the data, probably due to its intense broad N–H bending absorption with respect to the other compounds in the data set. Along PP2, one can observe two outliers, compounds no. 44 (camphor) and 49 (lysergide). These objects can not be observed as outliers in the resulting PCA-plots (Fig. 1), with the exception of compound No. 49 that can slightly be detected in the negative direction of PC4. This substance is marked by a broad high intensity N–H stretching absorption and may be regarded as an inhomogeneity in the data.

The results show that SPP can be regarded as an interesting tool to find outliers in the data set that probably influence the clustering tendency in PCA.

#### 3.2.2. SPP on log transformed infrared spectral features

After a logarithmic transformation of the infrared spectral features, one can detect a layered

structure of two elongated clusters in the score plot of PP1–PP2 (Fig. 6). The bottom cluster contains all those substances that absorb very little due to many features set to zero, the others appear in the upper cluster. Those zero values were replaced by a small number that approaches zero, i.e. 0.0001. Since a logarithmic transformation converts them into a highly negative number, the substances in the bottom cluster are characterized by a high negative value. The group of β-blockers can slightly be identified in the upper right part of Fig. 6. It shows that SPP can be used to find out such artefacts in the data that can not be detected with PCA and might influence the results in PCA. Therefore, a logarithmic transformation does not seem to be the best approach in IR-spectroscopy. One can avoid the problem by using transmissions instead of absorptions. This is shown in the plot of Fig. 7. The layered effect disappears and one can observe the inverse of the plot in Fig. 5.

One outlier (compound no. 6) can be identified in the direction of PP1. This object is also clearly outlying along PC1 (Fig. 3a) and therefore can be regarded as an inhomogeneity in the data, probably due to its strong IR-absorption with respect to the other compounds in the data. Again, the results show that SPP indeed detects outlying objects in the data better than PCA does.

Since no extra information is supplied after a logarithmic transformation of the data, it seems better to work with raw spectral features instead of log transformed spectral features.

### 3.3. Qualitative comparison of upgma-clusterings

Hierarchical cluster analysis, based on unweighted pair-group average linkage and the correlation coefficient, was also used for data classification and the resulting dendrograms for raw and log transformed infrared spectral features can be observed in Figs. 8 and 9, respectively. An examination of both hierarchical upgma-clusterings clearly shows that smaller clusters of similar structures are nested within larger clusters containing progressively more diverse structures. At a cut-off value of $\pm 0.75$–$0.77$, both Figs. 8 and 9 present the formation of two main clusters and some smaller ones. The first main cluster includes most β-blockers in the classification based on raw infrared spectral features. Most amino-acids are found in one subgroup of this respective cluster. The second main cluster of chemical structures is considerably more heterogeneous, but consists of smaller, more homogeneous groups of similar compounds, for example maltose and glucose, as well as morphine and lysergide are linked together in one smaller subcluster. The steroids appear more dispersed over the tree.

In the classification, based on log transformed spectral features, most β-blockers are found together in the first main cluster. Also, most amino-acids are contained in a small subgroup of this cluster. Almost all steroids are located near each other in one subgroup of the other main cluster, as well as maltose and glucose that are linked together in a smaller subgroup of this cluster.

The upgma-classification of the Daylight structural fingerprints is presented in Fig. 10. The Tanimoto coefficient was used as similarity measure. In the resulting tree-structure, some clusters of very similar compounds can be observed, for example, most amino-acids (L-aspartic acid, L-asparagin, DL-Leucin, L-isoleucin) are included in one cluster, as well as most steroids. Also, the group of β-blockers is found as such in one cluster. Another example is given by the alkaloids, codein, morphin and heroin, as well as melatonin and serotonin, camphor and menthol and the purine derivatives, caffein, pentoxifyllin and purin that are linked together in the tree-structure.

### 3.4. Quantitative comparison of upgma-clusterings

The measure of Wallace is applied to obtain more quantitative information about the similarity between two different classifications of the same set of compounds. The upgma-clusterings have been quantitatively compared between them and with an expert's classification of the same set of objects, according to known structure and pharmacological activity. In Fig. 11 the expert's classification, composed of six groups, is presented. It has to be stressed that other classifications might be proposed by other experts due to the selection of the set of compounds, some of

which are quite different from one another. From the comparison of the different upgma-clusterings with the expert's classification, no marked differences can be noted between the different comparisons (Table 2): all three upgma-classifications seem to compare equally with the classification, based on expert judgement. However, the classification, based on Daylight structural fingerprints produces the best results.

The results of the quantitative comparison of the different upgma-clusterings between them are shown in Table 3. From this comparative study, it seems that the different classifications are quite similar to each other. However, comparing the five largest clusters of the clustering, based on Daylight fingerprints with the six largest clusters of the clusterings, based on raw and log transformed spectral features, the classification of the log transformed spectral features compares most with the classification, based on Daylight structural fingerprints.

From the results, it seems that the classification of infrared spectral features is very similar to the classification based on structural characteristics. Therefore, it seems that mid infrared spectra can provide enough characteristic information to group compounds into structurally similar classes. Also, a logarithmic transformation pretreatment does not seem necessary for good clustering.

## 4. Conclusion

In this preliminary work, an evaluation whether FT-IR spectroscopy provides enough structural information for determining the molecular similarity/diversity of chemical compounds is presented. A comparative study was carried out between an upgma-clustering, based on Daylight structural fingerprints or infrared spectral features and both results were compared with an expert's classification of the same set of compounds. From our results, it seems that the upgma-clusterings based on Daylight structural fingerprints and IR spectral features are of similar quality and therefore it seems one does not loose much information using spectral characteristics instead of structure. However, a logarithmic transformation of the data does not seem necessary for good clustering.

## Acknowledgements

The authors are grateful to Mr F. Parmentier, Head of the Department of Bromatologie, Instituut voor Hygiene en Epidemiologie, for placing their FT-IR spectrometer at our disposal and Mr C. Jacques for collaboration in the IR-measurements.

## References

[1] J.P. Atherton, T.J. Van Noord, B.-S. Kuo, Sample pooling to enhance throughput of brain penetration study, J. Pharm. Biomed. Anal. 20 (1999) 39–47.
[2] J.-L. Fauchère, J.A. Boutin, J.-M. Henlin, N. Kucharczyk, J.-C. Ortuno, Combinatorial chemistry for the generation of molecular diversity and the discovery of bioactive leads, Chemometrics Intelligent Lab. Syst. 43 (1998) 43–68.
[3] D.H. Drewry, S.S. Young, Tutorial: approaches to the design of combinatorial libraries, Chemometrics Intelligent Lab. Syst. 48 (1999) 1–20.
[4] J.S. Delaney, Assessing the ability of chemical similarity measures to discriminate between active and inactive compounds, Mol. Divers. 1 (1995) 217–222.
[5] D. Gorse, A. Rees, M. Kaczorek, R. Lahana, Molecular diversity and its analysis, Drug Discov. Today 4 (1999) 257–264.
[6] J. Zupan, M.E. Munk, Hierarchical tree based storage, retrieval, and interpretation of infrared spectra, Anal. Chem. 57 (1985) 1609–1616.
[7] D.S. Frankel, Pattern recognition of Fourier transform infrared spectra of organic compounds, Anal. Chem. 56 (1984) 1011–1014.
[8] P.M. Dean (Ed.), Molecular Similarity in Drug Design, Blackie Academic Professional, London, 1995.
[9] J.M. Barnard, G.M. Downs, Clustering of chemical structures on the basis of 2-D similarity measures, J. Chem. Inf. Comput. Sci. 32 (1992) 644–649.
[10] D.L. Massart, L. Kaufman, The Interpretation of Analytical Chemical Data by the Use of Cluster Analysis, Wiley, New York, 1983.
[11] E.W. Robb, M.E. Munk, A neural network approach to infrared spectrum interpretation, Mikrochim. Acta 1 (1990) 131–155.
[12] H. Scsibrany, K. Varmuza, Common substructures in groups of compounds exhibiting similar mass spectra, Fresenius J. Anal. Chem. 344 (1992) 220–222.

[13] W. Werther, K. Varmuza, Exploratory data analysis of infrared spectra, Fresenius J. Anal. Chem. 344 (1992) 223–226.
[14] F. Ehrentreich, Representation of extended infrared spectrum-structure-correlations based on fuzzy theory, Fresenius J. Anal. Chem. 357 (1997) 527–533.
[15] E.W. Robb, M.E. Munk, A neural network approach to infrared spectrum interpretation, Mikrochim. Acta [Wien] I (1990) 131–155.
[16] 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.
[17] W. Vogt, D. Nagel, H. Sator, Cluster Analysis in Clinical Chemistry: A Model, Wiley, New York, 1987.
[18] I.E. Frank, R. Todeschini, Data Handling in Science and Technology: The Data Analysis Handbook, Elsevier, Amsterdam, 1994.
[19] P. Andersson, P. Haglund, C. Rappe, M. Tysklind, Ultraviolet absorption characteristics and calculated semi-empirical parameters as chemical descriptors in multivariate modelling of polychlorinated biphenyls, J. Chemometrics 10 (1996) 171–185.
[20] 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.
[21] R.D. Brown, Y.C. Martin, Use of structure-activity data to compare structure-based clustering methods and descriptors for use in compound selection, J. Chem. Inf. Comput. Sci. 36 (1996) 572–584.
[22] 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.
[23] 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, accepted for publication in Analytical Chemistry.
[24] K. Baumann, J.T. Clerc, Computer-assisted IR spectra prediction — linked similarity searches for structures and spectra, Anal. Chim. Acta 348 (1997) 327–343.
[25] R.D. Brown, Y.C. Martin, Use of structure-activity data to compare structure-based clustering methods and descriptors for use in compound selection, J. Chem. Inf. Comput. Sci. 36 (1996) 572–584.
[26] 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.
[27] C.A. James, D. Weininger, Daylight Theory Manual, Daylight Chemical Information Systems, 3951 Claremont St, Irvine, California 92714, USA, 1993.
[28] 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.