Journal of Pharmaceutical and Biomedical Analysis
21 (2000) 1197–1214

JOURNAL OF
PHARMACEUTICAL
AND BIOMEDICAL
ANALYSIS

www.elsevier.com/locate/jpba

Use of mass spectrometry for assessing similarity/diversity of natural products with unknown chemical structures

V. Schoonjans [a], F. Questier [a], A.P. Borosy [b], B. Walczak [c], D.L. Massart [a,*], B.D. Hudson [d]

[a] *ChemoAc, Pharmaceutical Institute, Vrije Universiteit Brussel, Laarbeeklaan 103, B-1090 Brussels, Belgium*
[b] *Department of Chemical Informatics, Faculty of Chemical Engineering, Technical University of Budapest, Gellért tér 4, Budapest, H-1119, Hungary*
[c] *Silesian University, 40-006 Katowice, 9 Szkolna Street, Poland*
[d] *Glaxo Wellcome, Medicines Research Centre, Compound Diversity Unit, Stevenage, Herts, SG1 2NY, UK*

Received 5 May 1999; received in revised form 7 September 1999; accepted 17 September 1999

Abstract

An evaluation whether mass spectral data contain useful information for assessing similarity/diversity of drug compounds is presented. A comparative study was carried out between Ward's hierarchical agglomerative clustering, based on the 2D Daylight fingerprints or on the mass spectra, of a small database of 66 synthetic substances. The influence of normalization of the mass spectral data on the clustering result has also been studied. The results were subsequently compared with an expert's classification of the same small dataset, based on own evaluation according to known structure and pharmacological activity. © 2000 Elsevier Science B.V. All rights reserved.

Keywords: Similarity; Mass spectra; Fingerprint; Ward's hierarchical clustering; Wallace's measure

1. Introduction

Because of the interest in both directed database searching and compound selection for high throughput screening of many hundreds of thousands of compounds available in in-house databases, quantification of chemical similarity between compounds has become an important subject in pharmaceutical research. However, there is no generally agreed quantitative or even qualitative definition of chemical diversity. Much of the early work on similarity searching was concerned with the development of quantitative measures of structural resemblance between chemical molecules and the use of such similarities for clustering chemical databases. A fundamental

* Corresponding author. Present address: Department of Pharmaceutical and Biomedical Analysis, Pharmaceutical Institute, Vrije Universitet Brussel, Laarbeeklaan 103, 1090 Brussels, Belgium. Tel.: + 32-2-477-4737; fax: + 32-2-477-4735.
E-mail address: fabi@vub.vub.ac.be (D.L. Massart)

PII: S0731-7085(99)00236-8

question to be addressed is: what is the most appropriate representation of a molecule for assessment of chemical similarity? The selection of suitable properties with which to characterize every structure in a dataset has widely been studied and 2D connectivity based structural descriptors seem most effective for use in similarity calculations [1–3]. Two-dimensional structural descriptors can be divided in two classes: structural keys and hashed fingerprints. Structural keys were first developed and rely on the use of a predefined fragment dictionary. Molecular fingerprints dispense with the fragment dictionary and define a set of patterns to index. They constitute a representation of the molecular structure generated from the hashing of unique substructural paths. Hashed fingerprints are for instance used in the Daylight clustering package. Studies by Brown and Martin showed that the Daylight 2D hashed fingerprints encode a considerable amount of relevant information. The Daylight fingerprints encode each atom's type, all augmented atoms and all paths of length 2–7 atoms [4–6]. Similarity measures, based on such 2D structural fingerprints, are the most commonly used in both similarity searching and in clustering procedures [2]. A range of numerical similarity definitions, which are chemically meaningful, have been suggested and implemented. The most commonly used measures are the Euclidean distance and the Tanimoto coefficient. The latter is specific for binary data and yields better results than distance measures for measuring the similarity between fragment bit-strings [7].

Table 1
List of synthetic substances

1 Menthol	34 Testosterone
2 Maltose	35 Caffeine
3 Glucose	36 Codeine
4 Saccharin	37 Mexiletine
5 Penicilline	38 Lysergide
6 Tetracycline	39 Morphine
7 Amphetamine	40 Cocaine
8 Ephedrine	41 Lidocaine
9 Cholesterol	42 Lobeline
10 Aspartic	43 Lormetazepam
11 L-Asparagine	44 Pentoxifylline
12 DL-Leucine	45 Sulfapyridine
13 Isoleucine	46 4-Benzyl-phenol
14Tyrosine	47 Miconazole
15 Phenylalanine	48 Fenfluramine
16 Histamine	49 Nicardipine
17 Parathion	50 Oxeladin
18 Digitoxigenin	51 Flurazepam
19 Digitoxin	52 Terbutaline
20 Amiodarone	53 Phenglutarimide
21 Melatonin	54 Procaine
22 Camphor	55 Sotalol
23 Strychnine	56 Pindolol
24 Laurine	57 Timolol
25 Guanidine	58 Propranolol
26 Estradiol	59 Metoprolol
27 Nicotine	60 Nadolol
28 1H-Purine	61Acebutolol
29 Dopamine	62 Prenalterol
30 Serotonin	63 Oxprenolol
31 Heroine	64 Atenolol
32 Progesterone	65 Betaxolol
33 Androsterone	66 Alprenolol

There are many different methods that can be used to cluster a dataset, but there are no a priori guidelines as to which will be most appropriate for a particular application domain. The most widely used are the hierarchical agglomerative methods and several have been applied to cluster chemical structures [1,4,10].

Compound clustering techniques are applied frequently in pharmaceutical industry laboratories for performing diversity analyses on combinatorial libraries containing a large number of compounds with known structure, to aid in the selection of a representative subset of all the compounds available. When the chemical structures are unknown, as may be the case in natural product collections, the current clustering techniques are, however, inapplicable and therefore the knowledge of diversity of natural product samples, which are often not completely pure as well, is severely limited. The different compounds must consequently be characterized by other descriptor variables, e.g. experimental parameters. These parameters must be easy to measure. Techniques as mass spectrometry (MS), infrared spectroscopy (IR), nuclear magnetic resonance spectroscopy (NMR), in combination with chromatographic techniques are capable of providing structure-related information and these techniques are thus likely candidates [11].

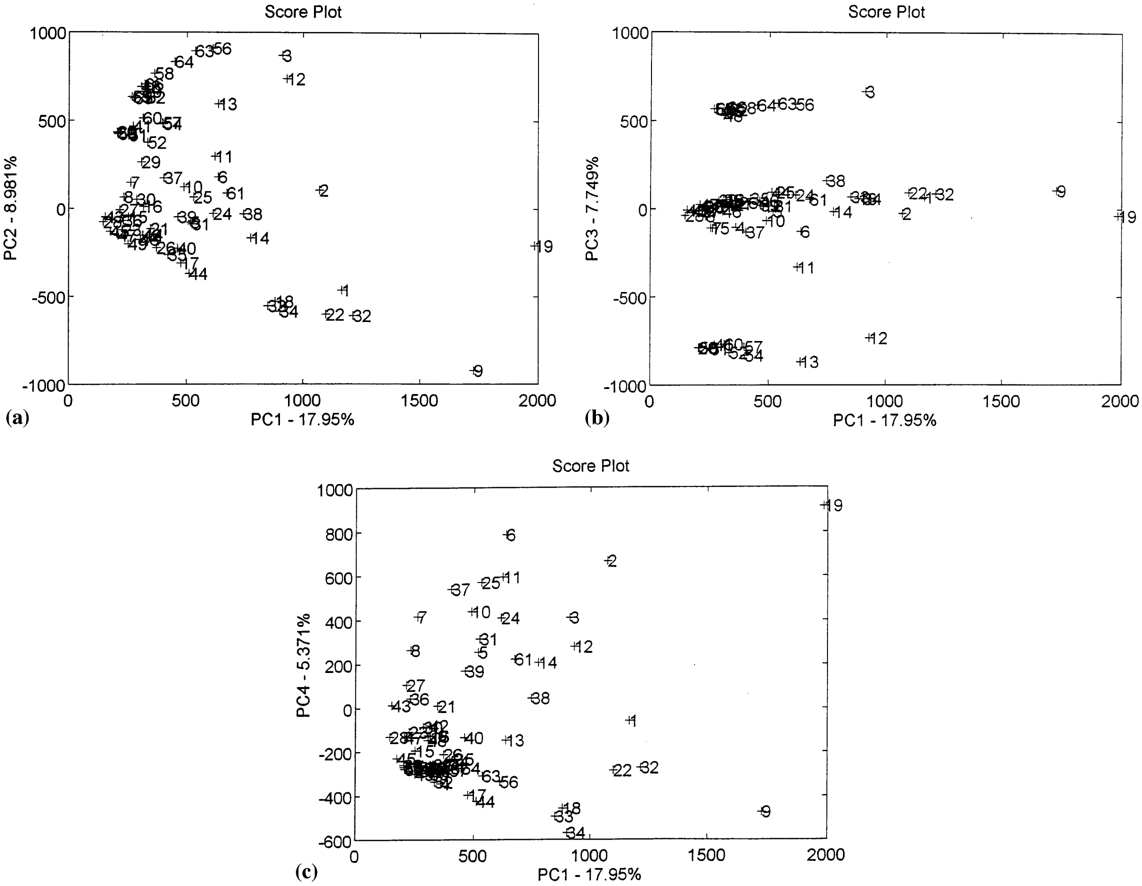

Fig. 1. (a) Score plot from the PCA of the raw mass spectral data, showing PC2 against PC1. For the numbering of the compounds, see Table 1. (b) Score plot from the PCA of the raw mass spectral data, showing PC3 against PC1. Notation as in Fig. 1(a). (c) Score plot from the PCA of the raw mass spectral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

In this paper we present a study aimed at investigating whether clustering techniques, using mass spectral data, can be applied for assessing similarity/diversity of chemical compounds and we investigate how much information is lost by using these analytical characteristics instead of the chemical structure for characterizing similarity/diversity. Ward's hierarchical method was used to cluster a small dataset of 66 synthetic substances, using the Euclidean distance (mass spectral data) and the Tanimoto coefficient (structural fingerprints) as similarity measures, and the clusters produced are evaluated to see whether spectral data give a similar grouping of the set of compounds as structural data.

2. Theory

2.1. Data

A small dataset of 66 synthetic substances, among which a number of structurally similar compounds (e.g. β-blockers, amino-acids) and a number of arbitrarily chosen substances was selected. Both mass spectra and structure are known for all substances. All compounds are listed in Table 1.

The mass spectra are electron impact spectra, obtained from the NIST/EPA/NIH Mass Spectral Database for PC. The mass spectra were available in the form of peak lists, with the relative intensity of each fragment ion (*m*/*q*-ratio) mentioned. A data matrix (66*390) was created from these peak lists, where the rows correspond to the 66 compounds and the columns to the 390 *m*/*q* (mass to charge)-values. The values in the matrix are the intensities of the respective ions and range from 0 (no peak for this ion) to 1000 (mass most important peak).

The 2D structural fingerprints for the same substances and the Tanimoto distance matrix were obtained using the Daylight clustering software.

2.2. Analysis of the data

2.2.1. Transformation of the mass spectral data

Transformation of the original data is widely used in multivariate analysis to ensure that all the variables under consideration are measured on a comparable scale. To eliminate the possibility that the contribution of a few of the variables that are being used to characterize the molecules will mask the contributions of all of the other variables, the raw data matrix of descriptor values can be normalized [1,7,13]. Because the intensities of the

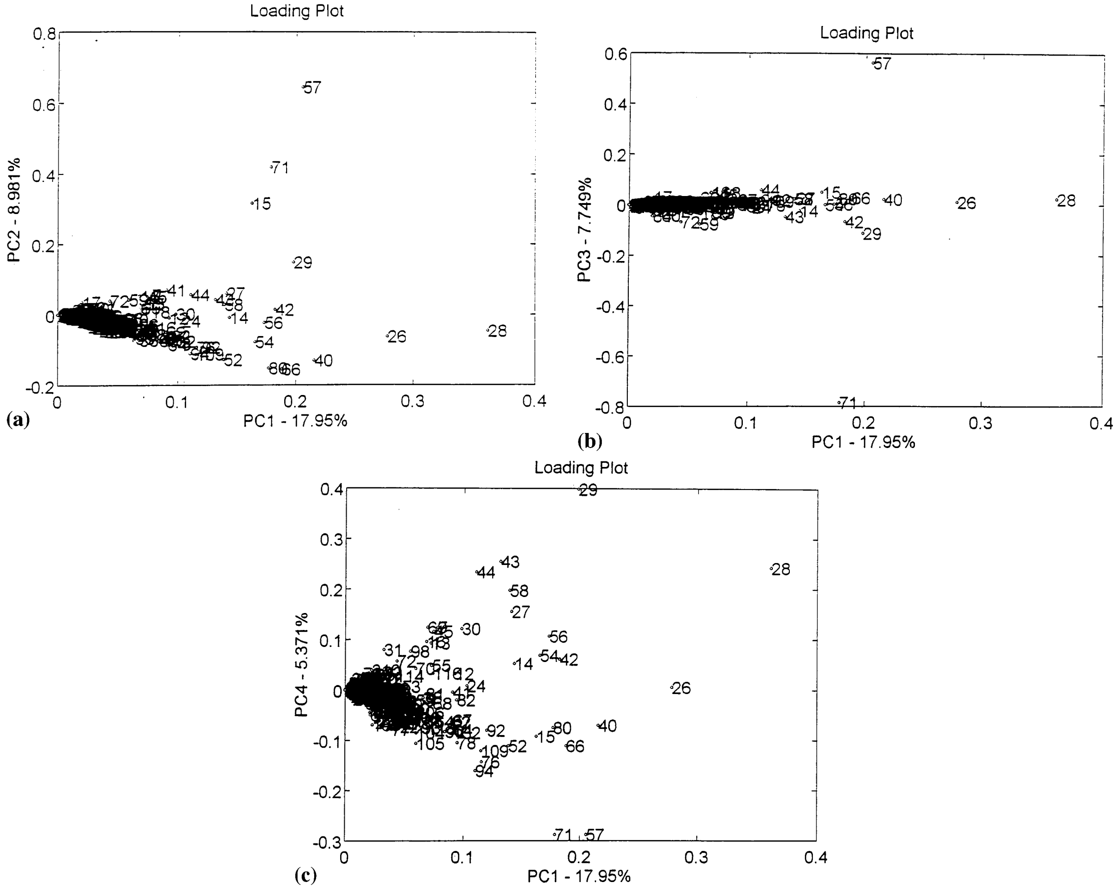

Fig. 2. (a) Loading plot from the principal component analysis of the raw mass spectral data. The second loading vector is plotted versus the first vector. (b) Loading plot from the PCA of the raw mass spectra, with the third loading vector plotted against the first loading vector. (c) Loading plot from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

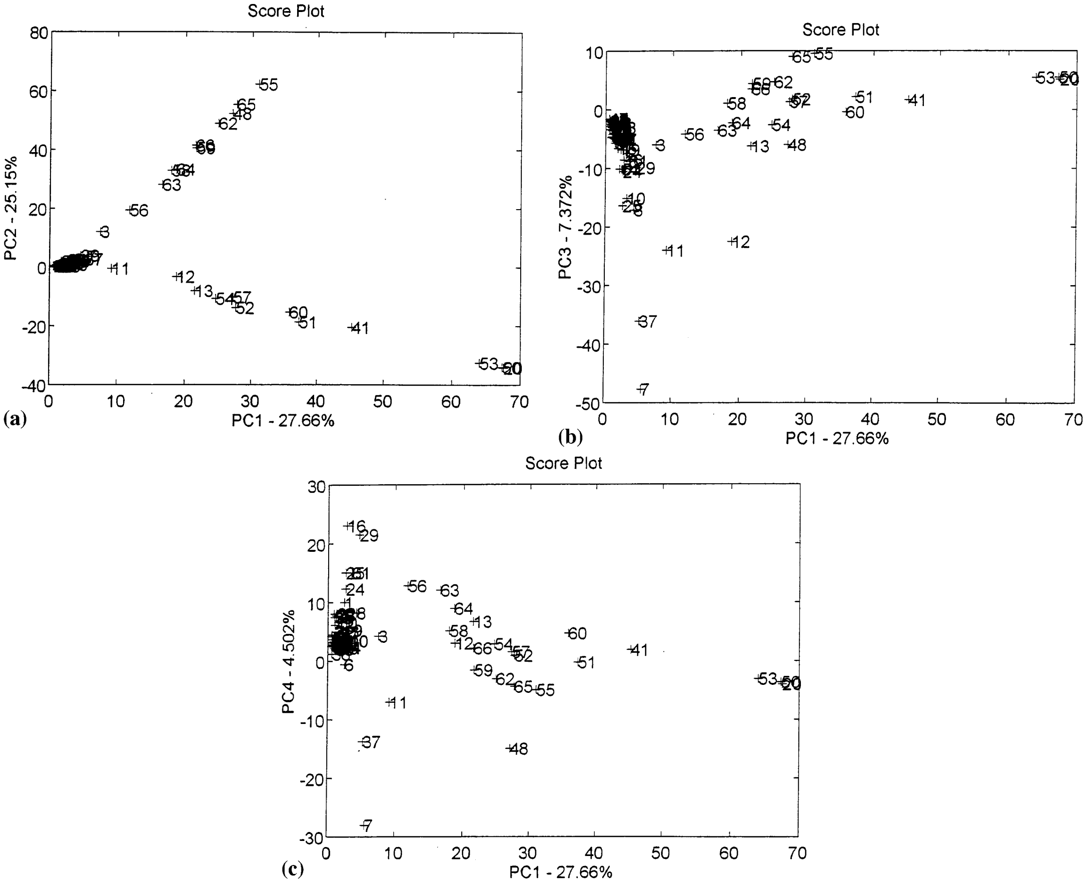

Fig. 3. (a) Score plot from the PCA of the normalized mass spectra, with PC2 plotted against PC1. For the numbering of the compounds, see Table 1. (b) Score plot from the PCA of the normalized mass spectra, showing PC3 against PC1. Notation as in Fig. 3(a). (c) Score plot from the PCA of the normalized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

respective ions of each substance in our mass spectral data matrix range from 0 to 1000, a normalization of the raw data for total mass equal to 100 is performed. Both the original and normalized data were analyzed.

A logarithmic transformation has the advantage that differences in variation are reduced so that variables with similar relative variation will have equal importance [14]. This transform is applied on both raw and normalized data.

2.2.2. Principal component analysis (PCA)

To achieve a first visualization of the information content in the data set, the data matrix, with and without normalization, was analyzed with principal component analysis (PCA) [14,15].

2.2.3. Cluster analysis

Numerous clustering methods have been described in the literature, among whom the most widely used are the hierarchical agglomerative methods, which are non-overlapping and therefore no structure appears in more than one cluster. A hierarchical clustering method produces a classification in such a way that small clusters of very similar compounds are included in larger and larger clusters of less similar molecules. The cluster hierarchy is mostly visualized as a dendrogram. Comparisons have shown Ward's hierarchical agglomerative method, which is used in the present study, to be best able to separate similar and dissimilar structures [1,4,7]. The objective of this method is to find at each stage those two clusters whose fusion gives the minimum increase in the total within-groups error sum of squares, which means minimum loss of information at every step of fusing two groups [13]. Ward's method is applied with the Euclidean distance, which is a measure of the geometric distance between two structures in a multidimensional space, as similarity measure. This approach is only used for the mass spectral classification. Similarity assessments of binary molecular representations commonly use the Tanimoto coefficient, based on the comparison of common bits in compared bitstrings or fingerprints, as a measure of chemical similarity [8,9,12]. The Tanimoto coefficient for each pair of fingerprints gives a symmetrical matrix of dissimilarities [16]. This approach is also implemented in the Daylight software package. Therefore, the Tanimoto distance matrix was used as input for the hierarchical Ward's clustering program (Statistica, version 5.x) for the classification, based on the 2D structural fingerprints.

2.3. Comparison of classifications

To measure quantitatively the similarity between two different clusterings of the same set of objects, a numerical measure of correspondence between classifications is needed. The methodology used in this study is based on the Wallace's measure s_w (1983), based on a $(k \times l)$ contingency table for two different partitions H (k groups) and G (l groups) of a same set S of n objects.

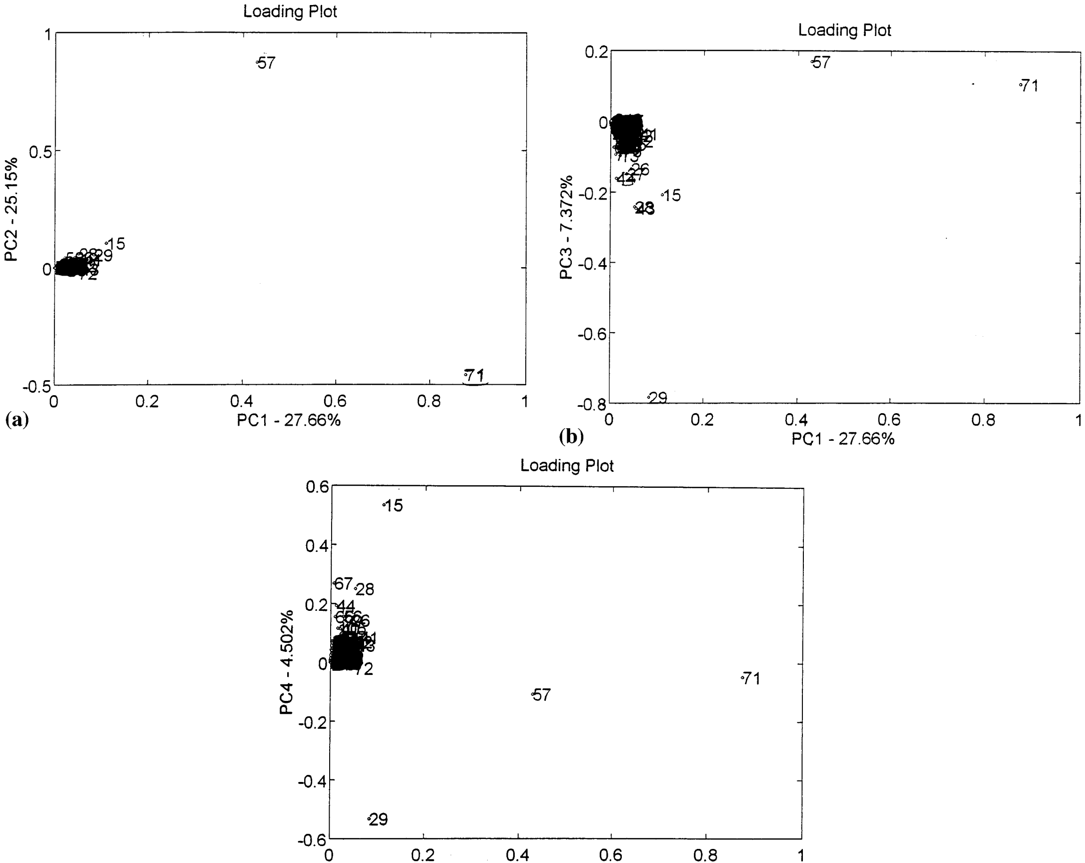

Fig. 4. (a) PCA loading plot of the normalized mass spectra, showing the second loading vector against the first loading vector. (b) PCA loading plot of the normalized mass spectra, with the third loading vector versus the first loading vector. (c) PCA loading plot of the normalized mass spectra, with the fourth loading vector plotted against the first loading vector.

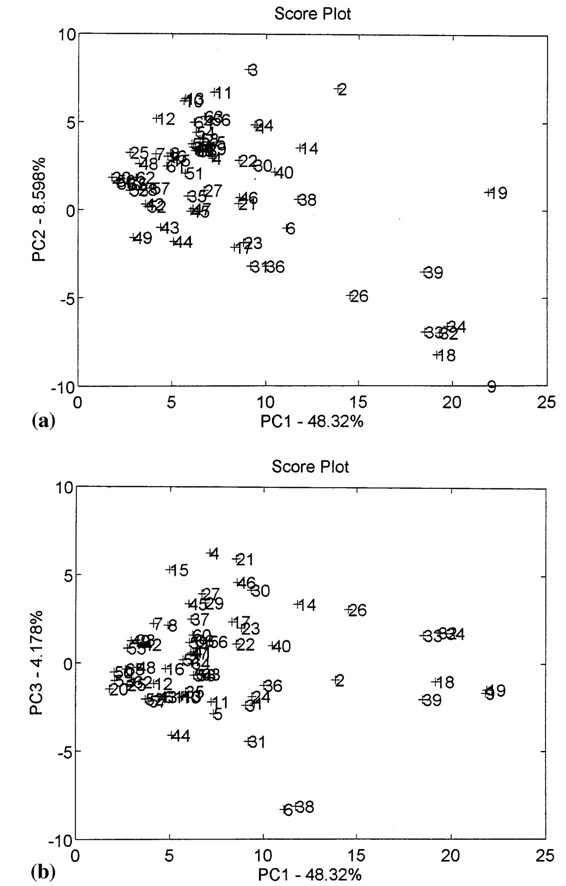

Fig. 5. (a) Score plot from the PCA of the log transformed mass spectra, showing PC2 against PC1. For the numbering of the compounds, see Table 1. (b) Score plot from the PCA of the log transformed mass spectra, showing PC3 against PC1. The numbering of the compounds is the same as in Fig. 5(a).

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

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

This similarity measure gives the probability that a randomly chosen pair of objects that is within the same class in one clustering (*H*) is also in the same class in the second clustering (*G*). The measure is not symmetric, so that it should be used only in cases where one solution can be considered to be the correct one.

An alternative which avoids this asymmetry and does not differ very much is the measure proposed by Fowlkes and Mallows s_{FM} (1983).

$$s_{\mathrm{FM}}G, H = \sqrt{s_wG, Hs_wH, G}$$

Both measures have an upper bound 1 when the two partitions are identical, and decrease with increasing number of clusters [13].

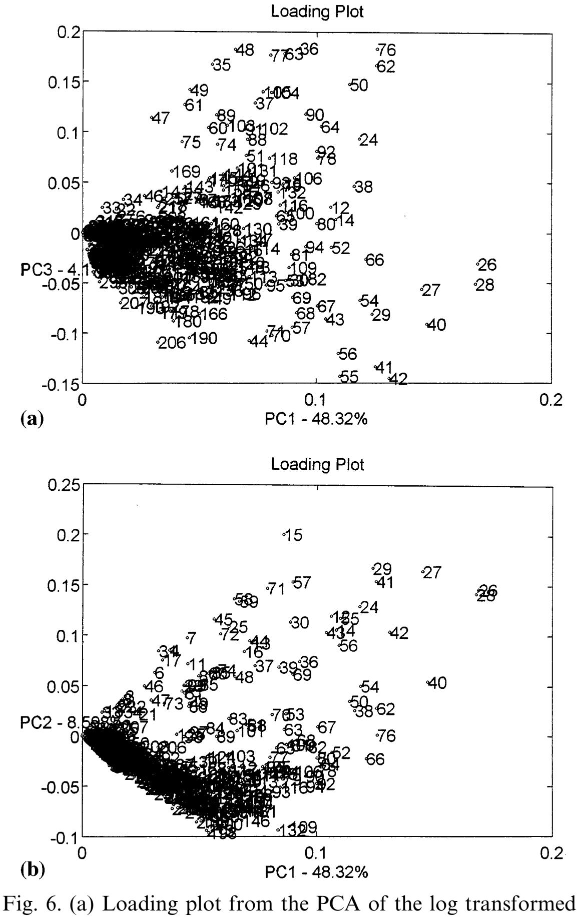

Fig. 6. (a) Loading plot from the PCA of the log transformed mass spectra, with the second vector plotted against the first loading vector. (b) Loading plot from the PCA of the log transformed mass spectra, with the third loading vector plotted against the first loading vector.

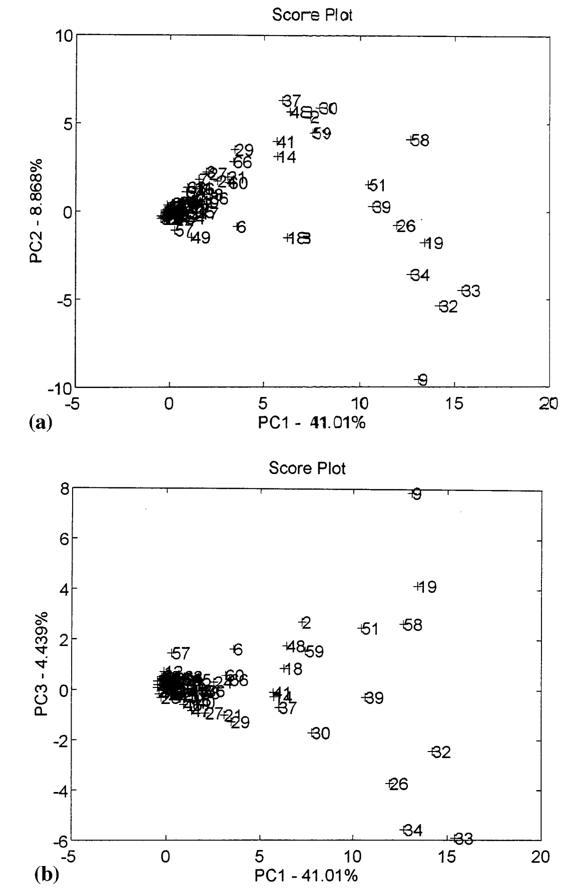

Fig. 7. (a) Score plot from the PCA of the log transformed normalized mass spectral data, showing PC2 against PC1. The numbers correspond to the numbering in Table 1. (b) Score plot from the PCA of the log transformed normalized mass spectral data, showing PC3 against PC1. The numbering of the compounds is the same as in Fig. 7(a).

3. Results and discussion

The data set of 66 synthetic compounds was selected in such a way that it consists of a relatively large class of highly similar compounds, e.g. the β-blockers, some smaller or more vague groups of similar substances and furthermore, some arbitrarily chosen compounds, which are structurally diverse.

3.1. PCA of the mass spectral data

To obtain a first impression of the information content of the mass spectral data, the raw data matrix was subjected to PCA. The first four principal components (PCs) explain 40.05% of the total variance. The first PC described 17.95% of the variance and the second, third and fourth PC 8.98, 7.74 and 5.37%, respectively.

The score plot of PC1 against PC2, PC3 and PC4 is shown in Fig. 1(a), (b) and (c), respectively. The corresponding loadings are plotted in Fig. 2(a), (b) and (c), respectively.

The score plots and the corresponding loading plots show that the first PC is related to the degree of total intensity of the fragment ions of

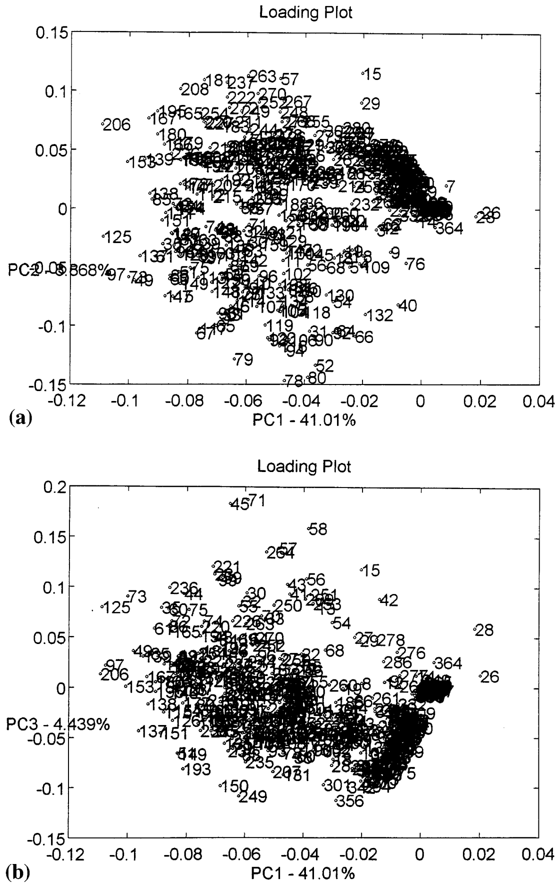

Fig. 8. (a) Loading plot from the PCA of the log transformed normalized mass spectral data, with the second loading vector plotted versus the first loading vector. (b) Loading plot from the PCA of the log transformed normalized mass spectral data, with the third loading vector plotted versus the first loading vector.

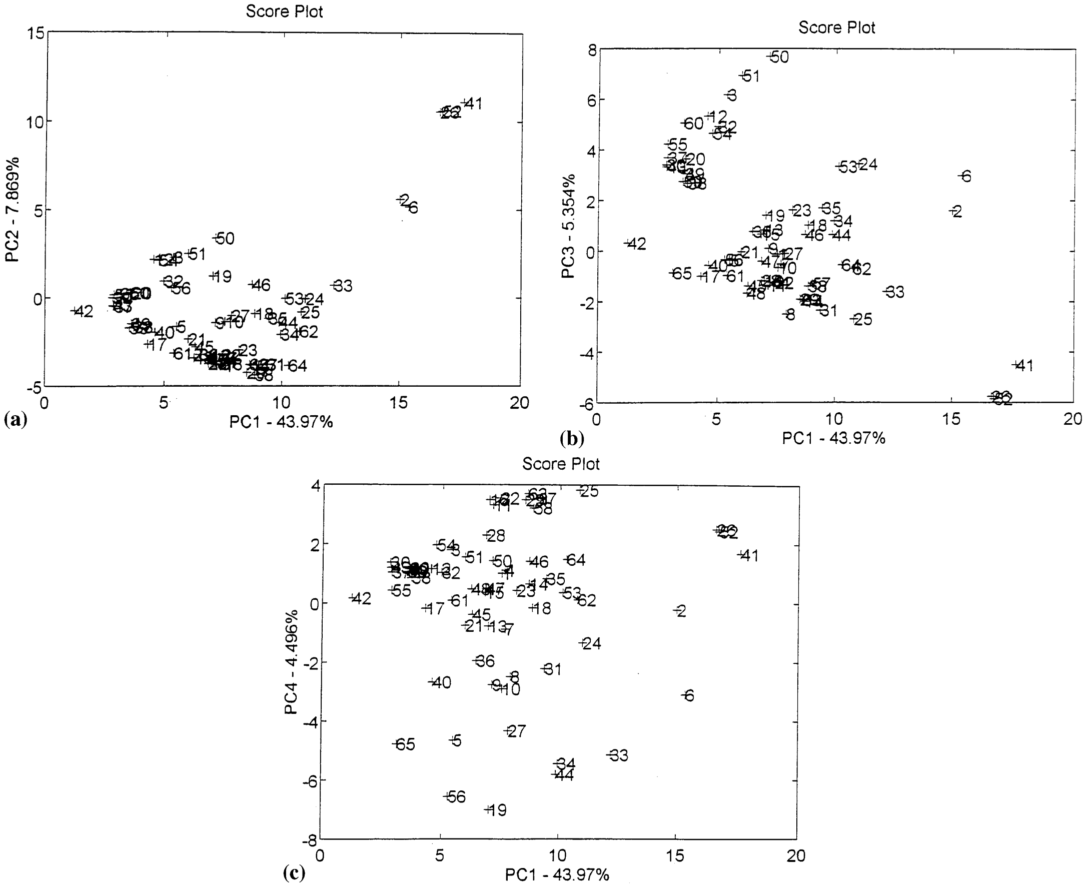

Fig. 9. (a) Score plot from the PCA of the 2D structural data, showing PC2 against PC1. The numbers correspond to the numbering in Table 2. (b) Score plot from the PCA of the 2D structural data, showing PC3 against PC1. Notation as in Fig. 9(a). (c) Score plot from the PCA of the 2D structural data, showing PC4 against PC1. Notation as in Fig. 9(b).

the substances or PC1 equals Σ (loading * intensity) since all loadings are positive. The more peaks of intensity near the most intense peak (999) a substance has, the higher its total intensity. Going from left to right in Fig. 1, one first finds substances with mostly very low intensity peaks, then substances with some abundant peaks, and finally substances with several high intensity peaks besides the basepeak.

An inspection of the score plot in Fig. 1(a) shows that all β-blockers are clustered together in the upper half of the plot. Also, most amino-acids (aspartic acid, L-asparagine, DL-leucine, isoleucine) are grouped together in the top region of the plot. The group of steroids is situated in the lower part of the same plot. Camphor and menthol appear closely clustered in the lowest region. The second PC reflects the difference between those substances with most prominent peaks at *m/q* 72 (variable 57), 86 (variable 71) and 30 (variable 15) (positive scores) and the substances that do not have such peaks or for which they are not prominent. 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. The resulting ions further break down, giving rise to $H_2N^+ = CH_2$, *m/q* 30. PC2 also represents a N-axis, since all chemical structures that contain aliphatic nitrogen appear at the top of the plot, whereas compounds that do not have nitrogen in their chemical structure appear in the lowest region. The substances appearing in the central region comprise either N-heterocycles and aliphatic N or N-heterocyclic structures only. PC3 describes the contrast between substances with prominent peaks at *m*/*q* 72 (variable 57) and *m*/*q* 86 (variable 71). This is seen in the loading plot of Fig. 2(b). where variable 57 is at the top and variable 71 is at the bottom of the plot. Correspondingly, compounds with basepeak at *m*/*q* 72 (variable 57) appear in the upper half of Fig. 1(b), such as, for instance, the β-blockers (No. 55, 56, 58, 59, 62, 63, 64, 65, 66), whereas compounds with basepeak at *m*/*q* 86 (variable 71) appear in the lower part of the same figure, for example compounds No. 20, 41, 50, 51, 52, 54, 57. The fourth PC only explains about 5.37% of the total variance. The variables that we have already encountered seem also to be important for PC4. The only new variable that is of great importance for the fourth dimension is *m*/*q* 44 (variable 29). This is seen in Fig. 2(c) where this variable has a high positive loading. Correspondingly, substances with basepeak and/or very intense peaks of *m*/*q* 44 (variable 29) and with very weak or negligible peaks at *m*/*q* 86 (variable 71) and 72 (variable 57) are situated together in the upper part of Fig. 1(c), while substances that are characterized by very intense peaks of high mass, basepeaks of high mass or at *m*/*q* 72 and 86 appear in its lower part. Fragment ions of *m*/*q* 44 can be associated with $(CH_3CH=NH_2)^+$, $(O=C=NH_2)^+$, $(CH_2=CHOH)^+$.

Table 2
Numbering of compounds, used for PCA on Daylight fingerprints

1	Tyrosine
2	Tetracycline
3	Testosterone
4	Terbutaline
5	Sulfapyridine
6	Strychnine
7	Sotalol
8	Serotonin
9	Saccharin
10	Nicotine
11	Propranolol
12	Progesterone
13	Procaine
14	Pindolol
15	Phenylalanine
16	prenalterol
17	IV-benzyl-Phenol
18	Phenglutarimide
19	Pentoxifylline
20	Penicilline
21	Parathion
22	Oxprenolol
23	Oxeladin
24	Nicardipine
25	Nadolol
26	Morphine
27	Miconazole
28	Mexiletine
29	Metoprolol
30	Menthol
31	Melatonin
32	Maltose
33	Lysergide
34	Lormetazepam
35	Lobeline
36	Lidocaine
37	Laurine
38	L-Asparagine
39	Isoleucine
40	Histamine
41	Heroine
42	Guanidine
43	Glucose
44	Flurazepam
45	Fenfluramine
46	Estradiol
47	Ephedrine
48	Dopamine
49	DL-leucine
50	Digitoxin
51	Digitoxigenin
52	Codeine
53	Cocaine
54	Cholesterol
55	Camphor
56	Caffeine
57	Betaxolol
58	Atenolol
59	Aspartic
60	Androsterone
61	Amphetamine
62	Amiodarone
63	Alprenolol
64	Acebutolol
65	1H-Purine

The PCA-analysis shows that the raw mass spectral data indeed contains at least some characteristic information, since similar compounds appear adjacent to each other in the resulting PCA-plots.

The mass spectral data matrix, after normalization, was also subjected to PCA, which resulted in 4 principal components. These PCs explained 27.66, 25.15, 7.37 and 4.50% of the total variance, respectively, i.e. together 64.68%. Fig. 3(a), (b) and (c) show the scores PC2 against PC1, PC3 against PC1 and PC4 against PC1, respectively and Fig. 4(a), (b) and (c) the respective loadings.

The score plot of Fig. 3(a) shows that the β-blockers are clustered together in the upper part of the plot. The amino-acids appear together in the lower part of the same figure.

An inspection of the scores and loadings plotted in Figs. 3 and 4, respectively, shows that the first PC no longer reflects the size differences between the compounds investigated. After normalization of the mass spectral data, PC1 describes the same features as the second PC for the raw mass spectral data. The same holds for PC2 and PC3, that explain the same characteristic information as described by PC3 and PC4, respectively, for the raw spectral data. PC4 represents the contrast between basepeak at *m*/*q* 30 (variable 15, $CH_2=NH_2^+$) and most prominent peak at *m*/*q* 44 (variable 29).

Since the size effect of PC1 disappears after normalization of the mass spectral data, it seems better to work with normalized spectral data instead of raw spectral data.

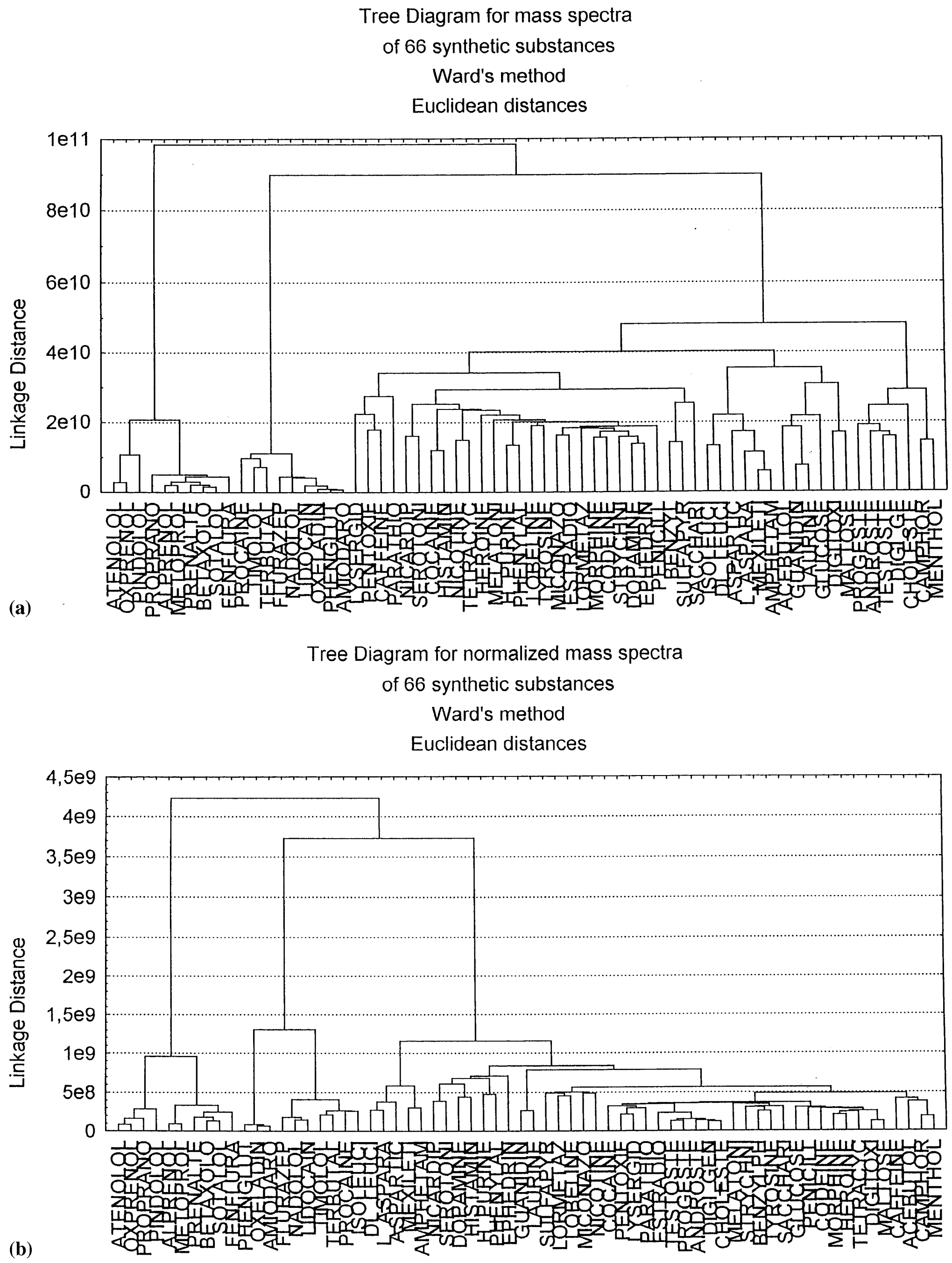

Fig. 10. (a) Ward's hierarchical classification, based on raw mass spectral data. (b) Ward's hierarchical clustering, based on normalized mass spectral data. (c) Ward's hierarchical clustering, based on 2D Daylight structural fingerprints. (d) Ward's hierarchical clustering obtained after applying a logarithmic transformation to the raw mass spectral data. (e) Ward's hierarchical clustering obtained after applying a logarithmic transformation to the normalized mass spectral data.

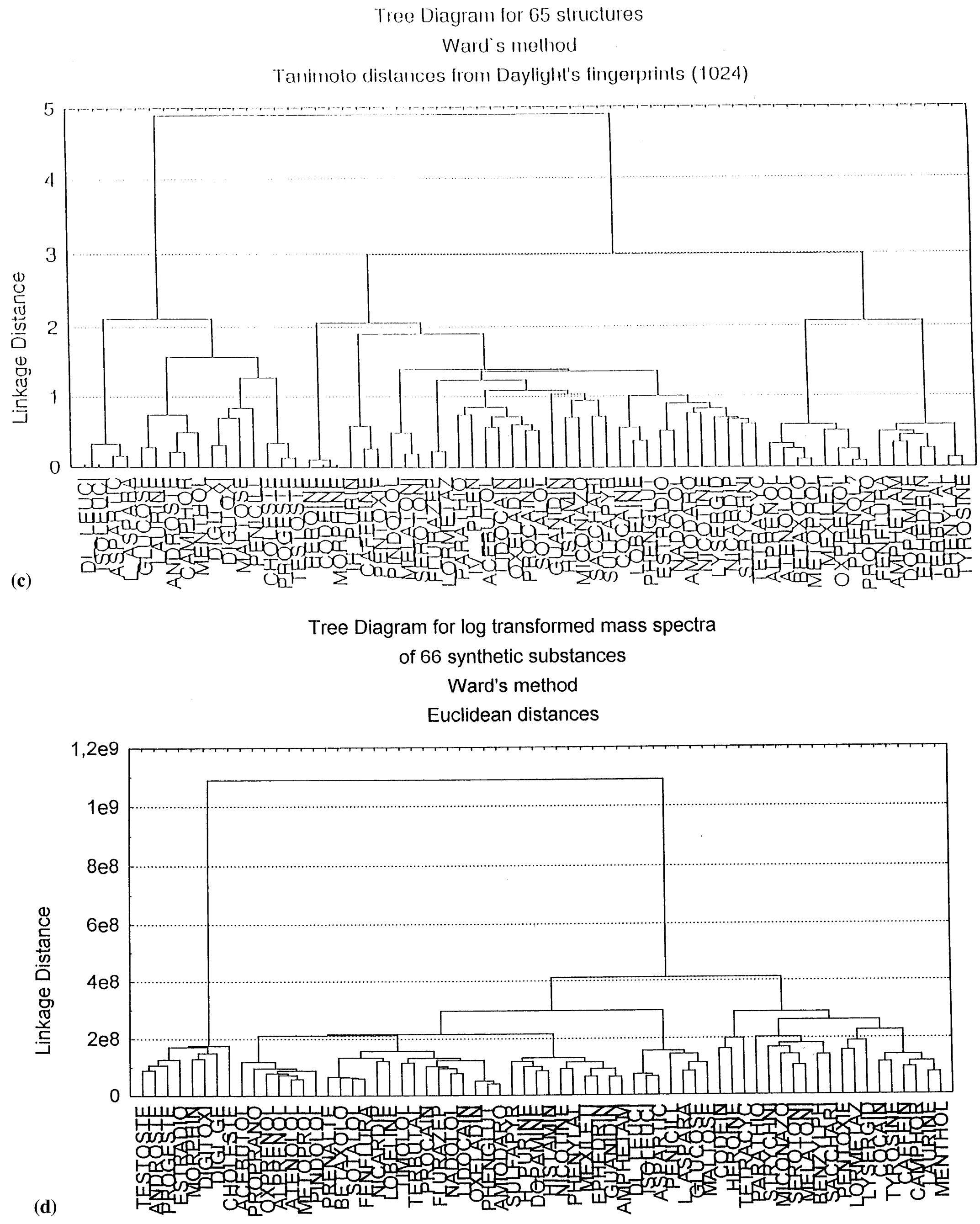

Fig. 10. (*Continued*)

To prevent that the large values will have too much influence on the classification and to enhance the effect of the smaller ones, a logarithmic transformation was applied on the (raw and normalized) mass spectral data matrix. In the transformed space, the variables with a comparable coefficient of variation will all have equal importance.

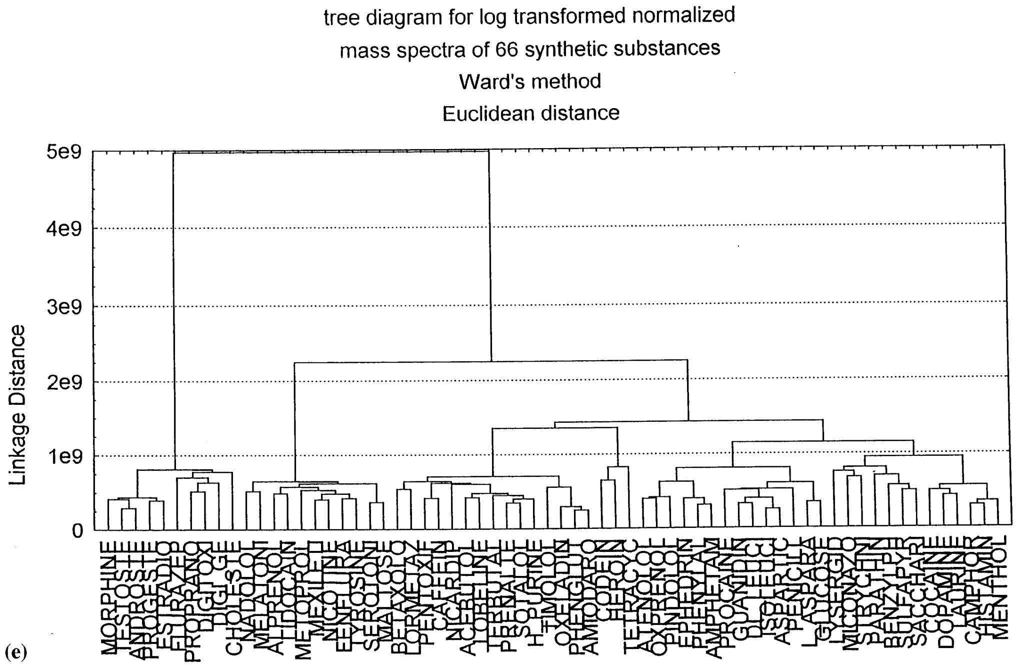

Fig. 10. (*Continued*)

The raw mass spectral data matrix, after log transformation, was subsequently analyzed with PCA. The PCA yielded a three-component model that explained 61.10% (48.32%, 8.60%, 4.18%) of the total variance. Figs. 5 and 6 (a) and (b) show the score- and loading plots of the analysis.

An examination of the score plot in Fig. 5(a). shows that all steroids are positioned in the lower right part of the plot. The amino-acids are grouped together in the upper left part of the score plot. Both sugars, maltose and glucose, lie closely clustered at the top of the plot and the β-blockers appear together in the upper half of the same score plot. The group of alkaloids (codeine, heroine, morphine, lysergide) is situated in the central region. Camphor and menthol, as well as serotonin and melatonin, are located near each other in the central part of the plot.

Looking at the score plots and loading plots (Figs. 5 and 6) shows that the first PC is related to the degree of fragmentation of the compounds investigated. Compounds with many fragment ions are clustered together to the far right in Fig. 5(a), whereas compounds with few fragment ions appear in the left part of the same figure. The second PC shows the contrast between substances that are primarily characterized by fragment ions of high mass and substances that are characterized by fragment peaks at low m/q-values. This is also seen in Fig. 6(a) since all variables corresponding to high m/q-values lie at the bottom of Fig. 6(a), whereas the variables related to low m/q-values appear at the top of the plot. The third PC only explains about 4.18% of the total variance. The interpretation of this component is difficult since many variables seem to be important for PC3. This is seen in the loading plot of Fig. 6(b). Correspondingly, PC3 discriminates compounds that have characteristic high intensity peaks at m/q 63 (variable 48), 51 (variable 36), 78 (variable 63) or 91 (variable 76) from compounds with specific high intensity peaks at m/q 56 (variable 41), 57 (variable 42), 70 (variable 55).

The PCA-analysis carried out on the normalized mass spectral data, after log transformation, resulted in a three-component model explaining 54.32% of the total variance, the individual PC's describing 41.01%, 8.87% and 4.44% of the vari-

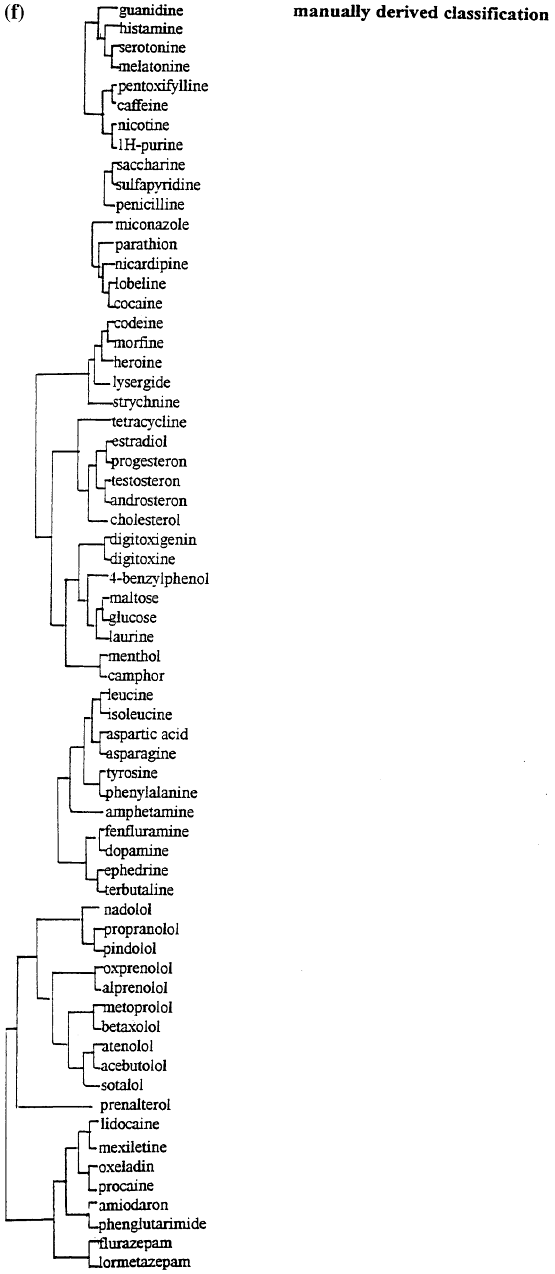

Fig. 10. (*Continued*)

ance, respectively. Fig. 7(a) and (b) shows the score plot of PC2 against PC1 and PC3 against PC1, respectively. The corresponding loadings are plotted in Fig. 8(a) and (b), respectively.

The score plot from Fig. 7(a) shows that all steroids are closely clustered in the lower right corner of the plot.

An examination of the scores and loadings plotted in Figs. 7 and 8, respectively, shows that the first PC describes the overall size of fragmen-

Table 3
1: Comparison with two largest clusters of the respective clusterings; 2: comparison with four largest clusters of the clusterings, based on mass spectra, and with six largest clusters of the clustering, based on Daylight fingerprints

	Manually derived/mass spectra	Manually derived/normalized mass spectra	Manually derived/Daylight fingerprint
1	0.4089	0.4089	0.4247
2	0.3638	0.4125	0.4272

tation of the compounds investigated since substances characterized by a lot of fragment ions are situated together in the right part of Fig. 7(a), such as, for instance, compounds No.9, 19, 32, 33, 34, 26, 58. Many variables enter into PC2 and PC3 so that it is not possible to regard any one of them as being mainly responsible for the formation of, respectively, the second and third PC. PC2 primarily reflects differences in the fragmentation patterns. Compounds that are characterized by a lot of high intensity fragment peaks at low *m*/*q*-values appear in the lower part of Fig. 7(a), whereas compounds with few fragment ions of high mass appear in the upper half of the same figure. The third PC discriminates substances with most intense peak at *m*/*q* 60 (variable 45), 72 (variable 57), 73 (variable 58) or 86 (variable 86) from the rest. This is seen in Fig. 8(b) where these variables are in the top region.

3.2. PCA of the 2D structural data

PCA was also carried out, using the 2D structural fingerprints. The result was a PC-model with four principal components. These components described in total 61.68% of the variance in the data set. The separate components described 43.97, 7.87, 5.35 and 4.49% of the variance in the data, respectively. Fig. 9(a) shows the score plot of PC1 against PC2. The third and fourth components are visualized in the same way in Fig. 9(b) and (c), respectively. The numbering of the compounds is reported in Table 2.

An inspection of the score plot of Fig. 9(b) shows that some small groups of similar substances can be found. The alkaloids morphine, codeine and heroine are situated together at the bottom edge of the figure. The steroids (testosteron, progesteron, androsteron, cholesterol, digitoxin, digitoxigenin) are clustered together to the upper right, as well as some amino-acids (L-asparagin, isoleucine, aspartic acid). The β-blockers appear in the central region of the same plot.

The first PC is again a size component which reflects the differences in molecular structure between the compounds investigated. Going from left to right, the order of complexity goes from linear and small size (e.g. guanidine, No. 42) to larger molecular structures, containing more complex aromatic groups, such as, for instance compounds No. 2, 6, 20, 52, 41. A fingerprint describes the overall size and chemical complexity of a molecule, since the hashed fragments encode all unique linear, branched, and cyclic fragments, including overlapping fragments. If a particular fragment is present in a molecule, then a corresponding bit is set to 1 in the bit string. Typically, small molecules set few bits to 1, and, as molecules get larger and more complicated, more bits are 1, so that the count of ones is higher.

PC2 separates alkaloid compounds with typical ‘morphine’ structure (compounds No. 20, 41, 52), corresponding to a more circular shape of ring structures all linked together, and other aromatic substances, with more linearly connected ring structures, like, for example compounds No. 50,

Table 4
1: Comparison with two largest clusters of the respective clusterings; 2: comparison with three largest clusters of the respective clusterings; 3: comparison with five largest clusters of the respective clusterings

	Manually derived/ log mass spectra	Manually derived/log normalized mass spectra
1	0.4210	0.3913
2	0.4547	0.3188
3	0.4617	0.2928

Table 5

1: Comparison of two largest clusters; 2: comparison of three largest clusters; 3: comparison of six largest clusters of the clustering, based on Daylight fingerprint with four largest clusters of the clustering, based on mass spectra; 4: comparison of four largest clusters

	Mass spectra/Daylight fingerprint	Normalized mass spectra/Daylight fingerprint	Mass spectra/normalized mass spectra
1	0.6472	0.6472	1
2	0.5040	0.4966	0.9761
3	0.4476	0.4206	
4			0.8361

51, 54, 12, from substances that are for the most part aliphatic. PC3 discriminates those compounds with a 'corticoid' structure from the morphine-derivatives. This is seen in Fig. 9(b), since the steroids appear in the upper half of the figure, whereas the alkaloids appear in the lower part. PC4 clusters all the β-blockers in the upper part of Fig. 9(c), while substances with N-containing aromatic 5- and 6-ring structures are situated more in its lower part, for instance, lormetazepam and flurazepam, are closely grouped in the bottom region, as well as melatonin and serotonin.

3.3. Qualitative comparison of hierarchical Ward's classifications

Hierarchical clustering methods produce classifications in which small clusters of very similar objects are nested within larger clusters containing more diverse structures [4]. The resulting classifications are, to a certain extent, in accordance with this. The hierarchical classifications, based on Ward's method and on raw mass spectral data and normalized mass spectral data, are shown in Fig. 10(a) and (b), respectively. The Ward's clustering, based on 2D Daylight structural fingerprints is shown in Fig. 10(c). Obviously, the variables used to describe the objects also have a great influence on how the objects will be classified. In order to determine the differences, the composition of the clusters of all three Ward's hierarchical clusterings has been mutually compared.

An investigation of the classification results shows that in all three clusterings, three large clusters, each consisting of many small subclusters are formed. The group of β-blockers is found as such, in one big cluster, in all three clusterings. Both other large clusters are more heterogeneous from the point of view of chemical classes, but consist each of smaller, more homogeneous subgroups, containing similar compounds. This holds for camphor and menthol, the amino-acids (asparagine, leucine and isoleucine), codeine and morphine, maltose and digitoxine, which are similar and closely clustered.

In both clusterings, based on mass spectra, as well as in the clustering, based on Daylight fingerprints, most of the corticoids (progesteron, androsteron, testosteron, digitoxigenin, cholesterol) are found in one subcluster. Flurazepam and lormetazepam, as well as melatonine and serotonine are contained in one subgroup in the classification, based on structural fingerprints.

The Ward's hierarchical classifications obtained after applying a logarithmic transformation to the (raw and normalized) mass spectral data are shown in Fig. 10(d) and (e), respectively. By examination of the clustering results, it can be seen readily that in both clusterings, three large clusters, each con-

Table 6

1: Comparison of two largest clusters; 2: comparison of three largest clusters; 3: comparison of six largest clusters of the clustering, based on Daylight fingerprint with five largest clusters of the clustering, based on mass spectra after log transformation

	Log mass spectra/Daylight fingerprint	Log normalized mass spectra/Daylight fingerprint
1	0.7907	0.7541
2	0.4431	0.4876
3	0.4136	0.3176

sisting of many small subclusters are formed. Clearly separated structural groups are present in the tree, for example all corticoids are in one big cluster in both clusterings. The composition of both other large clusters of chemical structures is considerably more heterogeneous, however with some small subgroups of very similar compounds. An example of this kind is the small cluster of amino-acids (leucine, isoleucine, aspartic acid, asparagine) or camphor and menthol, codeine and heroine. Another example is given by serotonin and melatonin as well as glucose and maltose that are linked together in the clustering, based on log transformed raw mass spectral data. Also, all β-blockers are located near each other in one smaller subcluster in this respective Ward's classification. However, in the clustering, based on log transformed normalized mass spectra, they appear more dispersed over the tree-structure.

In conclusion, it seems that the clustering with mass spectra is equally good as that obtained with Daylight structural fingerprints and that normalization of the mass spectral data does not have a major influence on the clustering result. At first sight therefore, it seems that one does not lose a lot of information using mass spectral data instead of structural data. Finally, it can be said that the clustering of the log transformed mass spectral data looks somehow different than the clustering of the original mass spectral data. However, the results are not better or worse than the classifications based on the original mass spectral data.

3.4. Quantitative comparison

A more quantitative comparison is possible using the measure of Wallace as similarity measure. It was performed between the different Ward's clusterings mutually and a classification, based on expert judgement, according to known structure and pharmacological activities of the set of 66 synthetic substances.

3.4.1. Comparison of Ward's clustering and the classification, based on expert judgement

The different cluster solutions were evaluated by comparing them with the expert's classification of the same data set, shown in Fig. 10(f). This classification consists of six clusters. It should be noted that, due to the choice of substances, some of which are relatively little related to others, other classifications might be proposed by other experts.

The stability of a cluster solution is probably different for different numbers of clusters, but, as already mentioned (2.3), the interpretation of the resulting partitioning depends on the characteristics of the chosen measure. It is known that the greater the number of clusters, the more Wallace's measure tends to yield a smaller similarity. However, whether a measure of similarity of, for example, 0.8 for two clusterings is good or not cannot be answered with this one single value. It can only be answered by user evaluation. The results of this comparative study are presented in Tables 3 and 4.

From a comparative study between the expert's classification and the different computer-assisted Ward's hierarchical clusterings, no major distinction can be made between the different comparisons, so that we can conclude that a classification, based on mass spectra compares well with that based on 2D Daylight fingerprints. Only a slight difference appears between the clusterings, based on raw and normalized mass spectral data, with the latter producing better results. However, the clustering of the log transformed mass spectral data compares most with the expert's classification and the Ward's hierarchical classification, based on log transformed normalized mass spectral data worst.

Therefore, mass spectral classification seems to be of similar quality as classification, based on structural fingerprints, and it seems that not much information is lost using analytical characteristics instead of structure for characterizing similarity. And, log transformation does not seem necessary for good clustering of mass spectral data.

The numerical results for the comparison of the different Ward's hierarchical clusterings between them are reported in Tables 5 and 6.

Comparing the two largest clusters of each classification, both clusterings, based on raw and normalized mass spectra are nearly identical. Also, their similarity to the Ward's hierarchical classification, based on 2D structural Daylight fingerprints is almost the same. The same conclusion can be made when comparing the three largest clusters or the six largest clusters of the clustering, based on 2D structural fingerprints with the four largest clusters of the clustering, based on (raw and normalized) mass spectra without log transformation. Generally speaking, after log transformation, the respective Ward's hierarchical classifications seem to compare less with the clustering, based on 2D Daylight fingerprints.

4. Conclusion

This study illustrates the use of clustering techniques to analyse whether experimental mass spectral parameters can be used for assessing similarity/diversity of chemical compounds. This was done by comparing a Ward's classification, based on the structural fingerprints, with a Ward's classification, based on the mass spectral data of the same compounds, and by validating both against an expert's classification of the same data set.

From our results, it seems that both structural parameters (using Daylight fingerprints) and experimental mass spectral parameters are able to group compounds into structural similar classes, so that it seems no information is lost, using mass spectra instead of fragments based 2D fingerprints. Also, normalization of the mass spectral data seems to have a (slight) positive effect on the clustering result and a logarithmic transformation does not seem necessary for good clustering.

Briefly we may conclude that there is a reasonable chance that mass spectrometry, in combination with other spectroscopic and chromatographic techniques, can give good results for assessing similarity/diversity to a library of natural products, the structure of which is not known

Acknowledgements

We thank 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.

References

[1] P.M. Dean (Ed.), Molecular Similarity in Drug Design, Blackie Academic Professional, London, 1995.

[2] D.R. Flower, On the properties of bit string-based measures of chemical similarity, J. Chem. Inf. Comput. Sci. 38 (1998) 379–386.

[3] V.S. Rose, E. Rahr, B.D. Hudson, The use of procrustes analysis to compare different property sets for the characterization of a diverse set of compounds, Quant. Struct.-Act. Relat. 13 (1994) 152–158.

[4] 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.

[5] C.A. James, D. Weininger, Daylight Theory Manual, Daylight Chemical Information Systems, 3951 Claremont St, Irvine, California 92714, USA, 1993.

[6] 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.

[7] G.M. Downs, P. Willet, Similarity searching and clustering of chemical-structure databases using molecular property data, J. Chem. Inf. Sci. 34 (1994) 1094–1102.

[8] K. Baumann, J.T. Clerc, Computer-assisted IR spectra prediction-linked similarity searches for structures and spectra, Anal. Chim. Acta 348 (1997) 327–343.

[9] N.E. Shemetulskis, D. Weininger, C.J. Blankley, J.J. Yang, C. Humblet, Stigmata: an algorithm to determine structural commonalities in diverse datasets, J. Chem. Inf. Comput. Sci. 36 (1996) 862–871.

[10] 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.

[11] W.A. Hardcastle, Towards a strategy for structure elucidation, Spectrosc. Eur. 10 (1998) 1.

[12] D.L. Massart, L. Kaufman, The Interpretation of Analytical Chemical Data by Use of Cluster Analysis, Wiley, New York, 1983.

[13] W. Vogt, D. Nagel, H. Sator, Cluster Analysis in Clinical Chemistry: A Model, Wiley, New York, 1987.

[14] 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, in: Handbook of Chemometrics and Qualimetrics: Part A-B, Elsevier, Amsterdam, 1997.

[15] L. Eriksson, A strategy for ranking environmentally occurring chemicals, Chemom. Intell. Labor. Syst. 5 (1989) 169–186.

[16] E.J. Martin, J.M. Blaney, M.A. Siani, D.C. Spellmeyer, A.K. Wong, W.H. Moos, Measuring diversity: experimental design of combinatorial libraries for drug discovery, J. Med. Chem. 38 (1995) 1431–1436.