ELSEVIER

Journal of Chromatography A, 897 (2000) 23–36

JOURNAL OF CHROMATOGRAPHY A

www.elsevier.com/locate/chroma

Exploratory chemometric analysis of the classification of pharmaceutical substances based on chromatographic data

A. Detroyer, V. Schoonjans, F. Questier, Y. Vander Heyden, A.P. Borosy, Q. Guo, D.L. Massart*

ChemoAC, Department of Pharmaceutical and Biomedical Analysis, Pharmaceutical Institute, Vrije Universiteit Brussel, Laarbeeklaan 103, B-1090 Brussels, Belgium

Received 11 January 2000; received in revised form 20 July 2000; accepted 27 July 2000

Abstract

A chemometric study has been conducted on a published data set consisting of the retention times of 83 substances, from five pharmacological families, on eight HPLC systems. Principal component analysis, clustering and sequential projection pursuit were applied. In this way it was investigated to what extent the combination of chromatography and chemometrics allows one to make conclusions about pharmacological activities of (candidate) drugs and what the contribution is of the different HPLC systems considered. © 2000 Elsevier Science B.V. All rights reserved.

Keywords: Quantitative structure–activity relationship; Principal component analysis; Clustering; Sequential projection pursuit; Chemometrics

1. Introduction

According to Valko [1] and several other authors [2] the chromatographic retention as well as the biological activity of a molecule are connected to its chemical structure. One can therefore hope to establish a relationship between chromatographic retention and biological activity. This has already been successfully done for instance by relating the retention on C_{18} stationary phases to the hydrophobicity parameter (log P), which plays a role in many quantitative structure–activity relationships (QSARs) [3].

Recently several new stationary phases or chromatographic systems have been proposed that may help to predict biological activity. They include amongst others the immobilized artificial membrane (IAM) stationary phase [4] consisting of cell membrane phospholipids, and micellar liquid chromatography [5,6], where the properties of the passage over a cell membrane are mimicked by using as the mobile phase aqueous solutions of surfactants at concentrations above the critical micellar concentration, thus creating two phases with different polarities.

Biological activity is a very complex matter, determined by many variables; i.e., it is multivariate by nature. Consequently one should try to include more than one chromatographic system when attempting to relate chromatographic and biological data. Such an interesting investigation was conducted by Nasal et al. [7]. They studied, on eight chromato-

*Corresponding author. Tel.: +32-2-4774-734; fax: +32-2-4774-735.

E-mail address: fabi@vub.vub.ac.be (D.L. Massart).

PII: S0021-9673(00)00803-7

graphic systems (CSs), 83 drugs belonging to several families according to an established pharmacological classification. The CSs were HPLC systems using several of the latest stationary phases at different pH values of the mobile phase. After applying principal component analysis (PCA) they concluded that the obtained logarithms of the retention parameters (log k) allow classifying of the substances according to their pharmacological properties. This is not surprising since multivariate statistical methods (like PCA) have also been successfully applied for characterising similarity/diversity of compounds given knowledge of the chemical structure [8].

Next to the chromatographic retention one of the first experimental data available about a (new) molecule is its molecular mass (M_r). The retention of candidate drugs is very often determined in a combinatorial synthesis context and, in such cases, one often applies liquid chromatography–mass spectrometry (LC–MS), with the electrospray ionization technique, where the MS essentially yields the molecular mass [9]. Being a descriptor of the molecule, the M_r together with the chromatographic data may allow better relationships with biological data and may thus lead to a better classification.

Nasal et al. [7] were only interested in establishing that the classification of the 83 drugs based on the retention data was possible. In this paper a more complete chemometric analysis of their data is performed in order to extract as much information as possible. This way for instance it is studied whether the data can give indications about the underlying physicochemical phenomena responsible for the retention on a given stationary phase. Furthermore it is evaluated if the results of all CSs are really needed to make the classification. Since in LC–MS the electrospray MS yields very little fragmentation and is used when one essentially wants to know the molecular mass, it is examined to what extent the M_r yields additional information to the chromatographic retention parameters for classification purposes.

2. Theory

2.1. Principal component analysis

Nasal's data set can be considered as a large $n \times m$ matrix where n represents the objects (the drugs) and m the variables (the CSs). With PCA the amount of original variables is reduced to a few latent variables or principal components (PCs) that still represent the main information from the original data set. The first new variable (PC1) is chosen in the direction of the largest variance in the data. The second PC is defined in such a manner that it is orthogonal to the first one and it represents a maximum of variance that was not explained by PC1, etc. Mathematically each PC can be described as a linear combination of the original variables where the importance of each original variable is given by the so-called loading of that variable. This yields for each object values, called the scores, on each PC. With PCA two main types of plots are obtained, namely the score plots which give information about the objects, here the substances, and the loading plots representing the variables, in this case the CSs.

2.2. Cluster analysis

Cluster analysis is the collective name for several techniques that are able to partition objects or variables into different groups. Most used are hierarchical clustering methods. They produce a classification in such way that any small cluster of a partition is fully included in one of the bigger clusters of the consecutive partition. Graphically the hierarchy can be represented by a dendrogram.

Before one starts the partition of n objects or variables it is necessary to determine the similarity between all objects. Most of the time the Euclidean distance, which is a measure of the geometric distance in a multidimensional space, is determined for each pair of objects. When clustering the variables the correlation coefficient between variables is used more frequent.

In this article Ward's hierarchical agglomerative clustering, which is often considered to be the method best able to separate similar and dissimilar structures [10–12], is also applied. With this method the two clusters whose fusion gives the minimum increase in the total within groups error sum of squares is used at each stage.

2.3. Weighted holistic invariant molecular descriptors

Weighted holistic invariant molecular (WHIM) descriptors are three-dimensional molecular indices that contain information about size, shape and symmetry. The essential characteristic of the method is that a PCA is made of the three-dimensional space in which the atoms are situated [13]. The first PC describes the direction of the largest length of the molecule, the second PC the direction orthogonal to the first and the largest variation around PC1, etc. A quantitative measure is the eigenvalue associated to each PC. If for instance the eigenvalue of PC3 is small compared to the others, then it means that the molecule has a planar structure. PC1 and PC2 describe the main axes in the planar molecule, PC3 the thickness of that planar molecule, which will be small compared to the eigenvalues of PC1 and PC2 and to what would have been found for a more globular molecule [14].

2.4. Sequential projection pursuit

Like PCA, projection pursuit (PP) is a chemometric method that projects an original multivariate space onto a few latent variables. However the aim of PCA is to choose these new variables in such way that they represent the maximal variance in the data, while PP looks for the most ‘‘interesting’’ directions. This means that the latent variables, in this case the PP factors, have a direction that leads to a non-uniform distribution of the projected data. These factors then show the inhomogeneities present in the data. A way of measuring non-uniformity and thus an index for the ‘‘interesting’’ directions is obtained by optimising entropy [15,16].

Because PP searches for all PP factors together, it is computationally very intensive. That is why in this article the sequential projection pursuit (SPP) method as described by Guo et al. [17] is applied. Here the latent variables or SPP factors are sought sequentially in the order of their importance as measured by the entropy index. Consequently the first SPP factor is the one that describes the maximum entropy of the projected data, the second SPP is constructed in such way that it is orthogonal to the first and maximises the remaining entropy of the data, etc. Parallel to PCA it is also possible to make ‘‘score’’ and ‘‘loading’’ plots, representing, respectively, the objects and the variables.

2.5. Transformations

Often some type of simple transformation (or scaling) is applied to the original data before it is chemometrically analysed with methods like PCA, clustering or SPP. The rows (objects) as well as the columns (variables) can be transformed solely or both, one after the other. Among the many possibilities column-wise autoscaling and centering are the most used transformations [18]. Column-centering just gives a scale shift in the data matrix, because for each variable (here CS) a constant (the mean) is subtracted from each of its measurements (here log k values). With autoscaling this difference is divided by the standard deviation, giving rise to variables which are independent of the unit of measurement, and which have equal range and therefore importance.

3. Experimental

The chromatographic data consisted of log k values and were taken as such from Nasal et al. [7]. The studied chromatographic systems included: a chiral α_1-acid glycoprotein (AGP) column at pH 6.5 (CS1), an IAM column at pH 7.0 (CS2), a Suplex pKb-100 column at pH 2.5 (CS3), a Suplex pKb-100 column at pH 7.4 (CS4), a RP-Spheri column at pH 2.5 (CS5), a RP-Spheri column at pH 7.0 (CS6), an Aluspher RP-select B column at pH 7.3 (CS7) and a Unisphere PBD column at pH 11.7 (CS8). More details concerning the experimental part can be found in Ref. [7]. The ninth variable added to this data analysis was the molecular mass. The drugs, their pharmacological classification (families A–E), the retention data together with their M_r and other physicochemical properties are shown in Table 1.

The log P values were estimated by applying the on-line interactive LOGKOW program of the Environmental Science Center of Syracuse Research Corporation, Syracuse, NY, USA (http://esc.syrres.com/~esc1/kowint.htm). The WHIM descriptors were calculated from the Cartesian coordinates of optimised structures (Hyperchem 3.0 [19]) using the software of Todeschini, namely the WHIM-3D package [20]. The transformations applied on the data were executed in the Matlab 4.2c.1 program from the MathWorks, Natick, MA, USA. For the PCA a

Table 1
The pharmacological classification, the retention data, the molecular masses and other physicochemical properties of the 83 drugs examined in Ref. [7]

No.	Drug	Log *k* AGP	Log *k* IAM	Log k_W Suplex		Log k_W RP Spheri		Log k_W Alusphere,	Log k_W Unisphere,	Log *P*	M_r
				pH 2.5	pH 7.4	pH 2.5	pH 7.0	pH 7.3	pH 11.7		
		CS1	CS2	CS3	CS4	CS5	CS6	CS7	CS8		9
Family A: Psychotropics and inactive phenothiazines											
1.	Acetopromazine	1.767	1.061	1.382	3.233	3.062	2.319	3.606	2.934	4.241	326.5
2.	2-Acetylphenothiazine	1.988	1.197	2.857	3.904	2.9	2.655	2.803	3.065	3.5054	241.33
3.	Carbamazepine	0.846	0.392	1.539	2.356	1.229	2.365	1.455	0.926	2.2484	236.27
4.	Chlorpromazine	2.131	1.435	1.595	4.051	1.935	2.632	3.309	4.076	5.2049	318.86
5.	Chlorprothixene	2.206	1.533	1.597	4.642	2.244	2.417	4.44	4.235	5.1445	315.86
6.	Clomipramine	2.005	1.391	2.134	4.144	2.353	2.473	4.115	3.91	5.6536	314.86
7.	Desipramine	1.595	1.031	1.616	3.02	2.015	2.341	3.171	2.888	4.7979	266.39
8.	Ethopropazine	2.066	1.213	1.418	3.241	2.443	3.761	2.832	4.181	5.4691	312.5
9.	Fluphenazine	2.159	1.496	1.683	4.554	2.922	2.688	4.067	3.352	4.1345	437.52
10.	Imipramine	1.67	1.097	1.391	3.535	2.082	3.158	3.133	3.02	5.0091	280.41
11.	2-Methoxyphenothiazine	2.151	1.282	3.048	4.094	3.097	3.056	3.397	3.4	3.1228	229.32
12.	Perphenazine	2.283	1.393	1.635	4.305	2.997	3.092	3.256	3.07	3.816	403.97
13.	Phenothiazine	1.854	1.354	3.06	3.949	2.769	3.167	3.263	3.375	3.8248	199.3
14.	Prochlorperazine	2.614	1.726	1.452	4.878	1.843	2.421	4.395	3.523	4.7897	373.94
15.	Promazine	1.89	1.165	1.556	3.492	2.338	2.808	3.794	3.294	4.5604	284.4
16.	Propiomazine	2.105	1.234	1.576	4.02	2.536	2.748	3.958	3.497	4.6586	340.48
17.	Thioridazine	2.448	1.752	2.113	4.26	2.055	2.924	3.182	4.655	6.4486	370.6
18.	*cis*-Thiothixene	2.273	1.359	1.417	3.971	2.098	3.365	3.58	2.77	3.1392	443.63
19.	Trifluoperazine	2.388	1.82	1.778	4.948	1.792	2.644	5.022	3.632	5.1082	407.49
20.	2-(Trifluoromethyl) phenothiazine	2.543	1.815	3.569	5.354	2.227	3.255	4.418	4.804	4.7878	267.27
21.	Triflupromazine	1.976	1.514	1.96	4.409	2.533	2.638	3.79	4.117	5.5234	352.44
22.	Trimeprazine	1.934	1.209	1.472	3.488	2.426	2.174	3.681	3.508	4.978	298.44
Family B: Agonists and antagonists of α-adrenoreceptors											
23.	Cirazoline	1.082	0.94	0.826	1.374	0.693	1.934	1.948	1.583	3.2186	216.28
24.	Clonidine	0.847	0.41	0.08	1.138	0.201	1.164	1.163	1.283	1.888	230.1
25.	Detomidine	1.073	1.018	1.097	2.582	0.758	1.337	2.255	1.627	3.2915	186.26
26.	Doxazosin	1.798	1.983	1.524	3.874	1.876	3.204	2.694	2.823	2.0853	451.48
27.	Indoramin	1.454	1.594	1.442	3.218	1.298	2.373	2.649	2.299	3.6021	347.46
28.	Lofexidine	0.965	0.879	0.791	1.479	0.509	1.704	1.581	1.41	3.5816	259.13
29.	Medetomidine	1.169	1.192	1.17	2.876	1.099	1.631	2.463	2.516	4.5026	200.28
30.	Moxonidine	0.528	−0.067	−0.24	0.385	−0.03	0.942	0.586	−1.125	0.2383	241.68
31.	Naphazoline	1.092	0.895	0.781	1.297	0.678	2.031	1.706	1.476	3.5174	210.27
32.	Oxymetazoline	1.108	1.216	1.151	2.312	1.578	1.666	2.319	1.274	4.8653	260.37
33.	Phentolamine	1.264	1.34	1.436	1.97	1.289	2.165	2.386	−0.834	3.3558	281.35
34.	Prazosin	1.39	1.594	0.863	2.948	0.909	1.639	1.442	1.172	1.2843	382.42
35.	Terazosin	1.051	1.119	2.204	2.266	0.405	1.818	1.249	0.167	1.4671	387.44
36.	Tetryzoline	0.822	0.553	0.247	0.917	0.671	1.259	1.001	0.68	3.685	200.28
37.	Tiamenidine	0.808	0.434	0.068	0.834	0.308	1.67	1	−0.231	0.7942	215.7
38.	Tolazoline	0.586	0.155	−0.292	0.1	0.404	1.353	0.58	−0.063	2.3414	160.21
39.	UK-14 304	0.831	0.269	0.401	1.493	−0.27	0.887	0.892	0.178	−1.3045	292.16
40.	Xylometazoline	1.158	1.362	1.468	2.38	1.92	2.412	2.475	2.385	5.3455	244.37

Table 1. Continued

No.	Drug	Log *k* AGP	Log *k* IAM	Log k_W Suplex		Log k_W RP Spheri		Log k_W Alusphere,	Log k_W Unisphere,	Log *P*	M_r
				pH 2.5	pH 7.4	pH 2.5	pH 7.0	pH 7.3	pH 11.7		
		CS1	CS2	CS3	CS4	CS5	CS6	CS7	CS8		9
Family C: β-Adrenolytics											
41.	Acebutolol	0.676	0.602	1.297	1.426	0.466	2.237	1.044	0.351	1.1909	336.43
42.	Alprenolol	1.49	0.918	1.594	2.229	1.308	2.831	1.971	1.72	2.8145	249.34
43.	Atenolol	0.499	−0.146	0.136	−0.01	0.297	0.414	0.226	−1.048	−0.0259	266.34
44.	Betaxolol	0.838	0.994	1.238	2.248	1.121	2.813	2.056	1.772	2.9817	307.44
45.	Bisoprolol	0.694	0.646	0.576	1.737	0.857	1.94	1.28	0.094	1.8375	325.45
46.	Bupranolol	0.981	0.269	1.178	2.379	1.22	2.484	2.474	2.055	3.0667	271.79
47.	Carteolol	0.706	−0.146	1.201	0.754	0.057	1.396	0.709	0.228	1.4165	292.38
48.	Celiprolol	0.7	0.723	1.645	1.45	0.775	0.854	1.037	0.232	1.9283	379.5
49.	Cicloprolol	0.735	1.012	1.465	1.994	0.937	2.757	1.674	0.573	2.0977	323.43
50.	Dilevalol	1.106	1.272	1.566	2.486	1.15	2.134	2.641	−1.258	1.9973	328.41
51.	Esmolol	0.649	0.646	1.24	1.569	0.742	1.687	1.429	0.916	2.0004	295.38
52.	Metoprolol	0.564	0.434	0.93	1.247	0.456	1.948	1.098	−0.553	1.6943	267.38
53.	Nadolol	0.606	0.269	1.044	0.685	0.404	2.849	0.778	−0.637	1.1691	309.42
54.	Nifenalol	0.639	0.269	0.387	1.214	0.343	1.707	1.316	0.075	0.9918	224.26
55.	Oxprenolol	1.21	0.586	1.018	1.674	0.82	1.672	1.647	1.218	1.8313	265.34
56.	Pindolol	0.87	0.586	0.675	1.084	0.415	1.623	1.126	0.331	1.4832	248.32
57.	Practolol	0.509	−0.067	0.565	0.294	0.541	1.014	0.365	−0.627	0.5281	266.34
58.	Propanolol	1.612	1.34	1.234	2.61	1.211	2.895	2.707	2.038	2.5974	259.34
59.	Sotalol	0.516	−0.146	−0.281	0.088	−0.07	2.024	0.325	−1.602	0.3693	272.36
60.	Timolol	0.696	0.385	0.956	1.271	0.333	1.688	1.19	0.171	1.7504	316.42
Family D: Histamine H1-receptor antagonists											
61.	Antazoline	1.154	1.043	1.363	2.169	1.003	2.128	2.272	1.888	3.3795	265.35
62.	Astemizole	2.408	1.437	1.779	4.902	1.492	2.36	4.425	3.508	6.4307	458.58
63.	Chloropyramine	1.431	1.33	0.798	3.299	1.058	2.216	3.013	2.767	3.3737	289.82
64.	(+)-Chlorpheniramine	1.19	1.063	0.701	2.912	0.726	2.811	2.687	1.899	3.8189	273.8
65.	(±)-Chlorpheniramine	1.202	1.055	0.701	2.895	0.794	2.788	2.7	2.043	3.8189	273.8
66.	Cinnarizine	2.148	2.25	2.242	5.12	2.476	3.253	4.842	4.665	5.4405	368.5
67.	Dimethindene	1.382	1.194	0.308	2.921	0.894	2.052	2.585	2.24	4.98	292.41
68.	Diphenhydramine	1.14	1.006	1.531	2.692	0.775	1.83	2.47	2.112	3.1063	255.35
69.	Isothipendyl	1.58	1.21	1.431	3.089	1.233	2.497	2.666	2.535	3.9405	285.42
70.	Ketotifen	1.459	1.168	1.24	3.105	1.002	1.946	2.707	1.95	3.6417	309.43
71.	Mepyramine	1.113	0.935	0.332	2.573	0.999	2.103	2.27	2.049	2.8101	285.39
72.	Pheniramine	0.926	0.602	−0.031	2.068	0.663	1.585	1.585	1.275	3.1744	240.34
73.	Pizotifen	1.898	1.588	1.455	4.091	2.154	3.032	2.203	3.465	5.5141	295.4
74.	Promethazine	1.833	1.508	1.693	4.081	1.132	3.169	3.069	3.216	4.4869	284.41
75.	Tripelennnamine	1.066	0.887	0.116	2.558	0.894	2.093	2.136	1.807	2.7292	255.35
76.	Triprolidine	1.185	1.084	0.667	2.818	0.834	2.359	2.294	2.618	3.704	278.38
77.	Tymazoline	1.306	1.024	−0.091	2.595	1.051	2.111	2.447	2.012	3.8815	232.32
Family E: Histamine H2-receptor antagonists											
78.	Cimetidine	0.482	−0.271	0.373	1.593	−0.301	0.069	0.412	0.724	0.574	252.34
79.	Famotidine	0.731	−0.271	0.416	0.755	−0.267	0.184	0.875	0.193	−0.6544	337.4
80.	Metiamide	0.517	−0.301	0.217	1.249	0.447	0.676	0.705	0.044	0.5201	244.4
81.	Nizatidine	0.46	−0.368	−0.006	0.832	0.114	0.209	0.089	−0.569	−0.671	331.5
82.	Ranitidine	0.6	−0.016	0.301	1.136	0.125	0.335	0.779	1.779	0.2938	314.41
83.	Roxatidine acetate	0.773	0.359	1.349	1.579	0.312	1.145	0.794	1.154	2.2099	306.4

laboratory Matlab toolbox for multivariate calibration was utilised. Clustering was executed by applying the Statstica 5.X program of Statistica, Gaithersburg, MD, USA. For the SPP of the data a laboratory-designed genetic algorithm running in Matlab 4.2c.1 has been applied.

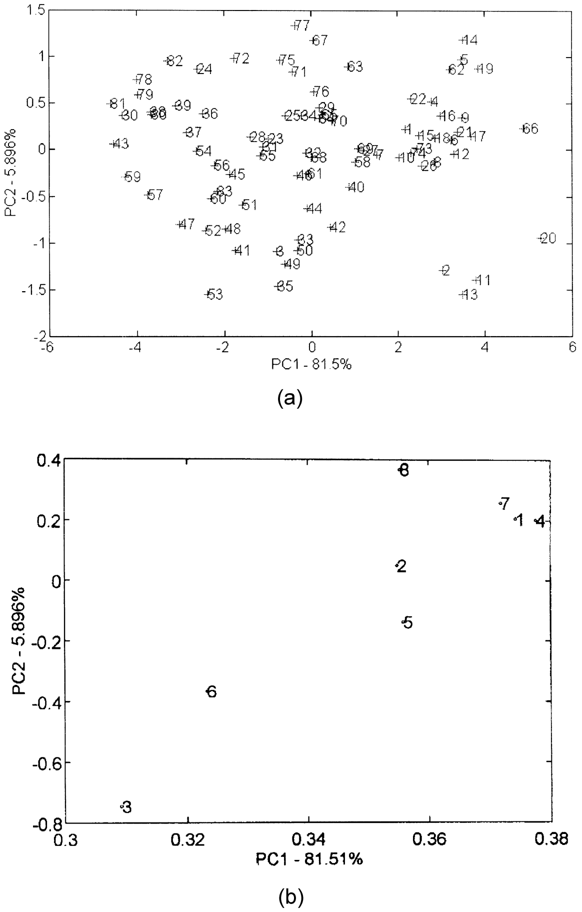

Fig. 1. (a) PC1–PC2 score plot of the autoscaled chromatographic data. The numbers represent the drugs 1–83 in Table 1. (b) PC1–PC2 loading plot of the autoscaled chromatographic data. The numbers represent CS1–CS8 in Table 1.

4. Results and discussion

4.1. Principal component analysis of the data

Nasal et al. only considered the score plot of the first and second PC, which was sufficient for their purpose, since it allowed them to show that five pharmacological groups can be partially discriminated. However, no reference to the loading plots was made ignoring possible conclusions about the role of the variables. Moreover PC3 and higher PCs were not investigated.

Although it was not mentioned in Ref. [7] which type of transformation was used on the original data, it must be autoscaling, since the PC1–PC2 score plot obtained by us after such a scaling procedure (Fig. 1a) is the same as the one shown by Nasal et al. [7] apart from the sign of the scores of PC2. The signs in PCA just indicate a direction, which is the result of an arbitrary choice depending on the used calculating program. The opposite direction can be chosen without it influencing the results drawn from score and loading plots [18].

Throughout these investigations autoscaling is applied. In this way, a scale effect due to an overall larger retention for all substances on one of the CSs is avoided [18].

4.1.1. Principal component analysis of the autoscaled chromatographic data

Fig. 1 shows both the score (a) and the loading (b) plot for PC1 and PC2. The score plot is the same as the one shown by Nasal et al. [7], the five different families (A–E) can be distinguished on it. Furthermore it is remarked that within family A several substances (Nos. 2, 11, 13, 20) have a somewhat deviating location and form a separate group. Although only 6% of the variance is explained by PC2 it is of importance to the group separation. As explained by Nasal et al. [7] the separation patterns themselves have a pharmacological resemblance.

On the loading plot the loadings of all the CSs along PC1 are very similar and positive. It should be remembered that the score of an object is the weighted sum of the original variables, with as weights their loadings. If, as is the case here, all the loadings are similar and have the same sign, then the scores of the substances along PC1 are more or less equal to a constant multiplied with the sum of the retention properties, as measured by the autoscaled log k on the eight CSs. Since the log k is often related to the log P [3,6], PC1 might possibly represent a hydrophobicity axis. As shown further, this is indeed the case. The largest contrast in loadings along PC2 is between CS3 (Suplex, pH 2.5) and CS8 (Unisphere, pH 11.7). Looking at the pH values of the CSs, it seems probable that along this axis a difference in acid/base behaviour of the substances is expressed. CS5, the other CS at low pH (RP Spheri, pH 2.5) has the same sign as CS3. The picture is mixed up however by CS6 (RP Spheri, pH 7.0) also having the same sign as CS3. Thus no simple interpretation for the separation along PC2 can be given. It should be noted that CS2 (the IAM column) has no influence at all on this PC, since it has a loading close to zero. In Fig. 2, which shows the loading plot of PC3 against PC2, PC3 essentially represents the contrast between CS3 and CS6, which were mixed up along PC2. Therefore PC2 is not a pure acid/base axis and PC2 and PC3 together are needed to express acid/base behaviour.

The main contrast in the loading plot for PC4 (Fig. 3) is between CS5 and CS2. It is not clear what this means, all the more so because PC4 explains only 3% of the variance in the data.

It is interesting to note that in all loading plots (PC4 included) the CS1, 4 and 7 are always found in each other's vicinity. This means that they give very correlated information. These findings are confirmed when calculating the correlation coefficients (r) between each of the CSs based on their log k (Table 2). It shows that CS1, 4 and 7 are indeed the most correlated stationary phases (r ±0.93). This becomes even more obvious in Fig. 4 where hierarchical single linkage clustering is applied on the CSs with as similarity measure $1-r$. In this way the CSs are classified according to their correlation with the other CSs. Indeed, there is a quite close cluster between CS1, CS4 and CS7 and to a somewhat lesser extent CS8. The most different from the others are CS3 and CS6. They are also the most different from each other ($r=0.632$). Small discrepancies between the dendrogram and PC1–PC2 loadings occur because the distances in the loading plot only relate to the variations accounted for by the first two components.

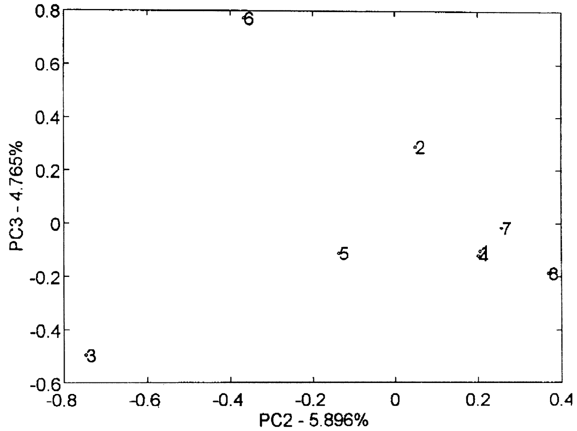

Fig. 2. PC2–PC3 loading plot of the autoscaled chromatographic data. The numbers represent CS1–CS8 in Table 1.

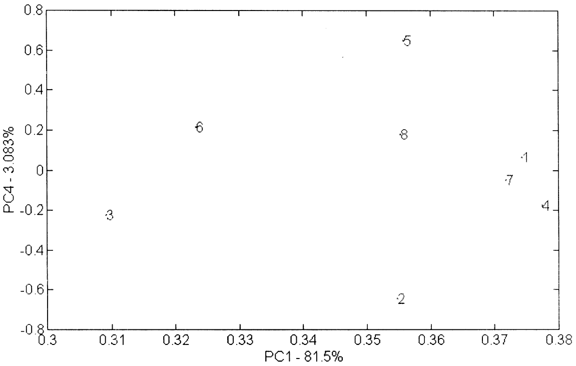

Fig. 3. PC1–PC4 loading plot of the autoscaled chromatographic data. The numbers represent CS1–CS8 in Table 1.

In the search for a physicochemical explanation for the observed relationships, the relationship between log P values and log k for each CS is investigated. In this way it should be possible to see whether hydrophobicity is a main factor for log k. The results are shown in Table 3. The highest correlations with log P are obtained for CS7 and 8 ($r=0.83$), the correlation coefficients for most other CSs are not much lower, except for CS3 ($r=0.54$) and CS6 ($r=0.68$). This implies that the CSs that resemble each other the most (CS1, 4 and 7), i.e., show the highest correlation between their log k values do so because they all describe mainly hydrophobicity. The worst correlation is obtained by CS3, which means that the log k on CS3 cannot be explained to the same extent by log P. This is also, to somewhat lesser extent, the case for CS6. Consequently some other factor(s) must be responsible.

The fact that CSs 1, 2, 4, 7 and 8 are highly correlated to the log P can be explained by the pH at which the measurements on these columns are performed. Knowing that the drugs are basic mole-

Table 2
The correlation (r) matrix between each of the CSs

r	CS1	CS2	CS3	CS4	CS5	CS6	CS7	CS8
CS1	1							
CS2	0.8399	1						
CS3	0.6971	0.6677	1					
CS4	0.9386	0.875	0.7223	1				
CS5	0.8594	0.7488	0.7343	0.8326	1			
CS6	0.7275	0.7669	0.632	0.7233	0.7435	1		
CS7	0.9228	0.8565	0.6662	0.9387	0.8383	0.7294	1	
CS8	0.8843	0.7732	0.6285	0.9004	0.7982	0.6637	0.8695	1

cules, which are non-dissociated at higher pH and knowing that the log P is determined for non-dissociated molecules it is logical that the CSs at high pH correlate better with the log P than the CSs at low pH. For CS3 the pH is low (2.5) and as a consequence the molecules are dissociated. It should be mentioned though that the low correlation found for CS3 might have another cause then dissociation, which is supported by the agreement between CS5 and CS6.

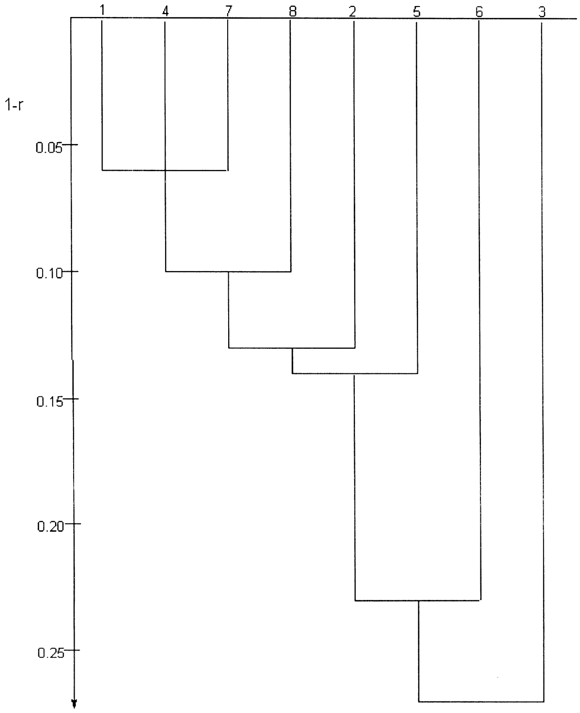

Fig. 4. Hierarchical single linkage clustering with similarity measure $1-r$ of the CSs 1–8.

Table 3
The correlation coefficients (r) between log P and log k for each CS

r	Log P
Log P	1
Log k CS1	0.7852
Log k CS2	0.7975
Log k CS3	0.5444
Log k CS4	0.7905
Log k CS5	0.751
Log k CS6	0.6875
Log k CS7	0.8372
Log k CS8	0.8352

The relationship between the autoscaled scores along PC1 of the drugs and their log P values (Fig. 5) is linear ($r=0.8386$). This confirms that the separation of the pharmacological groups along PC1 is due to differences in hydrophobicity. Since the

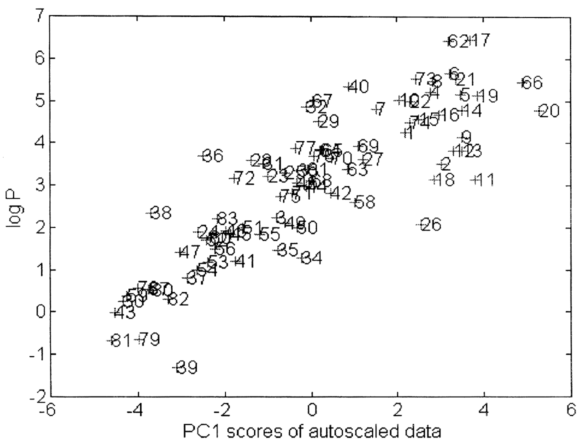

Fig. 5. The PC1 scores of the autoscaled data versus the log P. The numbers represent the drugs 1–83 in Table 1.

“true” hydrophobicity (log P) depends on the pK_a value of the analyte as well as on the pH of the CS, PC1 actually represents a combination of the “intrinsic” hydrophobicity and the acid/base properties of the drugs. However since all substances are basic, the acid/base properties cannot be very influential for PC1.

As a result the battery of CSs seems to yield essentially a plot of hydrophobicity (PC1) against (perhaps) acid/base characteristics (PC2–PC3). More subtle differences are apparently not observed. Certainly the AGP column (CS1) does not yield additional information compared to the RP column at pH 7.4 (CS4), although one would have hoped that this protein column would show more characteristic interactions. In fact the main information in the data can be reproduced by only two columns, CS8 and CS3. The sum of CS8 and CS3 represents the hydrophobicity while their difference (CS8−CS3) shows the acid/base contrast. However utilising three columns (CS4, CS8, CS3), plotting CS4 against CS8−CS3 (Fig. 6), has the advantage that CS4 represents conditions that are used very frequently by chromatographers. CS4 represents the hydrophobicity and the closely clustered CSs, while CS8−CS3 with their large pH difference show the acid/base contrast. These findings are confirmed by the high correlations that are found between the PC1 scores and the autoscaled log k of CS4 on the one hand (r=0.9642) and between the PC2 scores and the autoscaled log k of CS8 minus CS3 on the other hand (r=−0.8897).

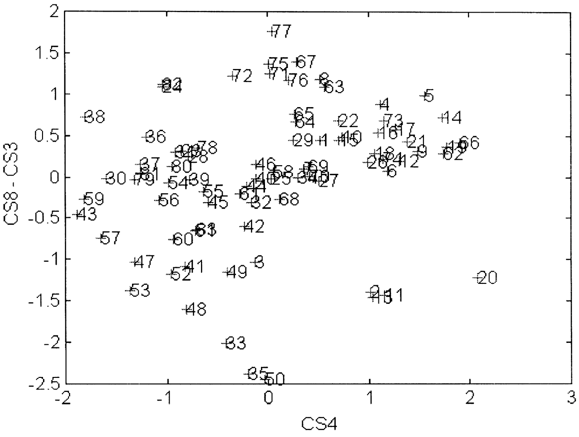

Fig. 6. The autoscaled log k of CS8 minus the autoscaled log k of CS3 versus the autoscaled log k of CS4. The numbers represent the drugs 1–83 in Table 1.

4.2. Cluster analysis of the autoscaled chromatographic data

Applying Ward’s clustering method, with the Euclidean distance as similarity measure, on the autoscaled data was much less successful than PCA meaning that the classification makes much less pharmacological sense, i.e., the big clusters do not contain all drugs from one and the same family. This is due to the fact that in clustering one works in the original variable space where the overriding influence is hydrophobicity, so that the clustering is merely based on distance along PC1. Since PC1 is not able to separate pharmacological classes by itself, the clustering is bad.

Clustering the first three PC’s scores on the other hand yields much better results (Fig. 7); a family can be assigned to each big cluster. The dendrogram is obtained by applying Ward’s method with Euclidean distances to the normalised scores of the substances. Thus the scores are not adjusted with the eigenvalues of the PCs and each PC has the same weight. In this way hydrophobicity is only one of the three variables.

The difficulty of this type of procedure is to decide how many PCs one should include. Criteria to decide on the number of significant PCs have been described in the literature [18]. However, different criteria yield different numbers. Since in the preceding PCA study the first three components were important, these have been investigated here too.

4.3. Principal component analysis of the autoscaled data including the molecular mass

The molecular mass has loadings (Fig. 8a) that differ appreciably from zero only on PC1 and on PC2. The loading on PC1 is positive as is the case for the chromatographic variables but it is less elevated. As a result the ranking of the substances along PC1 is nearly the same with or without M_r (Fig. 8b). The main effect is on PC2, which explains about 9.5% of the variance. The loading of M_r is positive and close to 1, the loadings of all the chromatographic variables are small, negative and

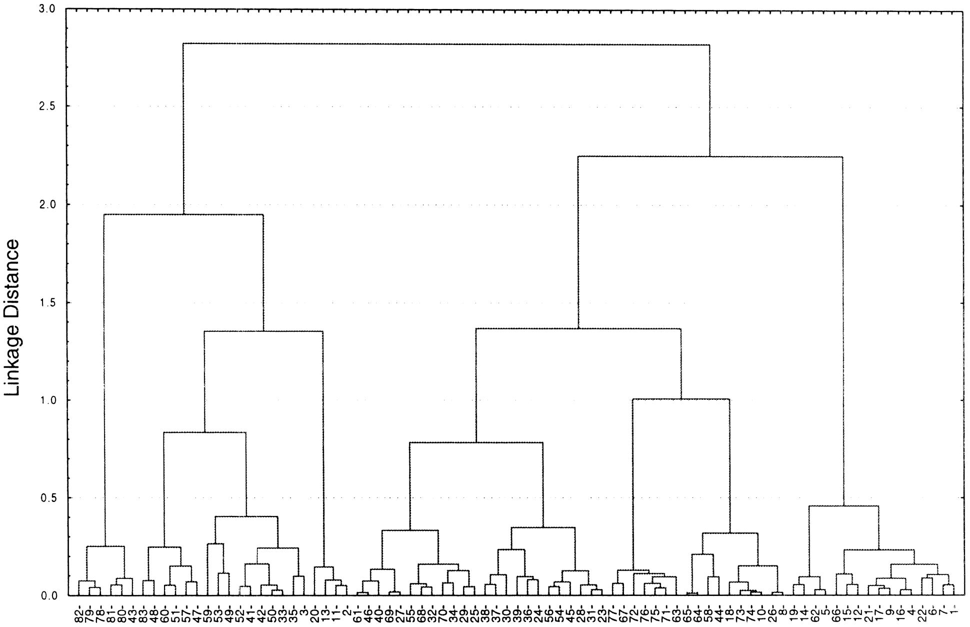

Fig. 7. Ward's hierarchical agglomerative clustering of the first three PC's normalised scores. The numbers represent the drugs 1–83 in Table 1.

similar to each other. It seems therefore that this axis is a nearly pure molecular mass axis. Moreover the loadings along PC3 of the CSs with M_r (Fig. 8c) are equal to the ones along PC2 without M_r (Fig. 2) apart from their sign (minus instead of plus). All this indicates that the information included in the molecular mass is little correlated to that given by the chromatographic data. Consequently it is concluded that the information from the molecular mass is different from that in the chromatographic variables. In that sense its addition to the chromatographic data is useful. However, it is also known that M_r is not extremely useful to describe diversity among substances [21], thus it does not improve the classification of the drugs (Fig. 8b).

4.4. Grouping in view of WHIM descriptors

The split of family A in two subgroups cannot be explained with log *P* and pH effects. In search of a physicochemical explanation for the split there is a need for some molecular descriptor to account for the observed phenomena. Thus the relationship between the chromatographic data and molecular descriptors as used in QSAR studies is investigated. There are many descriptors available, but eventually the WHIM descriptors as proposed by Todeschini and Gramatica [14] were chosen.

The computations show that the eigenvalue of PC3 is very small for drug Nos. 2, 11, 13 and 20 (family A1) and not for the other molecules of family A. Consequently the small outlying group of molecules has a very planar structure while the others have not (Fig. 9). In comparison to the others these planar drugs also show a stronger interaction (higher log *k* values) on CS3 and CS5 than is expected by their log *P* values (Fig. 10). This stronger interaction is responsible for their deviating behaviour as seen before on Fig. 1a and Fig. 6. On the other CSs this behaviour is not seen. Consequently the family A1's special retention properties should arise from a combination of the characteristic acidic environment of CS3 and CS5 and the planar structure of the molecules. At low pH all 83 molecules are ionized thus, considering hydrophobicity is the main partition force, they are not retained very strongly. The planar molecules however are retained more than the others. Possibly due to their shape they are well adsorbed on and/or folded between the hydrocarbon chains of the stationary phase. Removing the chro-

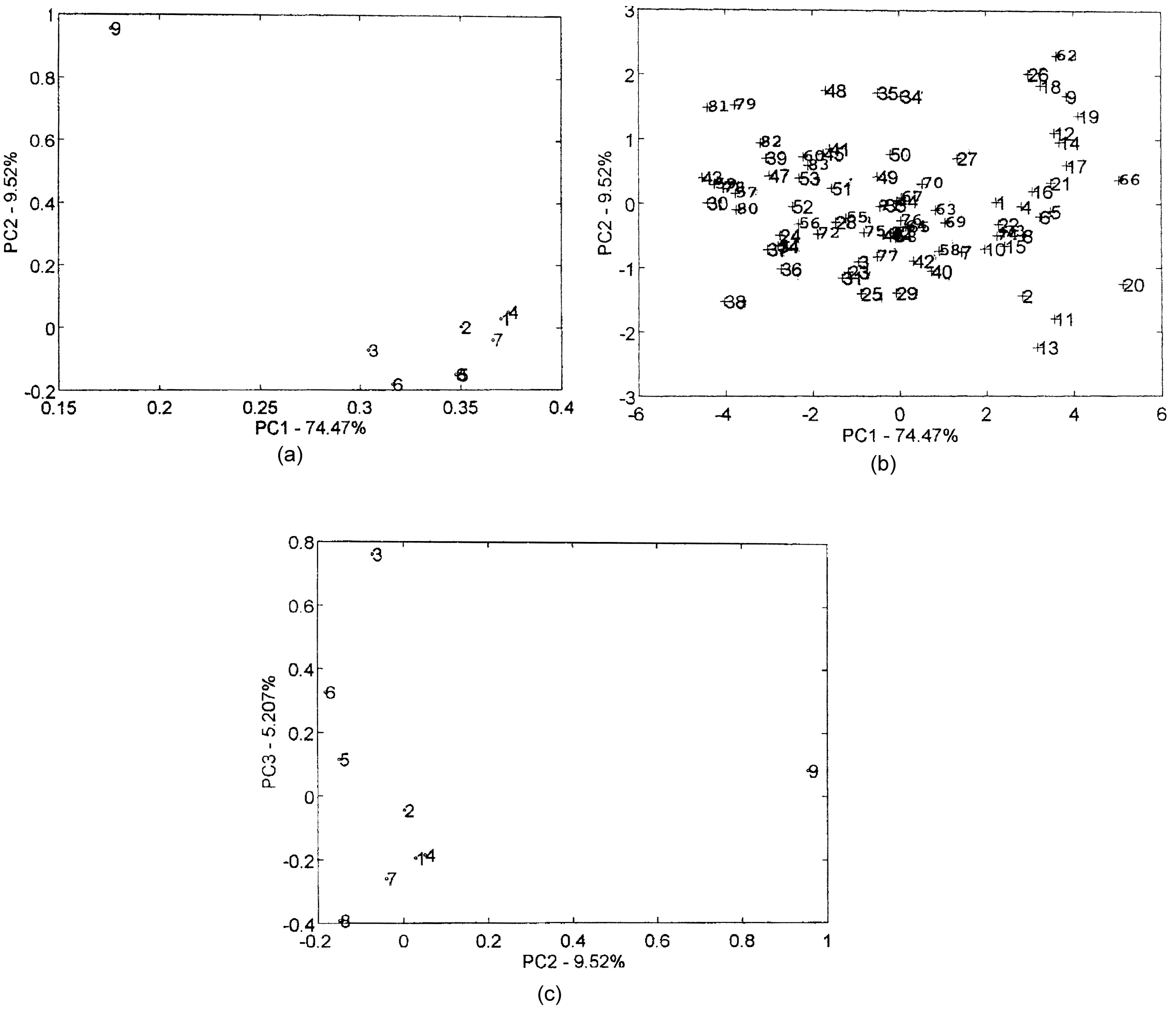

Fig. 8. (a) The PC1–PC2 loading plot of the autoscaled data including the molecular mass. The numbers represent CS1–CS8 and number 9 represents M_r in Table 1. (b) The PC1–PC2 score plot of the autoscaled data including the molecular mass. The numbers represent the drugs 1–83 in Table 1. (c) The PC2–PC3 loading plot of the autoscaled data including the molecular mass. The numbers 1–8 represent CS1–CS8 and number 9 represents M_r in Table 1.

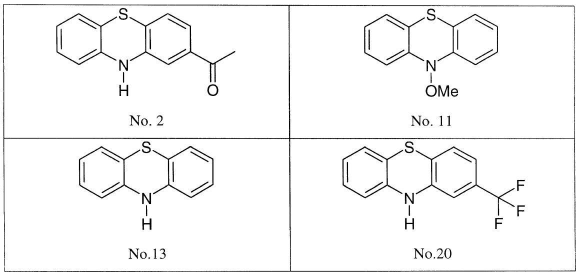

Fig. 9. Structures of the drug Nos. 2, 3, 11, 13 and 20 in Table 1.

matographic results of molecules 2, 11, 13 and 20 from the data improves, especially in the case of CS3 and CS5, the correlation with log *P* (Table 4). This demonstrates that the lower correlations of these CSs (Table 3) were due to some extent to the deviating retention mechanism of these molecules.

4.5. Sequential projection pursuit of the autoscaled data

On the SPP1–SPP2 "score" plot (Fig. 11) the

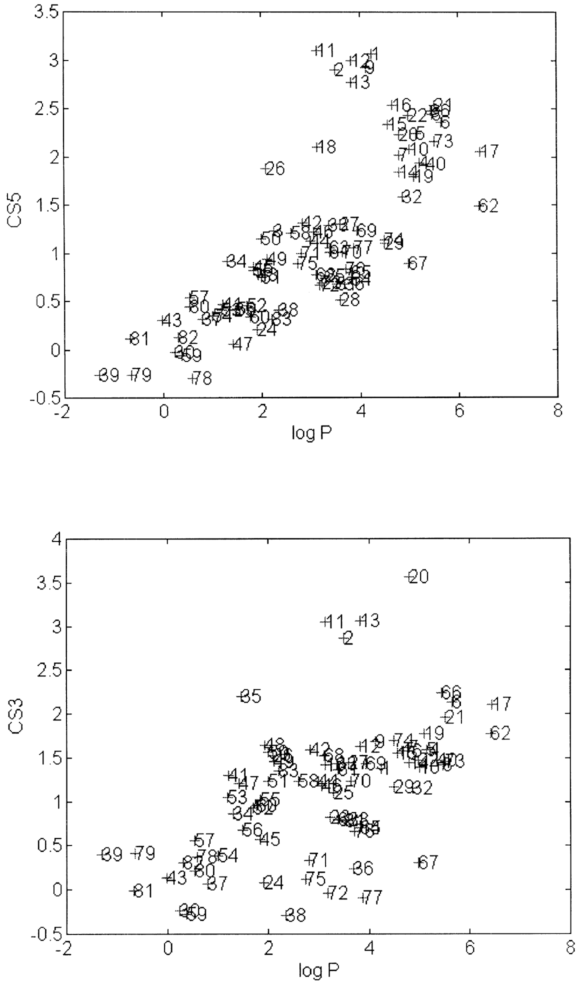

Fig. 10. The log P versus the log k of the drugs for CS3 and CS5. The numbers represent the drugs 1–83 in Table 1.

Table 4
The correlation coefficients (r) between log P and log k (except drug Nos. 2, 11, 13, 20) for each CS

r	Log P
Log P	1
Log k CS1	0.7983
Log k CS2	0.7960
Log k CS3	0.5998
Log k CS4	0.8009
Log k CS5	0.7912
Log k CS6	0.6895
Log k CS7	0.8414
Log k CS8	0.8437

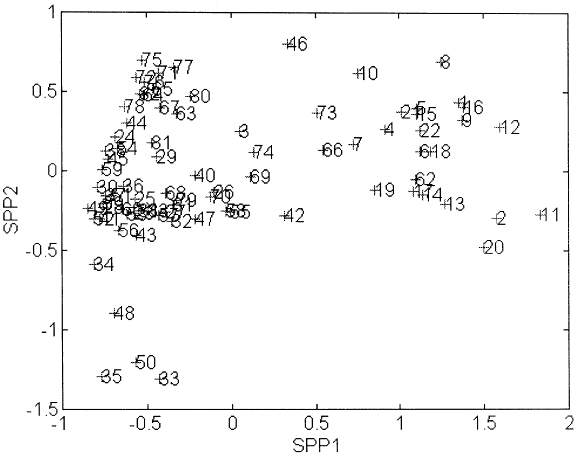

Fig. 11. SPP1–SPP2 ‘‘scoreplot’’ of the autoscaled data. The numbers represent the drugs 1–83 in Table 1.

psychotropic substances (family A) are better separated from the other substances along SPP1 then along PC1 (or any other PC). All columns have similar loadings along PC1 (Fig. 1b) and the psychotropics, which have generally higher log k values than the other substances, are separated from the rest but not completely (Fig. 1a). Looking at the corresponding SPP ‘‘loading’’ plot (Fig. 12) it seems that along SPP1 the distinction between CS1 (positive loadings) and CS2 (negative loadings) is made. SPP1 is therefore an axis which compares the autoscaled

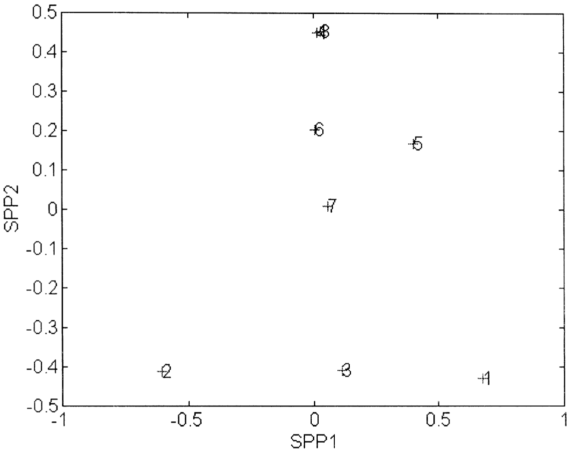

Fig. 12. SPP1–SPP2 ‘‘loadingplot’’ of the autoscaled data. The numbers represent CS1–CS8 in Table 1.

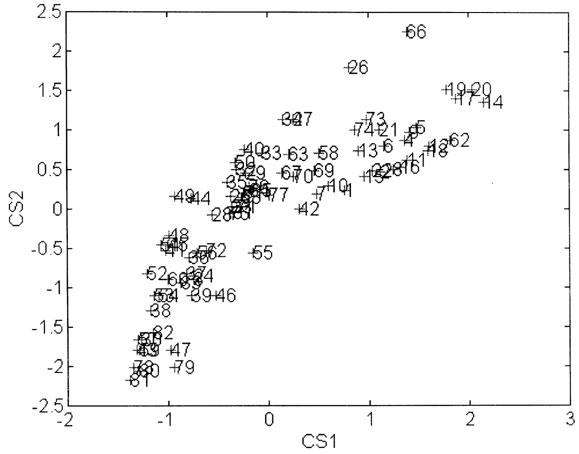

Fig. 13. The autoscaled results of CS1 versus CS2. The numbers represent the drugs 1–83 in Table 1.

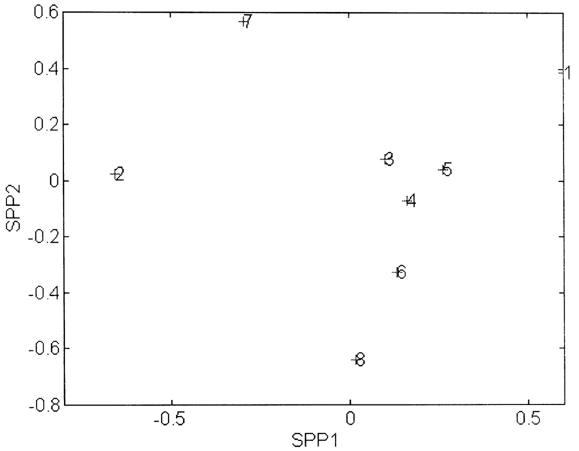

Fig. 15. SPP1–SPP2 “loadingplot” of the autoscaled data minus family A. The numbers represent CS1–CS8 in Table 1.

log *k* on CS1 to those on CS2. It can be verified that the ratio of the autoscaled values on CS1 to CS2 is larger for the psychotropics in comparison to the H1-receptor antagonists and the other substances showing overlap with the psychotropics along PC1 (Fig. 13). It is known that some psychtropics bind very well with the α-acid glycoprotein in blood [22] which may explain their high log *k* values on CS1.

Excluding the results from family A, SPP was repeated to investigate whether this would reveal less distinct inhomogeneities. On the “score” plot (Fig. 14) along SPP1 no new inhomogeneities are seen. Along SPP2 molecule Nos. 33 and 50 are clearly separated from the other substances. The “loading” plot (Fig. 15) indicates that both molecules are marked by their low values for CS8 as to CS7, which is confirmed in Fig. 16. It is found that this is not only the case for CS8 versus CS7, but also for CS8 versus all the other CSs. Generally, it can be concluded that SPP is able to find some contrasts more easily than PCA and that the two techniques complete each other.

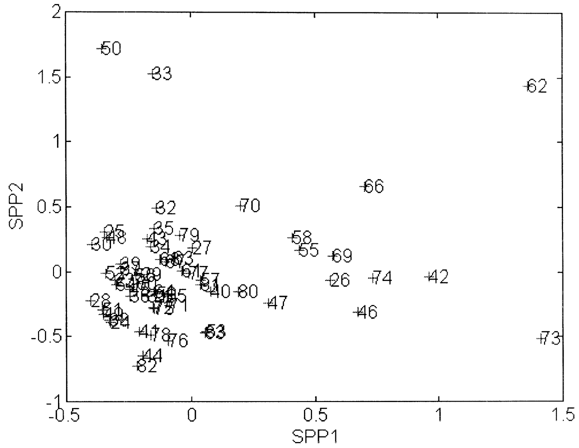

Fig. 14. SPP1–SPP2 “scoreplot” of the autoscaled data minus family A. The numbers represent the drugs 22–83 in Table 1.

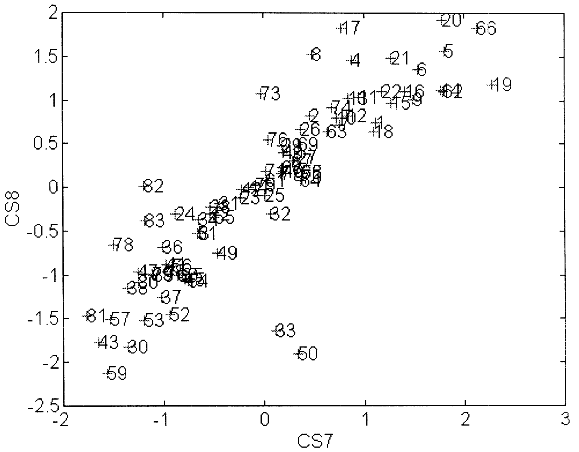

Fig. 16. The autoscaled results of CS7 versus CS8. The numbers represent the drugs 1–83 in Table 1.

5. Conclusion

The systematic chemometric analysis of chromatographic data with the use of complementary techniques such as PCA (with both score and loading plots), SPP, clustering and regression allows to uncover the information present in the data. It is possible for instance to conclude which combination of CSs seems to be the most useful. Here they are CS4, CS3 and CS8, and perhaps, CS2 (because of its contrast with CS1 for the psychotropics). However the number of substances investigated is limited and it may be that for other substances some of the other CSs also prove useful.

Other chemometric techniques might have been applied in this exploratory study. This is for instance the case for supervised pattern recognition methods such as soft independent modelling of class analogy (SIMCA) or linear discriminant analysis (LDA). We have chosen not to do so because the eventual intention of using a battery of CSs would be to classify new substances in an unsupervised fashion.

Acknowledgements

Investigation financed with a Specialisation Grant from the Institute for the Promotion of Innovation by Science and Technology in Flanders (IWT).

References

[1] K. Valko, Trends Anal. Chem. 6 (1987) 214.
[2] L. Bober, A. Nasal, A. Kuchta, R. Kaliszan, Acta Chromatogr. 8 (1998) 48.
[3] K. Valko, J. Liq. Chromatogr. 7 (1984) 1405.
[4] S. Ong, H. Liu, C. Pidgeon, J. Chromatogr. A 728 (1996) 113.
[5] E.D. Breyer, J.K. Strasters, M.G. Khaledi, Anal. Chem. 63 (1991) 828.
[6] M.C. Garcia Alvarez-Coque, J.R. Torres Lapasio, Trends Anal. Chem 18 (1999) 533.
[7] A. Nasal, A. Bucinski, L. Bober, R. Kaliszan, Int. J. Pharm. 159 (1997) 43.
[8] W. Werther, K. Varmuza, Fresenius J. Anal. Chem. 344 (1992) 223.
[9] J. Smedsgaard, J.C. Frisvad, J. Microbiol. Methods 25 (1996) 5.
[10] P.M. Dean (Ed.), Molecular Similarity in Drugs Design, Blackie Academic Professional, London, 1995.
[11] R.D. Brown, Y.C. Martin, J. Chem. Inf. Comput. Sci. 36 (1996) 572.
[12] G.M. Downs, P. Willet, J. Chem. Inf. Comput. Sci. 34 (1994) 1094.
[13] R. Todeschini, P. Gramatica, Perspect. Drug Design 9–10–11 (1998) 355.
[14] R. Todeschini, P. Gramatica, Quant. Struct.-Act. Relat. 16 (1997) 113.
[15] G.P. Nason, Entropy in multivariate analysis: projection pursuit, UK Chemometrics Discussion Group Newsletter, Issue 18, November 1991.
[16] M.C. Jones, R. Sibson, J. Royal Statist. Soc. A 150 (1987) 1.
[17] Q. Guo, W. Wu, F. Questier, D.L. Massart, C. Boucon, S. De Jong, Anal. Chem. (2000) in press.
[18] B.G.M. Vandeginste, D.L. Massart, L.M.C. Buydens, S. De Jong, P.J. Lewi, J. Smeyers-Verbeke, Data Handling in Science and Technology 20B: Handbook of Chemometrics and Qualimetrics – Part B, Elsevier, Amsterdam, 1998.
[19] Hyperchem 3.0, Hypercube Inc. and Autodesk Inc., 1993.
[20] R. Todeschini, WHIM-3D 3.3, Talete srl., Milan, Italy, 1997 (http://www.disat.unimi.it/chm/page4.htm).
[21] H. Matter, J. Med. Chem. 40 (1997) 1219.
[22] C.B. Eap, C. Cuendet, P. Baumann, Naunyn Schiedeberg's Arch. Pharmacol. 337 (1988) 220.