Available online at www.sciencedirect.com
ScienceDirect
International Journal of Pharmaceutics 339 (2007) 175–188
INTERNATIONAL JOURNAL OF PHARMACEUTICS
www.elsevier.com/locate/ijpharm

Statistical cluster analysis of pharmaceutical solvents

Dong Xu [a], Nancy Redman-Furey [b,*]

[a] *Computational Science Research Center, San Diego State University, 5500 Campanile Drive, San Diego, CA 92182-1245, USA*
[b] *Analytical Sciences Department, Procter & Gamble Pharmaceuticals, Inc., P.O. Box 191, Norwich, NY 13815, USA*

Received 11 December 2006; received in revised form 28 February 2007; accepted 1 March 2007
Available online 12 March 2007

Abstract

High efficiency in polymorph screening and crystallization optimization can be gained by judicious selection of solvents for the study design. Examination of all 57 (classes 2 and 3) pharmaceutical solvents may enable a complete study design but is costly in terms of time and resources. Based on a 17 descriptor dataset specifically constructed for all the classes 2 and 3 pharmaceutical solvents recognized by the International Conference of Harmonization (ICH), an optimal two-stage cluster analysis was carried out together with principal component analysis as a dimensionality and colinearity reduction pre-processor. Both hierarchical average linkage cluster analysis and non-hierarchical *K*-means cluster analysis converged on a 20-cluster solution with strong statistical criteria support and excellent agreement in cluster memberships, which can be reasonably interpreted from a chemical perspective. This 20-cluster solution is offered as an option for design of more efficient solid state screening studies. Rather than designing a polymorph screen to include all 57 solvents, the inclusion of a single member from each of the 20 clusters would be expected to adequately represent the full range of solvent properties exhibited by the entire 57 member solvent set.

Keywords: Cluster analysis; Principal component analysis; Solvent group; Polymorph screening

1. Introduction

Identification of the appropriate solvents to include in pharmaceutics study design is key to execution of meaningful and efficient studies. The choice of solvents to include in an efficient and effective polymorph screen, identification of solvents for optimizing crystallization yields and purity, selection of optimal solvent(s) for facilitating reaction rate and completion, and the choice of most efficient solvents for effecting analytical or preparative separations are typical examples of how solvent selection impacts everyday decisions and work efficiency. The choice of solvent(s) is traditionally based upon existing solvent classification schemes and often the experience of the chemist. Commonly used classifications recognize differences in solvent "strength" based upon acidity, polarity, volume and/or solubility or other characteristics. The insight that interactions involving numerous solvent properties may be important, particularly in the choice of solvents for polymorph screens, has recently resulted in the use of statistical tools to evaluate multiple solvent parameters and attempt to group solvents using cluster analysis. Gu et al. (2004) applied cluster analysis to 96 solvents using eight solvent parameters including hydrogen bond donor and acceptor propensity as well as several bulk solvent properties. This study considered significantly more parameters than those of previously reported studies Snyder (1978) and Carlson (1992) that were merely based on single or a small number of solvent properties.

Although the solvents used in the Gu study exhibited a broad range of properties, a significant portion were not pharmaceutically related and many solvents of pharmaceutical interest were not included. In this paper, we focus on the most commonly used pharmaceutical solvents, i.e. all of the classes 2 and 3 solvents as recognized by ICH (International Conference of Harmonization, U.S. FDA, 1997). The objective of this paper was to quantify chemical intuition in the grouping of pharmaceutical solvents assessing the use of a broader range of solvent properties than previously evaluated.

The statistical method used here does not lead to an all encompassing, rigid classification of solvents, rather, the exploration of a quantitative, logical approach on rational identification of

* Corresponding author. Tel.: +1 607 335 2601; fax: +1 607 335 2300.
E-mail address: redmanfurey.nl@pg.com (N. Redman-Furey).

0378-5173/$ – see front matter © 2007 Elsevier B.V. All rights reserved.
doi:10.1016/j.ijpharm.2007.03.002

solvents based on their similarities/dissimilarities. A grouping of 20 clusters is provided, justified by rigorous statistics. Optimally, an investigator may be able to use one solvent from each of the clusters within a polymorph to test the diversity anticipated from the entire set 57 pharmaceutically acceptable solvents, cutting the testing by more than half. For investigators willing to risk more or less in test reduction, the collection and choice of descriptors is fully described to enable future researchers to adjust statistical treatments to address more specific clustering goals. For some, the less diverse clustering of 12 groups may be adequate while for highly sensitive systems, the 22 cluster approach may be more appropriate. The full hierarchy of clustering is shared to enable the reader to fully explore the utility of the proposed clusters.

2. Database and analysis methods

2.1. Solvent subjects

The 57 classes 2 and 3 pharmaceutical solvents recognized by ICH and water were included in this study. The xylene mixture was divided into *m*-xylene, *p*-xylene, *o*-xylene and ethyl benzene and they were treated as individual solvents in this study.

2.2. Data collection

In an effort to better represent the solvents in all possible aspects, 231 solvent physicochemical descriptors ranging from experimentally measured variables to molecular properties calculated from various quantitative structure property relationships (QSPR), empirical/semi-empirical and quantum mechanical methods were collected on the initial data set. The experimental data were taken from over 15 published journals, chemical handbooks and online chemical databases. A series of 170 molecular descriptors were computed using ADAPT (Stuper and Jurs, 1976; Jurs et al., 1979), a software package developed by Peter Jurs of the Pennsylvania State University. About 30 hydrogen bond related descriptors were calculated by HYBOT 3D, a hydrogen bond thermodynamic package developed by Raevsky (1997) and Raevsky and Skvortsov (2002).

2.3. Multiple imputation of missing data

The treatment of missing values was carefully considered. This was important because subjects with missing value(s) would have been excluded by commonly used statistical software packages (i.e. SAS) from cluster analytic procedures which require a complete data matrix. To avoid the potential for bias due to case deletion and the pitfalls associated with ad hoc imputation procedures such as mean substitution, multiple imputation was used to develop a "complete" dataset. Instead of replacing each missing value with a set of $m > 1$ plausible values drawn from their prediction distribution (Rubin, 1987), a QSPR type strategy was adopted to replace missing values with the average value of two structurally similar compounds whose data were available.

2.4. Data set refinement

Although the preliminary data set was extensive, a smaller data set was desired for more efficient and straightforward analysis. Since the time complexity for a typical hierarchical cluster analysis is $O(n^2)$ or greater, reducing the size of the data set would significantly improve the computation time. After careful review of technical background, the completeness and reliability of all the descriptors, 40 out of the initial 231 descriptors were selected as the first refinement of the data set. This reduction in size was the result of culling out the obviously redundant descriptors, the removal of descriptors for which too few data values were available, and the removal of calculated descriptors that included descriptors already chosen for inclusion. For instance, most of the references examined provided some measurement of fundamental solvent properties such as acidity, polarizability, or refractive index although they were often identified by differing symbols or names. Rather than duplicating descriptors, the data set that was most complete and that appeared most consistent with other available literature was chosen to represent that parameter. For parameters that appeared to be unique but for which there was incomplete data and no obvious means to reliably extrapolate from known compounds, the descriptor was discarded. The largest decrease in number of descriptors occurred by discarding those that were derived from calculations involving descriptors already included in the database. To the extent possible, data was validated by comparison across references and publicly available databases.

A closer look at the refined data matrix revealed further dimensionality reduction was possible due to the correlation between the descriptors. Data correlation can severely deteriorate the quality of the statistical analysis results, particularly to hierarchical cluster analysis. By clustering and removing closely correlated descriptors ($|r| >= 0.9$), we effectively brought the total number of descriptors from 40 down to 17. A final data set was built on the 17 variables (see Table 1) whose Pearson correlation coefficient matrix is shown in Table 2. Again, when eliminating descriptors identified as redundant upon statistical evaluation, efforts were made to retain those descriptors with the fewest extrapolated data points and to retain those with the greatest substantiation from the literature.

2.5. Detail of basis descriptors

The final set of solvent descriptors included experimental properties from literature and computed properties based on 3D molecular geometries. Both were indispensable and contributed a more complete picture of the solvents. Each of the 17 descriptors is discussed below.

The index of refraction (n) is often used to describe the polarizability and London interactions (or dispersion) of solvents. Because of the availability of the data, a standard temperature of 20 °C was selected. The cohesion of a solvent (as a measure of the energy required to create a cavity in the solvent) was approximated by using the δ parameter of Hildebrand, which equals the square root of cohesive energy density. The bulk properties of: boiling point, BP, viscosity (η) and surface tension (γ) pro-

Table 1
Final 17-descriptor data set for 58 pharmaceutical solvents

Solvent ID	SIAEprobe	SIEDprobe	VOL	PPSA-1	PNSA-1	CNTA	PNHS-1	n	π_2^H	γ	ε	μ	Φ	η	δ	PI	BP
1,1,2-Trichloroethene	−3.7714	0	350.1	44.02	212.3	0	0	1.477	0.37	41.45	3.42	0.77	0	0.55	18.76	3.2	87
1,2-Dichloroethene	−6.94	0	308.8	77.09	157.3	0	0	1.449	0.61	45.86	9.2	1.9	0	0.45	19.48	3.9	60.5
1,2-Dimethoxyethane	−51.7584	0	420.5	265.5	35.05	2	58.07	1.38	0.66	32.38	7.2	1.71	0	0.45	17.73	4.3	84.1
1,4-Dioxane	−48.687	0	360.4	211.9	46.21	2	59.47	1.422	0.75	47.14	2.2	0	0	1.21	20.16	2.2	101.3
1-Butanol	−34.6322	−10.3407	386	236.3	39.46	1	63.32	1.399	0.42	35.88	17.8	1.66	0	2.6	23.3	4.4	117.7
1-Pentanol	−35.2572	−10.9686	441.2	269	39.7	1	63.48	1.41	0.42	36.5	13.9	1.7	0	3.43	22.6	4.1	137.8
1-Propanol	−35.1683	−10.7457	325.4	204.4	39.2	1	63.67	1.385	0.42	33.57	20.33	1.68	0	1.94	24.6	4	97.2
2-Butanol	−23.228	−9.5546	376.7	233.6	36.03	1	47.02	1.398	0.36	32.44	16.68	1.66	0	3.25	23.8	4.3	99.6
2-Butanone	−20.405	0	356.6	215.3	45.24	1	43.49	1.379	0.7	34.5	18.4	2.76	0	0.4	18.8	4.4	79.6
2-Ethoxyethanol	−52.4471	−11.1439	420.7	253.6	48.34	2	75.61	1.405	0.5	40.8	2.41	2.08	0	1.87	21.54	2.7	135
2-Hexanone	−20.4209	0	480.7	279	46.5	1	43.23	1.401	0.68	36.63	14.6	2.68	0	0.56	18.14	4.1	127.7
2-Methoxyethanol	−60.3736	−10.5643	359.2	217	48.23	2	92.98	1.402	0.5	44.39	16.93	2.04	0	1.39	23.2	4.4	124.4
2-Methyl-1-propanol	−26.4942	−11.3138	377.4	231.1	36.58	1	56.5	1.396	0.39	32.38	17.7	1.64	0	3.27	23.8	4.4	107.7
2-Propanol	−32.6688	−10.8236	324.3	197.8	41.2	1	53.63	1.378	0.36	30.13	19.92	1.66	0	1.96	23.6	3.9	82.3
acetic acid	−27.2151	−27.7826	269.7	131.6	80.57	2	89.7	1.372	0.65	39.01	6.2	1.74	0	1.13	18.4	4.8	117.9
Acetone	−30.4565	0	299.9	182.1	47	1	53.66	1.359	0.7	33.77	20.7	2.88	0	0.31	19.8	5.1	56.3
Acetonitrile	−18.5655	0	240.6	132.5	61.99	1	0	1.344	0.9	41.25	37.5	3.92	0	0.35	24.09	5.8	81.6
Anisole	−19.9958	0	442.2	211.6	93.56	1	27.24	1.517	0.75	50.52	4.3	1.36	0.75	1.06	20.1	3.9	153.6
Butyl acetate	−14.1362	0	510.4	291.9	57.44	2	55.41	1.395	0.6	35.81	5.1	1.84	0	0.68	17.6	3.3	126
Chlorobenzene	−2.3643	0	401.6	131	150.4	0	0	1.524	0.65	47.48	5.6	1.69	0.86	0.73	19.26	2.7	131.7
Chloroform	−1.2276	0	313	25.17	209.1	0	0.38	1.446	0.49	38.39	4.8	1.01	0	0.54	19.03	4.1	61.2
Cumene	0	0	522.5	259.6	81.47	0	0	1.492	0.49	39.85	2.4	0.39	0.67	5.97	17.5	1.6	152.4
Cyclohexane	0	0	408.4	271.2	11.21	0	0	1.427	0.1	35.48	2	0.61	0	0.9	16.7	0.2	80.7
Dichloromethane	−2.6312	0	269	58.7	151.6	0	3.99	1.424	0.57	39.15	9.1	1.6	0	0.44	20.38	4	39.8
Dimethyl sulfoxide	−41.2269	0	322.3	192.3	46.11	1	63.11	1.417	1.74	61.78	46.68	3.96	0	1.97	26.3	5.8	189
Ethanol	−36.4423	−11.3162	267.2	172.4	38.43	1	64.33	1.361	0.42	31.62	24.55	1.69	0	1.06	26.4	4.3	78.3
Ethyl acetate	−15.7045	0	394.3	227.6	56.43	2	56.55	1.372	0.62	33.67	6.08	1.78	0	0.42	18.3	4.4	77.1
Ethyl benzene	0	0	470.7	232.4	84.54	0	2.24	1.496	0.51	41.38	2.45	0.59	0.75	0.63	17.4	2.4	136.19
Ethyl ether	−18.154	0	390.1	258	25.32	1	12.45	1.353	0.27	23.96	4.3	1.15	0	0.22	15.5	2.8	34.4
Ethyl formate	−23.5258	0	334.8	172.4	79.02	2	88.72	1.36	0.66	33.36	7.1	1.93	0	0.38	18.9	4.9	54.3
Ethyleneglycol	−68.9093	−21.8239	297.9	168.3	60.67	2	127.7	1.432	0.9	69.07	37.7	2.31	0	17.65	34.48	5.8	197.3
Formamide	−39.857	−28.5611	222.3	116	67.87	1	88.86	1.447	1.3	82.08	84	3.73	0	3.33	39.28	9.6	219.9
Formic acid	0	−17.5818	209	82.77	92.33	2	114.6	1.371	0.6	53.44	58	1.42	0	1.64	21.4	7.1	100.6
Heptane	0	0	529.1	336.4	19.02	0	0	1.388	0	28.28	1.92	0	0	0.5	15.2	0.1	98.4
Hexane	0	0	479.6	304.3	18.38	0	0	1.375	0	25.75	1.88	0	0	0.3	14.99	0.1	68.7
Isoamyl alcohol	−35.0654	−10.5121	438.5	255	42.86	1	62.93	1.407	0.39	34.13	15.63	1.8	0	3.69	22.3	4.2	131.2
Isobutyl acetate	−15.0155	0	506.9	284.3	56.01	2	49.86	1.39	0.57	33.19	5.6	1.87	0	0.65	17.2	3.4	116.7
Isopropyl acetate	−13.3658	0	442.4	254.3	53.4	2	44.67	1.377	0.57	31.32	6.3	1.75	0	0.53	17.1	3.7	88.5
Methanol	−41.5267	−22.9609	205.6	135.8	38.13	1	81.55	1.329	0.44	31.77	32.7	1.7	0	0.54	29.52	5.1	64.7
Methyl acetate	−15.1535	0	332.8	191.1	56.13	2	73.32	1.361	0.64	35.59	6.7	1.68	0	0.35	19.4	4.9	56.9
Methylcyclohexane	0	0	460.5	294.8	14.13	0	0	1.423	0.1	33.52	2.02	0	0	0.68	16.27	0.9	100.9
Methylisobutylketone	−20.1482	0	466.8	268.5	46.11	1	36.13	1.396	0.65	36.63	14.6	2.68	0	0.56	18.14	4.1	127.7
m-Xylene	0	0	468.7	235.4	83.79	0	0	1.497	0.52	40.98	2.3	0.3	0.75	0.58	17.84	2.6	139.1

Table 1 (*Continued*)

Solvent ID	SIAEprobe	SIEDprobe	VOL	PPSA-1	PNSA-1	CNTA	PNHS-1	n	π_2^H	γ	ε	μ	Φ	η	δ	PI	BP
NN-Dimethylacetamide	−23.5275	0	391.5	238.7	37.66	1	60.66	1.438	1.33	47.62	37.78	3.81	0	1.93	22.35	5.2	166.1
NN-Dimethylformamide	−30.4904	0	339.2	205.5	44.78	1	76.43	1.431	1.31	49.56	36.71	3.82	0	0.84	23.97	5.3	153
Nitromethane	−12.8519	0	254.3	99.32	103.7	2	15.12	1.382	0.95	52.58	39.4	3.46	0	0.62	25.76	5.2	101.2
N-Methyl pyrrolidone	−36.9815	0	413.1	242.5	45.11	1	57.46	1.47	0.86	43.18	32.2	4.09	0	1.48	23.35	5	202
o-Xylene	0	0	460.4	223.7	87.18	0	0	1.506	0.56	42.83	2.55	0.62	0.75	1.99	18.45	3	144.4
Pentane	0	0	413.4	272.4	18.14	0	0	1.358	0	22.3	1.8	0	0	0.25	14.4	0	36.1
Propyl acetate	−14.793	0	458	259.6	56.86	2	55.23	1.384	0.6	34.26	6.3	1.79	0	0.57	18	4	101.5
p-Xylene	0	0	469	237	81.48	0	0	1.496	0.52	40.32	2.27	0	0.75	0.6	17.84	2.5	138.4
Pyridine	−26.2461	0	343.2	174.2	74.62	1	25.12	1.51	0.84	52.62	12.5	2.19	0.83	0.9	21.8	5.3	115.3
Sulfolane	−54.8841	0	408.5	189.5	94.11	2	93.95	1.481	1.7	61.89	43.3	4.69	0	10.1	26.3	5.4	285
tert-Butyl methyl ether	−18.0964	0	424.1	260.6	28.97	1	21.51	1.369	0.29	24.41	2.6	1.36	0	0.33	15.1	1.3	55.2
Tetrahydrofuran	−28.9342	0	336.7	213.9	32.7	1	35.26	1.405	0.52	39.44	7.58	1.63	0	0.47	19.1	3.8	64.9
Tetralin	0	0	528.3	269.7	73.97	0	2.16	1.541	0.52	47.74	2.77	0.22	0.6	2.41	19.52	2.4	207.6
Toluene	0	0	413	198	89.37	0	0	1.496	0.52	40.2	2.33	0.36	0.86	0.62	18.35	2.4	110.6
Water	0	−45.7117	136.3	72.24	57.22	1	129.5	1.333	0.45	104.7	78.36	1.87	0	0.89	45.78	8.8	100

vide different descriptions of the ability of solvent molecules to self-associate and the physical consequences thereof.

The polarity of solvents is linked both to the dipole moment (μ) and to the dielectric constant (ε). Polarity index, PI, a relative measure of the degree of interaction of the solvent with various polar test solutes, was also included. The above properties are important with regard to the solvation abilities of the solvents (Marcus, 1993).

In the early 1990s, Abraham et al. published a series of studies on solvent hydrogen bonding properties. They established the hydrogen-bond acidity (α_2^H) and hydrogen-bond basicity (β_2^H) scales as well as dipolarity/polarizability scale (π_2^H), which had been determined experimentally for over 1000 compounds.The π_2^H scale was statistically chosen to be a part of the final descriptor set, while α_2^H and β_2^H are represented by SIAEprobe and SIEDprobe, respectively, because of their close correlations. The dipolarity/polarizability scale (π_2^H) represents the ability of electrons to move and be delocalized in a molecule and is a measure of the polar interaction between the compound and the electron receptor. In Abraham's papers, McGowan's characteristic volume (V_x) is also a very interesting descriptor as it is an estimate of the van der Waals volume and, unlike the widely used molar volume, is not dependent on hydrogen-bonding and other properties. There is a very close correlation between V_x and the VOL descriptor, a computer generated intrinsic volume using the SAVOL program (Pearlman, 1980). We chose to use VOL over V_x. because there were a few outliers and missing values in the V_x data.

PPSA-1and PNSA-1 are two of the charged partial surface area (CPSA) descriptors developed by Stanton and Jurs (1990) and Stanton et al. (1992, 2002). They combined molecular surface area and partial atomic charge information to form CPSA descriptors that encode structural features responsible for polar interactions between molecules. PNHS-1 is among the Hydrophobic Surface Area (HSA) descriptors that are designed to capture information regarding the structural features responsible for hydrophobic/hydrophilic intermolecular interactions. It is a better representation than the commonly used log P descriptor, the logarithm of the n-octanol/water partition coefficient, because the HSA descriptors capture regional and localized hydrophilic effects on a molecular surface (Stanton et al., 2004).

Hydrogen bonding plays a fundamental role in solvent–solute interactions. It is a strong interaction where the acceptor group provides an electron pair, and the donor group provides a proton. The proton is thought of as being shared between two atoms (usually heteroatoms). Interactions between H-bond donor and acceptor molecules are known to result in the formation of many molecular and ionic complexes which are of great importance in chemical and biochemical processes. Many scales have been proposed to quantify the H-bonding strength over the years. The two descriptors, SIAEprobe and SIEDprobe, calculated by the three-dimensional HYBOT (Hydrogen Bonding Thermodynamics) program (Raevsky and Skvortsov, 2002) were selected to account for H-bonding interactions. These 3D H-bond descriptors reveal the surface enthalpy values when a molecule's acceptor/donor atoms interact with a donor/acceptor probe (a small molecule such as water or ammonia). They are

Table 2
Pearson correlation coefficient of the final 17-descriptor data set

	SIAEprobe	SIEDprobe	VOL	PPSA-1	PNSA-1	CNTA	PNHS-1	n	π_2^H	γ	ε	μ	Φ	η	δ	PI	BP
SIAEprobe	1.00																
SIEDprobe	0.29	1.00															
VOL	0.25	0.60	1.00														
PPSA-1	−0.03	0.36	0.81	1.00													
PNSA-1	0.31	0.11	−0.19	−0.73	1.00												
CNTA	−0.63	−0.23	−0.23	0.01	−0.25	1.00											
PNHS-1	−0.68	−0.67	−0.48	−0.16	−0.26	0.75	1.00										
n	0.22	0.32	0.40	−0.02	0.47	−0.48	−0.42	1.00									
π_2^H	−0.44	0.01	−0.27	−0.28	0.17	0.37	0.39	0.22	1.00								
γ	−0.15	−0.55	−0.47	−0.48	0.24	0.11	0.44	0.25	0.56	1.00							
ε	−0.32	−0.62	−0.67	−0.44	−0.06	0.27	0.62	−0.21	0.57	0.75	1.00						
μ	−0.54	−0.11	−0.39	−0.24	−0.03	0.47	0.49	−0.16	0.80	0.38	0.66	1.00					
Φ	0.39	0.23	0.35	0.05	0.29	−0.47	−0.47	0.78	−0.01	0.11	−0.30	−0.37	1.00				
η	−0.48	−0.29	−0.03	−0.02	−0.01	0.20	0.40	0.20	0.31	0.37	0.29	0.22	−0.03	1.00			
δ	−0.42	−0.78	−0.67	−0.47	−0.01	0.23	0.64	−0.13	0.41	0.80	0.86	0.48	−0.19	0.42	1.00		
PI	−0.42	−0.57	−0.70	−0.57	0.14	0.48	0.70	−0.19	0.63	0.67	0.83	0.71	−0.24	0.22	0.78	1.00	
BP	−0.33	−0.12	0.22	0.12	0.02	0.09	0.27	0.58	0.63	0.54	0.38	0.40	0.27	0.58	0.37	0.28	1.00

Units: SIAEprobe, SIEDprobe (kcal/M Å^2); VOL (Å^3); PPSA-1, PNSA-1 (Å^2); CNTA (none); PNHS-1 (Å^2); n,π_2^H, γ, ε (none); μ (Debye); η (c.p.); δ (J/cm^3)$^{1/2}$; PI (Snyder); BP (°C).

the only scales that are created from a thermodynamic perspective and a force-field method is used to determine distance and angle dependencies in three-dimensional spaces. The new 3D descriptors are the extension of 2D HYBOT and they provide much higher accuracy and are believed to be superior to any previous 2D or empirical scales (including α_2^H and β_2^H). CNTA is another hydrogen bond related descriptor developed by Stanton et al. (1992) as an analog of the CPSA to capture information about structural features responsible for hydrogen-bonding interactions. They were computed similarly to the CPSAs, but only considered the heteroatoms and their attached hydrogens that participate in hydrogen bonds.

Lastly, the aromaticity data, Φ, was taken from Minnesota Solvent Descriptor Database (Winget et al., 1999). This descriptor encodes not only the structural information but also the hydrophobility property of a solvent through the fraction of non-hydrogenic solvent atoms that are aromatic carbon atoms.

In summary, the final set of descriptors represented a wide range of solvent physicochemical properties, from solvent bulk property to hydrophilicity, from H-bonding interactions to polar interactions. These were imperative if the following statistical cluster analyses are to provide insight into how similar/dissimilar these pharmaceutical solvents are to each other. The definitions and sources of the 17 solvent descriptors are listed in Table 3.

2.6. Data standardization

The objective of standardization of the data matrix is to remove arbitrary effects caused by choice of measurement unit and/or to account for differences in measurement scales across variables (Romesburg, 1990). Since the solvent descriptor data were obtained from different sources and are different in nature, they do not have equal variance. The variables with large variances tend to have more effect on the resulting clusters than those with small variances particularly when the dissimilarity measure, such as Euclidean distance, is sensitive to such differences. Thus, we employed the typical z-score standardization so that each column (variable) would have a mean of zero and variance of one.

2.7. Principal components analysis (PCA)

The next phase of the analysis involved principal components analysis (PCA), which was used to further eliminate the inherent collinearity in the final data set and effectively reduce the total number of variables. It is important to eliminate this collinearity, prior to clustering, to prevent variables that are still highly correlated from exerting a disproportionate amount of influence and thereby biasing the results (Eder et al., 1994). This was achieved through the application of PCA, which linearly transformed a collinear data matrix into a hierarchy of orthonormal and thereby independent components, each of which explained successively less and less of the total variance. The initial number of "principal components" (PCs) was 17, equal to the number of variables in the final data set. Various tests were adopted to determine the number of PCs to retain. The Cattell (1966) scree test and the Kaiser (1960) criterion are the most frequently used procedures. They are both based on inspection of the correlation matrix eigenvalues. Cattell's recommendation was to retain only those components above the point of inflection on a plot of eigenvalues ordered by diminishing size. Kaiser (1960) recommended that only eigenvalues at least equal to one are retained as one is the average size of the eigenvalues in a full decomposition.

2.8. Two-stage cluster analysis

The resulting PC scores were analyzed by a two-stage clustering approach in which hierarchical clustering was used in conjunction with an iterative partitioning (reallocation) method. This approach is often considered an optimal way to perform

Table 3
Definitions and sources of the 17 solvent descriptors

	Descriptor	Definition	Unit	Source
1	SIAEprobe	Surface integral for enthalpy values of interactions between acceptor atoms of a molecule and a donor probe on the surface	kcal/M Å^2	HYBOT 3D by Raevsky (1997) and Raevsky and Skvortsov (2002)
2	SIEDprobe	Surface integral for enthalpy values for interactions between donor atoms of a molecule and an acceptor probe on the surface	kcal/M Å^2	HYBOT 3D by Raevsky (1997) and Raevsky and Skvortsov (2002)
3	VOL	Intrinsic volume	Å^3	SAVOL by Pearlman (1980)
4	PPSA-1	Type 1 partial positive surface area	Å^2	ADAPT by Stuper and Jurs (1976), Jurs et al. (1979), Stanton and Jurs (1990) and Stanton et al. (1992, 2002)
5	PNSA-1	Type 1 partial negative surface area	Å^2	ADAPT by Stuper and Jurs (1976), Jurs et al. (1979), Stanton and Jurs (1990) and Stanton et al. (1992, 2002)
6	CNTA	Simple count of all hydrogen bond acceptor groups	None	ADAPT by Stuper and Jurs (1976), Jurs et al. (1979), Stanton and Jurs (1990) and Stanton et al. (1992, 2002)
7	PNHS-1	Type 1 hydrophilic surface area	Å^2	ADAPT by Stuper and Jurs (1976), Jurs et al. (1979), Stanton and Jurs (1990) and Stanton et al. (1992, 2002)
8	n	Index of refraction (20 °C)	None	Winget et al. (1999), Abboud and Notari (1999), Lide (2005), CRC Press (2005) and Riddick et al. (1986)
9	π_2^H	Abraham's dipolarity/polarizability	None empirical scale derived from gas-chromatographic measurements	Abraham et al. (1991), Abraham (1993a, 1993b, 1993c), Abraham et al. (1994) and Abraham and Rafols (1995)
10	γ	Surface tension	None. $\gamma = \gamma_m/\gamma_0$, where γ_m is macroscopic surface tension at a liquid-air interface at 298 K, and γ_o is 1 cal mol^{-1} Å^{-2}	Winget et al. (1999), Lide (2005) and CRC Press (2005)
11	ε	Dielectric constant	None. Note that dielectric constant is also called relative permittivity	Mirmehrabi and Rohani (2005); CHEMnetBase, CRC Press (2005)
12	μ	Dipole moment	Debye	Mirmehrabi and Rohani (2005); CHEMnetBase, CRC Press (2005)
13	Φ	Aromaticity	None. fraction of non-hydrogenic solvent atoms that are aromatic carbon atoms	Winget et al. (1999)
14	η	Viscosity	c.p. (10^{-3} Pa s)	Mirmehrabi and Rohani, 2005, Trimen Database (2005)
15	δ	Hildebrand solubility	(J/cm^3)$^{1/2}$	Mirmehrabi and Rohani (2005); ADAPT by Stuper and Jurs (1976), Jurs et al. (1979), Stanton and Jurs (1990) and Stanton et al. (1992, 2002)
16	PI	Snyder polarity index	Snyder	Mirmehrabi and Rohani (2005) and Chastrette et al. (1985)
17	BP	Boiling point	°C	Mirmehrabi and Rohani (2005)

classifications, primarily because it allows one to take advantage of the beneficial characteristics of both hierarchical and iterative methods (Punj and Stewart, 1983; Milligan, 1996).

The first stage involved using a hierarchical average linkage clustering procedure (Sokal and Michener, 1958) to develop an initial partition. Although there was no clear consensus as to which of the many hierarchical cluster analysis approaches to use, in an evaluation of three of the more common linkage methods (Ward's, average linkage, and centroid) Kalkstein et al. (1987) and Bunkers et al. (1996) found that average linkage, specifically the unweighted pair-group method using arithmetic averages (UPGMA), was far superior to both the centroid and Ward's method in producing clusters with the smallest within-cluster variance and large between-cluster separation.

With hierarchical average linkage, which is an agglomerative method, each solvent (as represented by its component scores) initially started out as an individual cluster. Similar solvents were then merged, where cluster similarity was determined by the mean distance between all objects in the clusters (weighted by the number of members), until all solvents were united into one cluster.

Different statistics were used in combination to determine the number of solvent groups (clusters) resulting from hierarchical analysis. The Calinksi and Harabasz (1973) and Duda and Hart

(1973) tests are the two stopping rules that have been shown to outperform other rules such as cubic clustering criterion (CCC) in simulation studies on data sets with known structure and different levels of introduced error (Milligan and Cooper, 1985; Cooper and Milligan, 1988). They are available in SAS and are known as the pseudo-F and pseudo-t^2 statistics, respectively (SAS OnlineDoc, 2000). Consistency in the number of clusters suggested by pseudo-F and pseudo-t^2 statistics in hierarchical average linkage clustering provided an initial basis for selecting the final number of solvent groups.

The second phase involved performing non-hierarchical K-means cluster analysis on the principal component data set using the centroids (means of the component scores) of clusters obtained from hierarchical analysis as initial seeds (Davis and Kalkstein, 1990). A non-hierarchical iterative clustering method requires the number of clusters and cluster seeds be specified up front. After all solvents had been assigned to the cluster with the nearest seed values, the centroid of each cluster was recalculated and then used as the new cluster seed. Using the FASTCLUS procedures (SAS 2004, version 8.2), these steps were repeated until the change in new seed values converge to within 0.02. This method allowed free movement of cluster members after they had been classified into a given group at each step of the cluster analysis, thereby acting to optimize the final cluster groupings. Milligan (1980) suggested that application of this second stage achieves even smaller within-group variation than using average linkage alone. This two-stage approach appeared to be superior in terms of cluster cohesiveness and had been used successfully in numerous studies (Eder et al., 1994; Zelenka, 1997).

3. Results and discussion

3.1. Retainment of number of PCs and final PCA data set

The nature of the decision on the number of PCs retained was arbitrary. Table 4 lists the eigenvalues of all 17 PCs. In this case, 4 PCs would be kept based on Kaiser criterion and

Table 4
Eigenvalues of the correlation matrix in PCA

PC	Eigenvalue	Difference	Proportion (%)	Cumulative (%)
1	6.979415	3.839357	41.06	41.06
2	3.140058	0.642601	18.47	59.53
3	2.497457	1.014956	14.69	74.22
4	1.482501	0.576676	8.72	82.94
5	0.905826	0.290258	5.33	88.27
6	0.615567	0.194653	3.62	91.89
7	0.420914	0.151283	2.48	94.36
8	0.269631	0.05748	1.59	95.95
9	0.212151	0.07642	1.25	97.2
10	0.135731	0.037703	0.8	98
11	0.098028	0.022669	0.58	98.57
12	0.075358	0.010784	0.44	99.02
13	0.064575	0.013077	0.38	99.4
14	0.051497	0.013901	0.3	99.7
15	0.037597	0.024668	0.22	99.92
16	0.012929	0.012163	0.08	100
17	0.000766		0	100

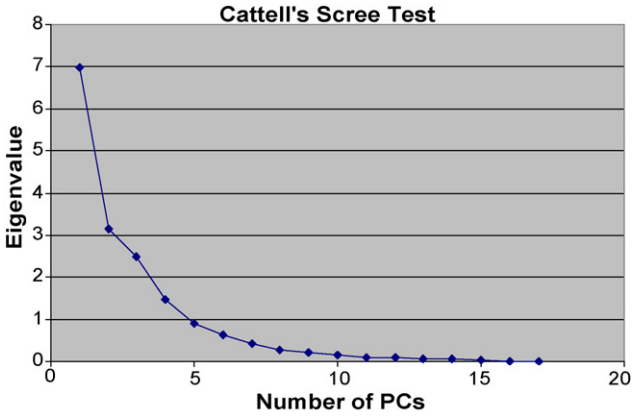

Fig. 1. Scree plot of the PC eigenvalues.

Cattell's scree test (Fig. 1) may suggest as many as 8 PCs. However, combining information of proportion of variance (>5%) that each component counts with the cumulative proportion of variance (>88%) from Table 4 enables justification for the use of five principal components as a reasonable summary of the data set. The loadings of the original variables on the 5 PCs are displayed in Tables 5 and 6 shows the final data set containing scores for each solvent on the five PCs, which was used to carry out the two-stage cluster analysis.

3.2. Determination of number of clusters

Selection of the final average linkage solution was achieved through examination of pseudo-F and pseudo-t^2 statistics. The pseudo-F test was considered a "global" stopping rule since it made use of all the information contained in a partition into n clusters for each value of n (Gordon, 1996). The pseudo-F describes the ratio of between-cluster to within-cluster variance and the pseudo-t^2 quantifies the difference between two clusters that are merged at a given step. Maximum value of the pseudo-

Table 5
PC loadings

Solvent descriptor	Prin1	Prin2	Prin3	Prin4	Prin5
SIAEprobe	−0.22967	0.165863	−0.30257	−0.06941	0.336436
SIEDprobe	−0.25543	0.070927	0.1677	0.500757	0.057481
VOL	−0.27013	0.055469	0.390699	−0.04452	0.045436
PPSA-1	−0.1861	−0.19202	0.455424	−0.17892	0.271932
PNSA-1	−0.00038	0.3649	−0.30447	0.27475	−0.43313
CNTA	0.207207	−0.27448	0.190673	0.215524	−0.27492
PNHS-1	0.312678	−0.18865	0.115829	−0.11319	−0.1774
n	−0.09326	0.508209	0.162104	−0.02035	−0.08994
π_2^H	0.245306	0.210972	0.212074	0.409647	0.138956
γ	0.275753	0.286873	−0.04625	−0.18384	0.153163
ε	0.336784	0.049451	−0.06601	−0.07362	0.362918
μ	0.27498	−0.00169	0.149574	0.425448	0.240952
Φ	−0.1432	0.411076	0.046079	−0.14397	0.043909
η	0.165527	0.126877	0.294526	−0.26149	−0.49258
δ	0.332592	0.077336	−0.08051	−0.27466	0.108937
PI	0.344991	0.046994	−0.09472	0.119589	0.097809
BP	0.149022	0.320358	0.416357	−0.09161	0.078905

Table 6
Final PCA data set

Solvent ID	Prin1	Prin2	Prin3	Prin4	Prin5
112-Trichloroethene	−1.54624	2.43959	−3.15652	1.12764	−1.85687
12-Dichloroethene	−0.70169	1.65152	−2.78152	1.41164	−0.674
12-Dimethoxyethane	−0.0799	−2.0605	1.304	0.90502	−0.53088
14-Dioxane	−0.14168	−0.95403	0.79869	0.05299	−1.33297
1-Butanol	0.57575	−1.01575	0.51615	−0.83708	−0.23687
1-Pentanol	0.29543	−0.81955	1.30532	−1.06444	−0.33045
1-Propanol	0.79262	−1.31046	−0.23727	−0.70172	−0.25869
2-Butanol	0.17105	−1.05327	0.14629	−0.8791	−0.14761
2-Butanone	0.04183	−1.07833	−0.14981	1.10021	0.65823
2-Ethoxyethanol	0.60042	−1.38016	1.62705	−0.31388	−1.10062
2-Hexanone	−0.54119	−0.62624	1.29831	0.71793	0.86281
2-Methoxyethanol	1.80292	−1.41803	0.98754	−0.20672	−0.98572
2-Methyl-1-propanol	0.42875	−1.07644	0.25411	−0.96674	−0.23509
2-Propanol	0.46421	−1.4902	−0.52684	−0.61333	−0.32585
Acetic acid	1.86541	−1.21594	−1.08546	−0.29057	−1.42786
Acetone	0.72009	−1.47655	−0.73494	1.33235	0.37771
Acetonitrile	1.75405	−0.43806	−1.54586	1.94755	1.37029
Anisole	−0.94475	2.71625	0.88431	0.09658	−0.215
Butyl acetate	−0.98165	−1.09438	1.52576	0.57815	0.04158
Chlorobenzene	−1.79602	3.96877	−0.98283	0.44571	−0.34271
Chloroform	−1.14842	2.0368	−3.7139	1.53344	−1.7141
Cumene	−2.93772	2.60967	1.31871	−1.38183	−0.50722
Cyclohexane	−3.33778	−0.75627	−0.1766	−1.19993	0.77278
Dichloromethane	−0.73653	1.1236	−3.40386	1.35658	−0.76766
Dimethyl sulfoxide	3.79002	1.24434	1.59095	1.74186	1.62813
Ethanol	1.22389	−1.67568	−1.05935	−0.54981	−0.14894
Ethyl acetate	−0.33323	−1.56282	0.06577	0.98447	−0.21998
Ethyl benzene	−2.87878	2.52216	0.1936	−0.63029	0.38765
Ethyl ether	−2.2186	−2.32043	−0.45622	0.20134	0.19396
Ethyl formate	0.52208	−1.71428	−0.71117	1.38907	−0.90614
Ethyleneglycol	6.15905	0.9775	2.74873	−2.69403	−3.46215
Formamide	7.61356	2.32341	−0.33637	−1.11954	1.71616
Formic acid	3.47362	−0.40874	−2.50305	−0.27808	−0.17237
Heptane	−4.24733	−1.29872	0.61718	−1.53219	0.89054
Hexane	−4.15656	−1.63046	−0.10505	−1.30568	0.7255
Isoamyl alcohol	0.30937	−0.87157	1.1421	−0.94573	−0.44723
Isobutyl acetate	−1.07144	−1.25192	1.36427	0.6674	0.00714
Isopropyl acetate	−0.93587	−1.53522	0.54451	0.84062	−0.07851
Methanol	2.52578	−2.15878	−2.00727	−1.11256	−0.1674
Methyl acetate	0.23873	−1.74375	−0.65719	1.03801	−0.39329
Methylcyclohexane	−3.58121	−0.70728	0.21676	−1.41267	0.82084
Methylisobutylketone	−0.54817	−0.6361	1.11168	0.73651	0.85515
m-Xylene	−2.90473	2.55289	0.17893	−0.73298	0.38466
NN-Dimethylacetamide	2.31494	0.41948	−1.55554	2.04606	0.49599
NN-Dimethylformamide	1.64465	0.76532	2.12571	0.97438	1.09099
Nitromethane	1.98597	0.67146	1.71396	1.35279	1.56965
N-Methyl pyrrolidone	2.48309	0.49276	1.09013	1.5067	1.35767
o-Xylene	−2.50441	2.87787	0.29883	−0.69301	0.17534
Pentane	−4.06879	−2.0466	−0.90844	−1.05183	0.49936
Propyl acetate	−0.6595	−1.32458	0.80083	0.7699	−0.07539
p-Xylene	−3.01254	2.49952	0.16939	−0.86008	0.33622
Pyridine	0.17534	2.51288	0.12142	0.55506	0.17097
Sulfolane	4.93966	2.65658	4.18419	1.46913	−0.9978
tert-Butyl methyl ether	−2.47132	−2.0547	0.07477	0.14133	0.06306
Tetrahydrofuran	−0.52933	−1.122	−0.28447	0.39723	0.08407
Tetralin	−2.65095	3.30058	1.58473	−1.387	0.40576
Toluene	−2.782	2.61637	−0.55799	−0.58388	0.12273
Water	7.53608	0.34847	−4.26718	−4.07296	1.99441

F statistic across hierarchy levels (or local "peaks" within the range of possible cluster solutions) indicated the optimal number of clusters. The pseudo-t^2 statistic was a "local" stopping rule based on tests of whether or not a pair of clusters should be merged or remain separate. A small value for the pseudo-t^2 statistic followed by a much larger pseudo-t^2 value for the next cluster fusion suggested that two clusters should not be merged. In addition to pseudo-F and pseudo-t^2 tests, R^2, the squared multiple

Table 7
Statistics used for determination of the number of clusters

Number of clusters	R^2	Pseudo-F	Pseudo-t^2	Pseudo-F (K-means)
27	98.1	62.7	–	62.69
26	97.9	60.7	5.9	60.71
25	97.6	54.9	7.5	58.11
24	97.3	54.2	3.3	57.16
23	97.1	52.7	6.3	55.26
22	96.9	52.9	2.2	55.11
21	96.4	49.8	12.1	51.65
20	96.2	50.1	2.8	51.83
19	95.4	45.4	7.4	49.78
18	95.1	45.9	4.7	49.96
17	94.9	47.3	–	51.41
16	93.8	42.1	10.2	44.14
15	93.4	43.5	1.4	44.41
14	91	34.1	13.8	35.33
13	90	33.8	11.2	34.82
12	89.5	35.8	3.7	36.81
11	88.1	34.8	9.3	35.69
10	86.3	33.6	12.5	34.36
9	82.5	28.8	14	29.65
8	75.8	22.4	15.8	27.76
7	68.9	18.8	27.2	22.59

Note: The shaded area with 12, 15, 17, 20, 22 clusters were potential solutions.

correlation, representing the proportion of variance accounted for by the clusters, was used as a supporting criterion.

In Table 7, the columns displays R^2, pseudo-F and pseudo-t^2 statistics in the average linkage hierarchical clustering procedure and pseudo-F in the non-hierarchical K-means clustering procedure from 7 to 27 cluster generations. Naturally, greater number of clusters implied less heterogeneity in any given cluster. Therefore, the pseudo-F was bound to increase as the number of clusters went up. However, large increments of pseudo-F relative to the previous and subsequent clusters indicated a stopping point. Hierarchies of 7–27 clusters were examined as a region of practical interest. It was assumed that a smaller number of clusters would likely be inadequate to fully represent all key solvent characteristics and that greater than 27 clusters would provide little advantage over testing all 57 solvents individually. The small values for the pseudo-t^2 statistic in conjunction with local peaks of pseudo-F suggested 12, 15, 17, 20, 22 might be the optimal numbers of clusters. To aid the selection of clusters, the squared multiple correlation, R^2, provided a measure of the proportion of variance accounted for by the clusters. With 12 clusters, nearly 90% of the variance was accounted for, 20 clusters yield 96% and 22 clusters provide 97%.

Table 7 was then visualized by the scree graph (Fig. 2) in which all clustering criteria were plotted versus the number of clusters. Looking at pseudo-F test in non-hierarchical K-means cluster analysis, the trend was very similar to that in hierarchical cluster analysis. There were local maxima at 12, 15, 17, 20 clusters with exception at 22 clusters. Due to this discrepancy, 22 clusters seemed inappropriate to be the final solution.

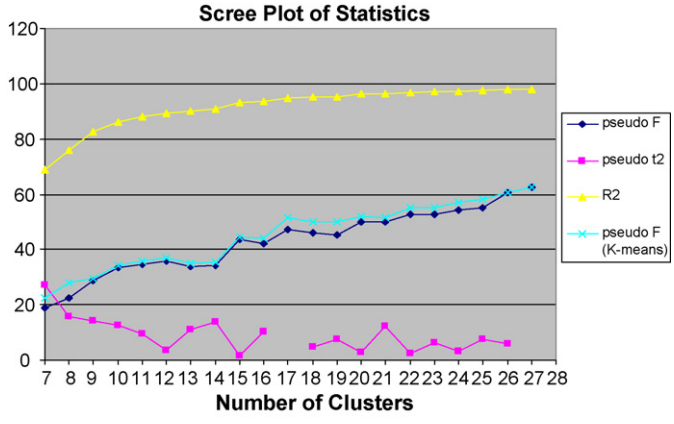

Fig. 2. Scree plot of the statistics in Table 7.

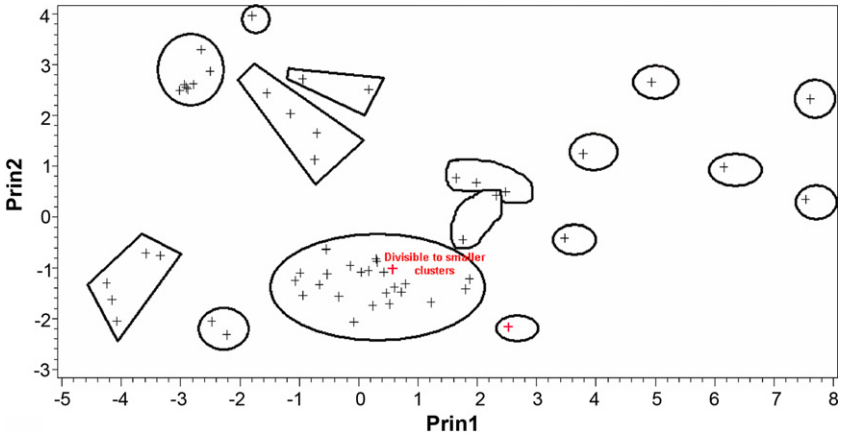

Fig. 3. Plot of PC1 vs. PC2 without solvent ID.

These statistics provided guidelines in choosing the most appropriate partition of clusters. However, they were only indicative and were not as authoritative as "proper" statistical tests. Indeed, one needed to take into account whether the taxonomy given by the chosen partition made any sense from a theoretical viewpoint. It was essential that the clusters formed could be justified by the perceived properties of the members and not just by the cluster analysis procedure.

To better understand the separation between solvent populations, plots of the second principal component Prin2 against the first principal component Prin1 were created. Since the first principal component, Prin1, was defined in the direction of maximum variance in the data set, and the subsequent components were orthogonal to one another and described the maximum of the remaining variance, the first two component scores represented 60% of the total variance (see Table 4). Thus, the plots could reveal a good deal of information about the underlying population structure. Fig. 3 (without solvent ID) and Fig. 4 (with solvent ID) were used for comparison. A manual segmentation (marked in Fig. 3) based on visual distance and solvent chemical properties showed that at least 16 clusters should be formed. In fact this exploratory partitioning not only produced surprisingly good agreement with the final solution from the two-stage cluster analysis, but also effectively excluded 12 and 15 clusters from the potential solution candidates.

The main problem with the 17-cluster solution was that it grouped acetic acid and methanol together, an unexpected grouping that did not make sense based upon the authors' experience. Figs. 3 and 4 also illustrate the large distance between the two solvents (see red crosses and arrows in the Figures). The unavailable pseudo-t^2 value brought additional uncertainty to the 17-cluster solution.

From a typological viewpoint, if the number of clusters becomes too big (>22), not much structure is imposed upon the objects. Hence, the cluster analysis would not reveal much in such a case. On the other hand, if the number of clusters becomes too small (<15), too much "structure" is forced on the data, which may not be in accordance with the actual properties. The combination of all the above analysis resulted in a 20-cluster solution where consensus of different statistics was found in both hierarchical and non-hierarchical analysis.

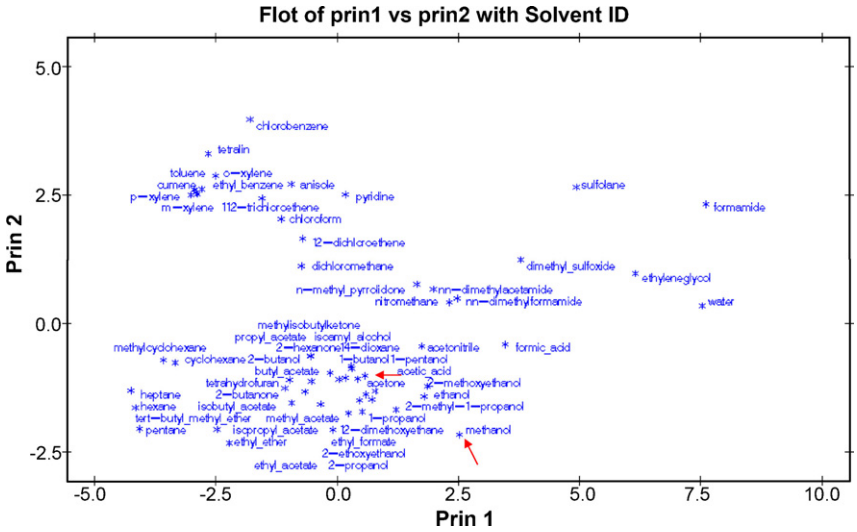

Fig. 4. Plot of PC1 vs. PC2 with solvent ID.

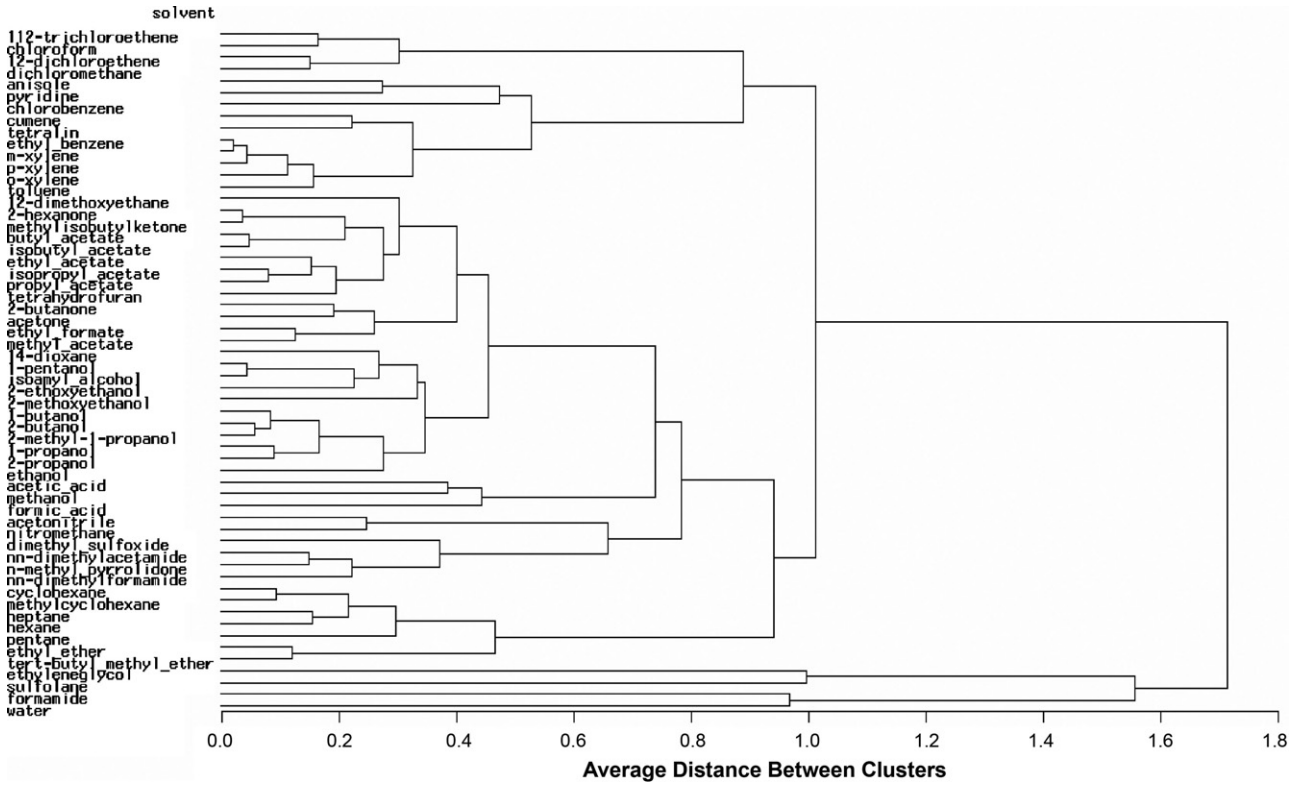

Fig. 5. Tree diagram of the hierarchical clustering procedure.

3.3. Final solution

The dendrogram of 20-cluster hierarchical solution is displayed in Fig. 5, showing the partition creation sequence at different cluster aggregation levels. Table 8 lists membership information of all the clusters obtained from hierarchical average linkage cluster analysis and non-hierarchical *K*-means cluster analysis. When comparing column 2 and 3, the overall agreement between the results from the two different clustering methods is excellent. The only difference is slight membership change between cluster 2 and 7. From a chemical perspective, the *K*-means cluster analysis solution seems more reasonable. It moves ethyl acetate and acetone, two solvents that have smaller hydrophobic components, from cluster 2 where most solvents have longer hydrophobic segments to the more similar cluster 7. This demonstrated when the hierarchical clustering method is complemented by the nonhierarchical method. The resulting clusters selection was refined by maximizing the similarity of the positions of solvents within the clusters. This was evidenced in Table 2 by the higher pseudo-*F* values obtained in non-hierarchical *K*-means procedure than in hierarchical average linkage procedure at all cluster levels.

More detailed information about the clusters was revealed by the non-hierarchical *K*-means procedure in Table 9. The column labeled "Frequency" displays the number of solvents in each cluster, whereas the next column labeled "RMS standard deviation" shows the root mean square standard deviation. Single member cluster implies no or zero standard deviation. Increasing the number of members in a cluster increases the standard deviation. A good cluster has a small within cluster standard deviation. In this respect, all clusters demonstrated relatively low standard deviation.

Maximum distance from seed to observation is the largest distance (Euclidean) from the cluster seed (mean component score) to any observation within the cluster. This measure showed the heterogeneity of the clusters and inter-cluster similarity when comparing to the next column labeled distance between clusters centroids, which displays the distance from the cluster mean to the mean of the nearest cluster. In general, this distance measure is indicative of the between cluster heterogeneity. For a more straightforward illustration, the ratio between these two columns was taken and recoded in the last column. Although column 4 did not give any information about the position of cluster members relative to its centroid, the percent ratio in column 5 could still indicate the likelihood of member movement between two nearest clusters. Two pairs of clusters, namely cluster (2, 7) and cluster (3, 4) stood out, with their ratios being 50% and higher. This means their members may be re-partitioned into the other nearest clusters as long as the change of cluster membership can be explained from chemical perspective or validated by experiments. The movement of ethyl acetate and acetone from cluster 2 to cluster 7 in non-hierarchical *K*-means procedure was a good example. A closer look at the members in cluster 3 and 4 shows all of them are alcohols. It would not be surprising if they were switched between the two clusters due to the high ratio scores (89 and 80%) and chemical similarities.

With a casual inspection of the final solution in Table 8, one can easily find special chemical features in each cluster. The chemical features may be structural similarities, functional

Table 8
Cluster members

Cluster	Hierarchical average linkage cluster analysis	Non-hierarchical *K*-means cluster analysis
1	*m*-Xylene, *o*-xylene, *p*-xylene, ethyl benzene, toluene, cumene, tetralin	*m*-Xylene, *o*-xylene, *p*-xylene, ethyl benzene, toluene, cumene, tetralin
2	Propyl acetate, isobutyl acetate, butyl acetate, isopropyl acetate, methylisobutylketone, 2-hexanone, 12-dimethoxyethane, ethyl acetate, acetone	Propyl acetate, isobutyl acetate, butyl acetate, isopropyl acetate, methylisobutylketone, 2-hexanone, 12-dimethoxyethane
3	2-Ethoxyethanol, isoamyl alcohol, 1-pentanol, 14-dioxane, 2-methoxyethanol	2-Ethoxyethanol, isoamyl alcohol, 1-pentanol, 14-dioxane, 2-methoxyethanol
4	1-Propanol, 2-propanol, 2-methyl-1-propanol, 2-butanol, 1-butanol, ethanol	1-Propanol, 2-propanol, 2-methyl-1-propanol, 2-butanol, 1-butanol, ethanol
5	Hexane, methylcyclohexane, cyclohexane, heptane, pentane	Hexane, methylcyclohexane, cyclohexane, heptane, pentane
6	Ethyl ether, *tert*-butyl methyl ether	Ethyl ether, *tert*-butyl methyl ether
7	Methyl acetate, 2-butanone, tetrahydrofuran, ethyl formate	Methyl acetate, 2-butanone, tetrahydrofuran, ethyl formate, ethyl acetate, acetone
8	*NN*-Dimethylacetamide, *N*-methyl pyrrolidone, *NN*-dimethylformamide	*NN*-Dimethylacetamide, *N*-methyl pyrrolidone, *NN*-dimethylformamide
9	Chloroform, 12-dichloroethene, dichloromethane, 112-trichloroethene	Chloroform, 12-dichloroethene, dichloromethane, 112-trichloroethene
10	Acetonitrile, nitromethane	Acetonitrile, nitromethane
11	Anisole, pyridine	Anisole, pyridine
12	Dimethyl sulfoxide	Dimethyl sulfoxide
13	Acetic acid	Acetic acid
14	Methanol	Methanol
15	Formic acid	Formic acid
16	Chlorobenzene	Chlorobenzene
17	Formamide	Formamide
18	Water	Water
19	Ethyleneglycol	Ethyleneglycol
20	Sulfolane	Sulfolane

group similarities or both. For example, all members in cluster 1 are benzene derivatives with different alkyl groups. Cluster (3, 4), 5 and 6 are obviously alcohol, alkane and ether groups respectively. All members in cluster 9 are alkyl halides (–Cl). No negative influence was observed when the components of the xylene mixture were treated separately. All of them always fell into the same group as their properties are very similar. Despite the seemingly "trivial" solution, the two-stage cluster analysis made important decisions on those not so distinctive solvents. It separated methanol from both alcohol groups (3, 4), isolated chlorobenzene from cluster 1 and cluster 9, and singled out water, acetic acid, ethyleneglycol, sulfolane and formic acid as single component clusters. Two sets of closely related clusters (3, 4) and (2, 7) were successfully isolated by the statistical

Table 9
Cluster summary

Cluster	Frequency	Nearest cluster	Maximum distance from seed to observation	Distance between cluster centroids	Ratio (%)
1	7	16	1.3922	2.5994	53.56
2	7	7	1.2661	1.7949	70.54
3	5	4	1.3264	1.4937	88.80
4	6	3	1.1899	1.4937	79.66
5	5	6	1.1973	2.3911	50.07
6	2	5	0.3306	2.3911	13.83
7	6	2	1.0751	1.7949	59.90
8	3	12	0.7612	1.932	39.40
9	4	16	1.0415	3.4775	29.95
10	2	7	0.6753	2.9982	22.52
11	2	16	0.7479	2.48	30.16
12	1	8	0	1.932	0
13	1	4	0	2.0292	0
14	1	13	0	2.1069	0
15	1	14	0	2.2143	0
16	1	11	0	2.48	0
17	1	12	0	5.2625	0
18	1	17	0	5.3064	0
19	1	20	0	5.4563	0
20	1	12	0	4.1244	0

treatment. These results seem very reasonable from chemical viewpoint and have strong statistical support.

4. Conclusion

An exhaustive review of current literature and databases resulted in the compilation of 231 solvent physiochemical descriptors. Removal of incomplete, redundant or cross correlative descriptors reduced the set to 17 descriptors. Using the dimension reduction procedure of PCA any remaining correlation in the 17 different solvent descriptors was removed and a compact quality data set was generated for cluster analysis. The two-stage cluster analysis produced excellent agreement and consistent results between different statistical clustering techniques. The non-hierarchical *K*-means method yielded slightly refined results than average linkage hierarchical method, but the main trends of the analysis remained unaffected.

Based upon cluster analysis, the 57 classes 2 and 3 solvents recognized by ICH (plus water) have been reduced to a set of 20 clusters. One of the proposed utilities for this reduction is the ability to design more efficient polymorph screens through the judicious inclusion of solvents expected to behave differently and exclusive of solvents expected to behave in a similar fashion to those already studied. For instance, in clusters of several members, testing of a single member would be expected to represent the solvent characterisitics for the group. For group number 2, this represents the testing of a single solvent rather than seven or more. In addition, the inclusion of a member from each different group in a screening study would ensure appropriate solvent diversity. Because the entire tree diagram and statistical design have been shared with the readers, experimentalists may choose to further collapse or increase the number of groups studied based upon individual needs and preferences.

References

Abboud, J.-L.M., Notari, R., 1999. Critical compilation of scales of solvent parameters, IUPAC. Pure Appl. Chem. 71, 645–718.

Abraham, M.H., Whiting, G.S., Doherty, R.M., Shuely, W.J., 1991. Hydrogen bonding XVI. A new solute solvation parameter, π_2^H, from gas chromatographic data. J. Chromatogr. 587, 213–228.

Abraham, M.H., 1993a. Scales of solute hydrogen-bonding: their construction and application to physicochemical and biochemical processed. Chem. Soc. Rev. 22, 73–83.

Abraham, M.H., 1993b. Hydrogen bonding. XXVII. Solvation parameters for functionally substituted aromatic compounds and heterocyclic compounds, from gas-liquid chromatographic data. J. Chromatogr. 644, 95–139.

Abraham, M.H., 1993c. Hydrogen bonding 31 Construction of a scale of solute effective or summation hydrogen-bond basicity. J. Phys. Org. Chem. 6, 660–684.

Abraham, M.H., Chadha, H.S., Whiting, G.S., Mitchell, R.C., 1994. Hydrogen bonding. Part 32. An analysis of water–octanol and water–alkane partitioning, and the DlogP parameter of Seiler. J. Pharm. Sci. 83, 1085–1100.

Abraham, M.H., Rafols, C., 1995. Factors that influence tadpole narcosis. An LFER analysis. J. Chem. Soc. Perkin Trans. 2, 1843–1851.

Bunkers, M.J., Miller, J.R., DeGaetano, A.T., 1996. Definition of climate regions in the Northern Plains using an objective cluster modification technique. J. Clim. 9, 130–146.

Calinksi, T., Harabasz, J.A., 1973. A dendrite method for cluster analysis. Comm. Stat. 3, 1–27.

Carlson, R., 1992. Design and Optimization in Organic Synthesis. Elsevier, Amsterdam.

Cattell, R.B., 1966. The scree test for the number of factors. J. Multiv. Behav. Res. 1, 245–276.

Chastrette, M., Rajzmann, M., Chanon, M., 1985. An approach to a general classification of solvents using a multivariate statistical treatment of quantitative solvent parameters. J. Am. Chem. Soc. 107, 1.

Cooper, M.C., Milligan, G.W., 1988. The effect of measurement error on determining the number of clusters in cluster analysis. In: Gaul, W., Shader, M. (Eds.), Data, Expert Knowledge and Decision. Springer, pp. 319–328.

CRC Press, 2005. CHEMnetBase, Chemical Databases Online, http://www.chemnetbase.com.

Davis, R.E., Kalkstein, L.S., 1990. Development of an automated spatial synoptic climatological classification. Int. J. Climatol. 10, 769–794.

Duda, R.O., Hart, P.E., 1973. Pattern Classification and Scene Analysis. Wiley, New York.

Eder, B.K., Davis, J.M., Bloomfield, P., 1994. An automated classification scheme designed to better elucidate the dependence of ozone on meteorology. J. Appl. Meteor. 33, 1182–1199.

Gordon, A.D., 1996. Cluster Validation. Data Science, Classification, and Related Methods. Springer, Tokyo.

Gu, C.H., Li, H., Gandhi, R.B., Raghavan, K., 2004. Grouping solvents by statistical analysis of solvent property parameters: implication to polymorph screening. Int. J. Pharm. 283, 117–125.

Jurs, P.C., Chou, J.T., Yuan, M., 1979. Studies of chemical structure-biological activity relations using pattern recognition. In: Olson, R.C., Christoffersen, R.E. (Eds.), Computer-Assisted Drug Design. American Chemical Society, Washington, DC, pp. 103–129.

Kaiser, H.F., 1960. The application of electronic computers to factor analysis. Educ. Psychol. Meas. 20, 141–151.

Kalkstein, L.S., Tan, G., Skindlov, J.A., 1987. An evaluation of three clustering procedures for use in synoptic climatological classification. J. Clim. Appl. Meteorol. 26, 717–730.

Lide, D.R., 2005. Handbook of Chemistry and Physics, 86th ed. CRC Press, Boca Raton, FL.

Marcus, Y., 1993. The properties of organic liquids that are relevant to their use as solvating solvents. Chem. Soc. Rev. 22, 409–416.

Mirmehrabi, M., Rohani, S., 2005. An approach to solvent screening for crystallization of polymorphic pharmaceuticals and fine chemicals. J. Pharm. Sci. 94, 1560–1576.

Milligan, G.W., 1980. An examination of the effect of six types of error perturbation on fifteen clustering algorithms. Psychometrika 45, 325–342.

Milligan, G.W., Cooper, M.C., 1985. An examination of procedures for determining the number of clusters in a data set. Psychometrika 50, 159–179.

Milligan, G.W., 1996. Clustering validation: results and implications for applied analyses. In: Clustering and Classification. World Scientific Publishers, New Jersey, 341-375.

Pearlman, R.S., 1980. In: Yalkowsky, S.H., Sinkula, A.A., Valvani, S.C. (Eds.), Physical Chemical Properties of Drugs. Marcel Dekker, New York, pp. 321–347.

Punj, G., Stewart, D.W., 1983. Cluster analysis in marketing research; Review and suggestions for application. J. Mark. Res. 20, 134–148.

Raevsky, O.A., 1997. Hydrogen bond description in framework of multiplicative approach. J. Phys. Org. Chem. 10, 369–376.

Raevsky, O.A., Skvortsov, V., 2002. 3D hydrogen bond thermodynamics (HYBOT) potentials in molecular modeling. J. Comput. Aided Mol. Des. 16, 1–10.

Riddick, J.A., Bunger, B.B., Sakano, T.K., 1986. Organic Solvents: Physical Properties of Purification, fourth ed. Wiley-Interscience.

Romesburg, H.C., 1990. Cluster Analysis for Researchers. Krieger Publishing Co., Malabar, FL.

Rubin, D.B., 1987. Multiple Imputation for Nonresponse in Surveys. Wiley, New York.

SAS OnlineDoc®, 2000. Version 8, the CLUSTER, FASTCLUS and PRINCOMP procedures. SAS Institute Inc.

Snyder, L.R., 1978. Classification of the solvent properties of common liquids. J. Chromatogr. Sci. 16, 223–234.

Sokal, R.R., Michener, C.D., 1958. A statistical method for evaluating systematic relationships. Univ. Kans. Sci. Bull. 28, 1409–1438.

Stanton, D.T., Jurs, P.C., 1990. Development and use of charged partial surface area descriptors in computer-assisted quantitative structure–property relationship studies. Anal. Chem. 62, 2323–2329.

Stanton, D.T., Egolf, L.M., Jurs, P.C., Hicks, M.G., 1992. Computer-assisted prediction of normal boiling points of pyrans and pyrroles. J. Chem. Inform. Comput. Sci. 32, 306–316.

Stanton, D.T., Dimitrov, S., Grancharov, V., Mekenyan, O.G., 2002. Charged partial surface area (CPSA) descriptors. QSAR applications. SAR QSAR Environ. Res. 13, 341–351.

Stanton, D.T., Mattioni, B., Knittel, J.J., Jurs, P.C., 2004. Development and use of the hydrophobic surface area (HAS) descriptors in computer-assisted quantative structure–activity and structure-property relationship studies. J. Chem. Inform. Comput. Sci. 44, 1010–1023.

Stuper, A.J., Jurs, P.C., 1976. ADAPT: a computer system for automated data analysis using pattern recognition techniques. J. Chem. Inform. Comput. Sci. 16, 99–105.

TriMen Chemicals Co., 2005. Physiochemical Properties of Popular Liquids, http://www.trimen.pl/witek/ciecze/liquids.html.

U.S. FDA, 1997. Companion Document for the International Conference on Harmonisation of Technical Requirements for Registration of Pharmaceuticals for Human Use (ICH) Guidance for Industry Q3C Impurities: Residual Solvents, http://www.fda.gov/cber/gdlns/ichq3ctablist.htm.

Winget, P., Dolney, D.M., Giesen, D.J., Cramer, C.J., 1999. Minnesota Solvent Descriptor Database, http://comp.chem.umn.edu/solvation/mnsddb.pdf.

Zelenka, M.P., 1997. Analysis of the meteorological parameters affecting ambient concentrations of acid aerosols in Uniontown, Pennsylavania. Atmos. Environ. 31, 869–878.