

Molecular Diversity Sample Generation on the Basis of Quantum-Mechanical Computations and Principal Component Analysis

H. Gutiérrez-de-Terán^{a,†}, J. J. Lozano^{a,#,†}, V. Segarra^b and F. Sanz^{a,*}

^aResearch Group on Biomedical Informatics, IMIM, Universitat Pompeu Fabra, C/ Dr. Aiguader 80, E-08003 Barcelona, Spain

^bAlmirall Prodesfarma, S.A., C/ Cardener 68-74, E-08024 Barcelona, Spain

Abstract: The present study introduces a new strategy of selection of a maximum diversity sample of n compounds from N available in a molecular database. This strategy can be useful in pharmacological screening, combinatorial chemistry or parallel synthesis planning. It consists of first describing the compounds by means of parameters derived from quantum mechanical computations (water solvation G , benzene solvation G , octanol solvation G , dipolar moment), as well as standard molecular parameters such as solvent-accessible surface area and molecular weight. Solvation parameters are used because of the importance of this phenomenon in the pharmacological behaviour. Redundant information in the description of the compounds is eliminated by using principal components (PC) instead of the original descriptors. Based on the similarity between the N compounds in the PC space, they are classified into n groups by k -means cluster analysis. The compounds that are nearest to the centroid of each cluster constituted the maximum diversity sample. When practical difficulties exist for the use of one of the proposed compounds, another also close to the cluster centroid can substitute for it. This strategy has been tested in the selection of a sample of 50 amines from the 923 available in the Aldrich catalogue. The results have been contrasted with those obtained from an optimal, distance-based experimental design, resulting in an 86% of agreement between both approaches. An R^2 -like diversity coefficient has been used to assess the quality of the proposed solutions.

INTRODUCTION

The selection of an assorted sample of compounds from those available in a molecular database is a typical task in pharmacological screening, combinatorial chemistry or parallel synthesis planning. For instance, in the framework of a lead optimization process, it is a common practice to systematically explore a certain position of a lead compound by testing the introduction of a series of substituents having diverse molecular properties. The introduction of combinatorial chemistry techniques has extended and accelerated such exploration. Such efforts are often targeted at producing a large series of structurally diverse molecules to be biologically tested through high throughput screening techniques in order to identify potential new drugs [1]. Despite the relatively reduced cost of parallel synthesis techniques, the selection of a relevant sample among all the possible substituents is required in order to adequately explore the experimental space using the minimum number of compounds [2]. The substituents introduced should produce the maximum coverage of the molecular properties,

avoiding the repetition of information. The primary aim of the mentioned sampling is to reduce the excessively large number of potential starting materials with the aim of generating a manageable synthetic plan [3]. For this purpose, we introduce in this article a new strategy of selection of compounds, which is based on their description by means of principal components obtained from variables including molecular weight, solvent-accessible surface area (SASA) and physicochemical descriptors estimated by quantum mechanical computations. Because of the simplicity of its computational process (PCA and cluster analysis) and the accessibility of the required software (a standard statistical package such as SPSS [4]), the present approach differs from other valuable approaches recently proposed, such as those based on self-organizing or Kohonen neural networks [5,6]. Our new approach has been tested in the selection of a sample of amines from those available in the Aldrich catalogue. This study constituted part of a real lead optimisation process in the framework of a drug development project.

The selection of representative members from a library of compounds in order to optimally cover the chemical diversity space is based on (usually computed) chemical properties [7]. With this purpose, many molecular descriptors have been proposed and thoroughly reviewed [8,9]. They can be classified in three groups:

- 1) Scalar molecular properties, which do not depend on a particular 3D conformation of the molecules, such as the molecular weight.

*Address correspondence to this author at the Research Group on Biomedical Informatics, IMIM, Universitat Pompeu Fabra, C/ Dr. Aiguader 80, E-08003 Barcelona, Spain; Phone: +034 932 257 587; Fax: +034-932-213-237; E-mail: fsanz@IMIM.ES

[†]J.J. Lozano and H. G. de Terán contributed equally to this work.

[#]Current address: Dep. of Physiology and Biophysics, Mount Sinai School of Medicine, New York, USA.

- 2) Topological descriptors (“2D descriptors”), which rely on the molecular formulae, such as the connectivity indices.
- 3) Descriptors derived from 3D molecular structures, such as the dipolar moment.

In molecular diversity studies within the pharmaceutical field, it is usual to employ physicochemical descriptors [9], since they are often related to pharmacologically relevant molecular interaction phenomena (transport, ligand-receptor recognition, etc.) [10, 11].

As the information contained in each molecular descriptor is somehow correlated with the information contained in the other ones, principal components analysis (PCA) is a useful tool to eliminate redundant information [12], and has been often used to derive new orthogonal descriptors to be applied in molecular diversity studies [11,13-15]. The main principal components (PCs) allow a useful visualization of the diversity of a molecular library, and they can be used as input for maximum diversity sampling methods.

The methods for the selection of maximum diversity samples described in the literature use different computational techniques, such as hierarchical and non-hierarchical cluster analysis [16], genetic algorithms [17], neural networks [5,6], clique detection [18] or specifically designed methods like the most descriptive compound (MDC) method [19]. Random selection of the intended sample of compounds has been demonstrated to be less effective than the aforementioned rational algorithms [8,19]. Among them, in the present study we propose the use of non-hierarchical k-means cluster analysis [20], which will be compared with a distance-based algorithm [21].

METHODS

Molecular Library

The 923 primary amines included in the Aldrich catalogue were used in the present study. A molecular database containing their formulae (in sdf format) was available.

3D Structures Generation

The aforementioned molecular database containing the sdf planar coordinates of the 923 amines was first transformed into 923 independent files. The 923 formulae contained in the files were transformed into feasible 3D structures using the CORINA software [22], which is one of the fastest and more efficient software for such a purpose [23].

Taking into account that the selected amines will be used for generating substituents in a certain position of a molecular scaffold, the most meticulous approach would imply the substitution of one of the hydrogens of the amino group by the scaffold or a molecular fragment representing it. However, this operation could be neglected since the values of the molecular descriptors would be influenced in a similar

way by a molecular fragment shared by all the compounds, as showed previously by others [11]. This argument might be erroneous in the case of vectorial properties, such as the dipolar moment (DM) employed in the present study. To test the reliability of our structural simplification on this property, we recomputed the DM for nine amines homogeneously distributed along the dipolar moment range. In this second computation, we substituted one of the hydrogens of the amino group by a pyrazinyl moiety that represented the common scaffold of the parallel synthesis process that motivated the present analysis. The comparison of both dipolar moment values, those of the unsubstituted amines and those of the corresponding extended compounds, is presented in the Results section.

Molecular Descriptors

The 923 3D molecular structures were used as input for quantum-mechanical computations, which were carried out at the AM1 level using AMSOL 6.3 software [24]. Molecules were fully optimised geometrically as solvated in water, using the SM5.4A solvation model [25]. Starting from the resulting geometries, single point computations were carried out to simulate the species solvated in octanol and benzene. Consequently, we compiled information about free energies of solvation in polar (water) and two non-polar (octanol and benzene) solvents.

The execution of quantum chemical computations produced strange results for 177 molecules because of one of the following reasons:

- a) They contained atoms such as Hg, Se, or B that make the amines uninteresting as drug-like fragments.
- b) The molecules were too large to be processed using quantum-mechanical (QM) methods.
- c) They were hydrochlorides, hydrobromides or other salts unsuitable for QM methods. On the other hand, the neutral forms of these compounds were also included in the database as different entries.
- d) In the case of enantiomers, which generate the same values of the physico-chemical descriptors considered, we selected only one of each pair.

The filtering of the above-described compounds could have been carried out by simple inspection of every compound included in the Aldrich database. Nevertheless, the present protocol allows for a certain automation of the process.

For the remaining 746 amines, six molecular descriptors were generated through the AMSOL computations. Two descriptors dealt with steric information (molecular weight and solvent-accessible surface area), one dealt with electronic information (dipolar moment), and three with solubility in solvents having different characteristics (free energies of solvation in water, benzene and octanol). The AMSOL program generates directly all these molecular descriptors

Table 1. Bivariate Correlation Matrix Between the Original Descriptors

	DM	MW	SASA	G_H ₂ O	G_Bz
MW	0.18				
SASA	-0.15	0.75			
G_H ₂ O	-0.44	-0.08	-0.01		
G_Bz	-0.19	-0.72	-0.76	-0.55	
G_Oct	-0.35	-0.46	-0.45	-0.87	-0.88

with no need of additional software, thereby simplifying the process.

Consequently, an initial data matrix having a 746 x 6 dimension was generated. As it is shown in the bivariate correlation matrix (see Table 1), all the molecular descriptors are correlated in different degrees. On the other hand, since they are expressed in different units, they cannot be mixed in certain computations (i.e., Euclidean distances).

Principal Components Analysis

To solve the aforementioned problems of co-linearity between descriptors and the subsequent information redundancy, we propose the use of principal components (PCs) as molecular descriptors instead of the original ones [11,12-15]. The SPSS software [4] was used for this part of the analysis. The relevant number of PCs to be considered was determined on the basis of the percentage of the total variability of the data that had to be retained. The scores were standardized for the considered compounds in the PC space. A Varimax rotation of the selected PCs is advisable to obtain the maximum separation between the contributions of the original descriptors to each PC. The rotation aims to obtain high loadings for a subset of the original variables in each PC, while having the minimum overlap between such subsets (see Fig. 1). The orthogonality of the resulting PCs

implies a total lack of information redundancy in the new variables describing the compounds.

Sampling Methods

The high diversity sample of n compounds was selected by a k-means cluster analysis [20] on the basis of the rotated PCs previously obtained. K-means is a non-hierarchical clustering method that starts with an arbitrary set of n centroids (vectors of descriptors), assigns the compounds to the closest centroid, recomputes each centroid as the mean of the vectors of descriptors of the corresponding compounds, and repeats iteratively the process until a stable classification is obtained. The compound closest to each final centroid is proposed as cluster representative. In the approach used in the present study, the initial set of centroids was randomly generated. The sample size (50 in this study) can be decided on the basis of statistical or feasibility criteria [16]. A critical aspect of k-means cluster analysis is its dependence of the initial set of centroids. We have studied the importance of this problem by comparing the results obtained from two different randomly generated initial sets of centroids.

In order to compare the proposed approach with another based on different algorithms, we used the longest minimum distance algorithm (LMD) developed by Marengo and Tordeschini [21]. The aim of this algorithm is iteratively

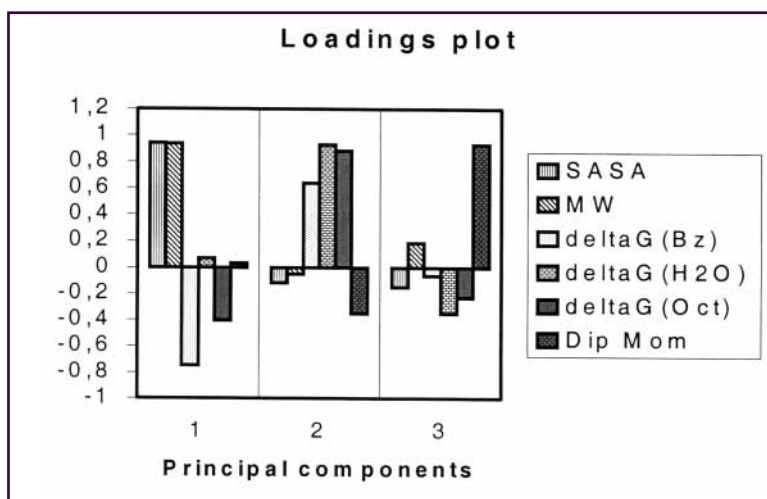


Fig. (1). PCA loadings plot. The contributions of each original molecular descriptor on each PC are depicted.

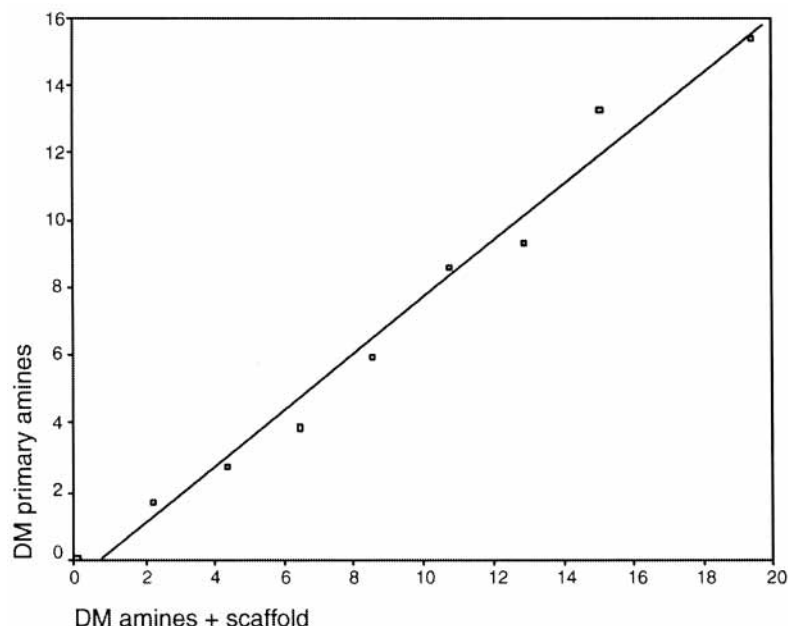


Fig. (2). Dipolar moment comparison: primary amines vs. secondary amines obtained by pyrazinyl substitution.

generating an n -sized sample of compounds having the largest possible distances between them. For this analysis, we employed the Q^2 software [26] using standardized PCs without applying the Varimax rotation, which is not allowed in this software.

Diversity Measures

In this kind of study, it is useful to assess the resulting solutions using diversity functions [27]. For such a purpose we used an R^2 -like diversity coefficient defined as follows:

$$\text{Total diversity prior clustering} = D_T = \sum_{i=1}^N d_{ic}^2$$

$$\text{Diversity lost within the clusters} = D_{WC} = \sum_{j=1}^n \sum_{i=1}^{n_j} d_{ic_j}^2$$

$$R^2 = \frac{\text{Remaining diversity after clustering}}{\text{Total diversity}} = \frac{D_T - D_{WC}}{D_T}$$

being d_{ic} Euclidean distances computed in the PCs space, N and n the number of compounds of the whole set and the sample respectively, c the centroid of the whole set of compounds, and n_j and c_j the sizes and the centroids of the n clusters respectively. A refined, but computationally more difficult, version of the present coefficient would consist in defining c_j as the coordinates of the compound representing each cluster instead of its centroid. In the present study we have used the first definition.

RESULTS AND DISCUSSION

Structural Simplification

As explained in the Methods section, we evaluated the effect on the dipolar moment (MD) of the structural simplification consisting of not substituting one of the amino hydrogens by the relevant scaffold. The scatter plot of both MD values (simplified vs. extended structures) is presented in Fig. (2). The plot and the corresponding R^2 value, which is equal to 0.980, show a good relationship between both series of values.

Original Descriptors

Table 2 summarizes the values of the original descriptors in the entire dataset. Their distributions are clearly asymmetrical. Table 1 shows the correlations between such descriptors.

The descriptors look appropriate for drug discovery projects. For instance, free energies of solvation are effective in distinguishing molecules on the basis of their polarity [28], a feature that has well known importance in the pharmacological behaviour. Another parameter related to molecular polarity is the dipolar moment (DM). On the other hand, properties related to the molecular size, here represented by solvent-accessible surface area (SASA) and molecular weight (MW), are also known to affect biological behaviour. Thus, their ensemble may be useful as diversity descriptors in the analysis of series of drug candidates [29].

Principal Component Analysis (PCA)

In the present case, considering the first three PCs retained 95% of the original variability of the compounds.

Table 2. Descriptive Statistics of the Original Descriptors

Descriptor	Maximum	Minimum	Mean	Median	SD
DM	19.39	0.08	4.48	3.82	3.04
MW	515.35	31.04	164.29	155.07	56.39
SASA	902.35	201.74	390.75	372.38	87.64
G(H ₂ O)	0.44	-45.91	-10.11	-8.56	6.37
G(Bz)	-2.86	-26.09	-9.93	-9.43	3.32
G(Oct)	-4.50	-36.15	-12.63	-11.63	4.96

As can be seen in the loadings plot shown in Fig. (1), the orthogonal PCs resulting from the Varimax rotation had a clear interpretation:

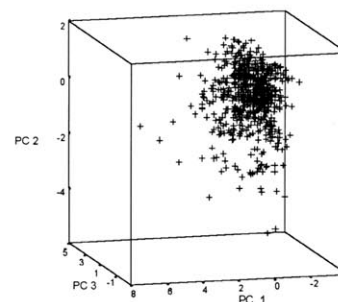
- PC1 was mainly constituted by steric information since the original variables that presented the loadings with the highest absolute values were solvent-accessible surface area (SASA) and molecular weight (MW).
- PC2 mainly included solvation information since the original variables that showed the coefficients with highest absolute values were the free energies (G) of solvation in water, benzene and octanol.
- PC3 was mainly related to electrostatic information since the only original variable that resulted in a high loading in this component was the dipolar moment (DM).

The resulting descriptors (PCs) defined a standardized orthogonal three-dimensional space, which keeps 95% of the original information on the compounds.

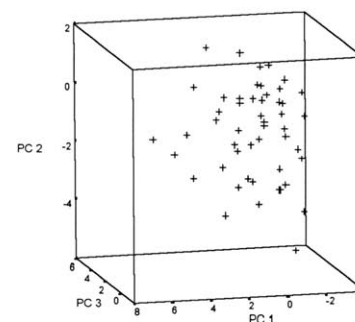
The distribution of the whole dataset in the space defined by the three PCs is visualized in the scores plot shown in Fig. (3a). It is worth noting that the density of compounds is greater in the intervals that correspond to low or moderated values of the PCs, contrasting with the low density of compounds having high values. This indicates that the initial set was not uniformly distributed in the molecular properties space.

Selection of Compounds

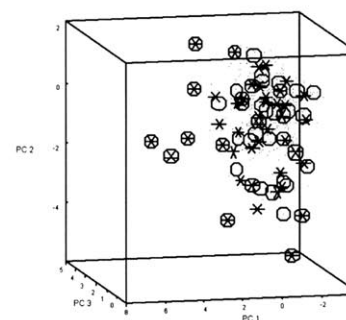
Starting from a set of 50 randomly generated centroids, the k-means cluster analysis produced a stable classification of the studied set of compounds into 50 groups. Obviously, the most populated clusters were those situated in the aforementioned high-density region, whereas the clusters created in the extreme zones were less populated or even single-compound groups. A high degree of chemical similarity between compounds belonging to the same cluster was observed by visual inspection. Even in the highly populated clusters, i.e., cluster 43 having 92 compounds-, a single representative of each cluster should be selected to



(a)



(b)



(c)

Fig. (3). Scatter plots in the PCs' space: (a) the 746 amines studied, (b) the 50 compounds selected by k-means cluster analysis, (c) superposition of the 50 compounds selected from two different runs of k-means cluster analysis (circles and asterisks); the entire initial population is represented by small points).

maximize the diversity of the final set [30]. The compound nearest to each final centroid was proposed as cluster representative. Fig. (3b) shows the position of each representative in the PCs space. Fig. (3c) depicts the comparison of the previous results with those obtained using an alternative initial set of centroids. Both solutions are coincident in the areas of the PCs' space with low density of compounds, whereas they differ in the particular compounds selected in the zones with high density of compounds. In

any case, the quality of both solutions in terms of R^2 -like coefficient were the same (0.923).

To compare the solutions produced by the strategy above, another maximum diversity sample of 50 compounds was selected using an alternative distance based algorithm, the Longest Minimum Distance (LMD) method. This approach contains an iterative procedure analogous to the exchange algorithm used in D-optimal design, but using

Table 3. Comparison Between k-Means and LMD Approaches

k-means				LMD	
Cluster	Size	Compound	Dist. to centroid	Compound	Dist. to k-means centroid
1	40	HGT150	0.170	HGT111	0.676
2	11	HGT096	0.140	HGT093	0.328
3	1	HGT860	0.000	-	-
4	1	HGT331	0.000	HGT331	0.000
5	6	HGT784	0.192	HGT348	0.948
6	6	HGT079	0.245	HGT080	0.532
7	14	HGT873	0.121	HGT672	0.605
7				HGT517	0.717
8	36	HGT757	0.153	HGT889	0.515
9	3	HGT550	0.135	HGT921	0.655
10	1	HGT853	0.000	HGT853	0.000
11	47	HGT311	0.093	-	-
12	4	HGT338	0.105	HGT699	0.403
12				HGT410	0.568
13	7	HGT028	0.304	HGT028	0.304
14	12	HGT076	0.285	HGT091	0.578
15	9	HGT001	0.436	HGT842	0.798
15				HGT833	0.837
16	1	HGT545	0.000	-	-
17	11	HGT469	0.172	HGT469	0.172
18	14	HGT160	0.165	HGT662	1,112
19	25	HGT043	0.188	HGT278	0.593
20	30	HGT287	0.159	HGT217	0.674
20				HGT559	0.710
21	32	HGT465	0.190	HGT915	0.676
22	1	HGT074	0.000	HGT074	0.000
23	4	HGT846	0.209	HGT547	0.635
24	4	HGT845	0.224	HGT877	0.303
25	1	HGT220	0.000	HGT220	0.000
26	4	HGT528	0.279	HGT654	0.748

(Table 8). contd.....

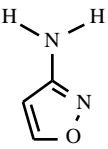
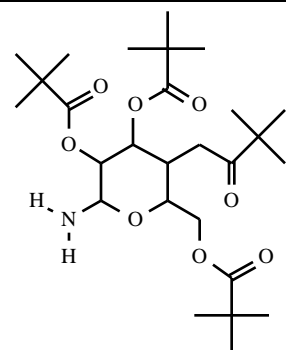
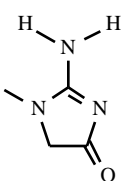
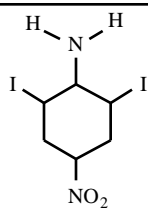
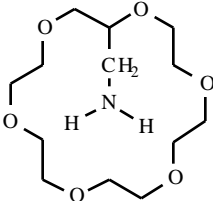
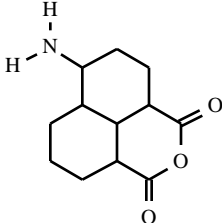
k-means				LMD	
Cluster	Size	Compound	Dist. to centroid	Compound	Dist. to k-means centroid
27	36	HGT693	0.114	HGT472	0.664
28	15	HGT823	0.171	HGT405	0.522
28				HGT356	0.540
29	24	HGT374	0.107	HGT389	0.501
30	22	HGT265	0.062	-	-
31	2	HGT755	0.474	HGT755	0.474
32	1	HGT066	0.000	HGT066	0.000
33	23	HGT121	0.318	-	-
34	17	HGT772	0.209	HGT535	0.513
34				HGT189	0.901
35	1	HGT409	0.000	HGT409	0.000
36	11	HGT330	0.370	HGT109	0.418
37	8	HGT092	0.291	HGT077	0.582
38	19	HGT378	0.173	HGT394	0.942
39	1	HGT522	0.000	HGT522	0.000
40	2	HGT897	0.340	-	-
41	1	HGT881	0.000	HGT881	0.000
42	4	HGT529	0.231	HGT027	0.930
43	92	HGT851	0.074	HGT495	0.681
44	42	HGT413	0.098	-	-
45	4	HGT171	0.399	HGT171	0.399
46	10	HGT308	0.169	HGT309	0.460
46				HGT557	0.917
47	13	HGT118	0.147	HGT319	0.706
48	2	HGT700	0.227	HGT035	0.227
49	31	HGT370	0.210	HGT864	0.633
50	40	HGT022	0.100	HGT796	0.638

distance considerations with the purpose of obtaining designs well distributed in the variable space [21]. The correspondence between the representatives obtained here and the groups generated with the k-means algorithm is shown in Table 3. This table shows the size of each cluster, the compounds selected by k-means and LMD that belong to the same cluster, as well as the distances to the corresponding centroids. In seven cases, LMD proposed two compounds belonging to the same k-means cluster. Vertical bold lines in Table 3 indicate them. Consequently, since the sample size was the same in both approaches, there were seven k-means clusters without LMD representative. In general, the clusters that contained two compounds proposed by LMD were the most populated. In such cases, it has to be pointed out that

both representatives showed long Euclidean distances to their centroid and relatively short distances to that of the nearest cluster non-represented by LMD. An 86% of agreement between both methods was found. Agreement was defined as the percent of clusters generated by k-means analysis that included at least one compound generated by the alternative LMD algorithm. Finally, it has to be pointed out that the solution obtained with the LMD method had the same R^2 -like diversity coefficient (0.923) as the results obtained with the strategy postulated in the present study.

As an example of the compounds selected as cluster representatives, the formulae of those having extreme values of each PC are shown in Table 4. It can be appreciated that

Table 4. Cluster Representatives with Extreme Values on the Principal Components

PC	Lowest value		Highest value	
	Cluster	Compound	Cluster	Compound
PC1 (steric)	48		10	
PC2 (solvation)	31		25	
PC3 (dip. moment)	41		39	

the proposed compounds have completely different values of the molecular properties represented by each PC. Thus, the first line of the table shows two compounds having very diverse steric characteristics, while the last line shows two compounds with very different charge balances. The two compounds with extreme values of the second PC, despite being diverse, are more difficult to interpret, which is logical taking into account that this PC is a mix of solvation thermodynamics in solvents exhibiting different polarities.

CONCLUSIONS

We have developed and tested a fast, semiautomatic distance-based strategy for the selection of a maximum diversity sample of compounds from a relatively large initial dataset. This approach uses chemical information that is relevant for the biological behaviour and avoids information heterogeneity and redundancy by using standardized Varimax-rotated principal components instead of the original molecular descriptors.

ACKNOWLEDGMENT

The authors thank Dr. M. Pastor for helpful discussions. CIRIT is also acknowledged for the grant provided to one of the authors.

LIST OF ABBREVIATIONS

- PCA = Principal components analysis
 MW = Molecular weight
 SASA = Solvent-accessible surface area
 DM = Dipolar moment
 $G(\text{H}_2\text{O})$ = Free energy of solvation in water
 $G(\text{Bz})$ = Free energy of solvation in benzene
 $G(\text{Oct})$ = Free energy of solvation in octanol
 LMD = Longest minimum distance

REFERENCES

- [1] Special issue on "Computational methods for the analysis of molecular diversity", *Perspect. Drug Discov. Design*, **1997**, 7/8.
- [2] Newton, C.G. In *Molecular diversity in drug design*; Dean, P. M.; Lewis, R. A. Eds.; Kluwer Academic Publishers: Dordrecht, **1999**; Ch 2, pp. 23-42.

- [3] Gund, P.; Sigal, N.H. *Immunol. Today*, **1999**, 2(Suppl. 1), 25.
- [4] SPSS 9.0, SPSS Inc. Chicago.
- [5] Polanski, J.; Jarzembek, K.; Gasteiger, J. *Comb. Chem. High Throughput Screen.*, **2000**, 3, 481.
- [6] Ros, F.; Audouze, K.; Pintore, M.; Chretien, J.R., *SAR QSAR Environ. Res.*, **2000**, 11, 281.
- [7] Spellmeyer, D.C.; Grootenhuis, P.D.J. *Ann. Rep. Med. Chem.*, **1999**, 34, 287.
- [8] Gillet, V.J. In *Molecular diversity in drug design*; Dean P. M.; Lewis, R. A. Eds.; Kluwer Academic Publishers: Dordrecht, **1999**; Ch 3, pp. 43-65. And references there in.
- [9] Downs, G. M.; Willett, P. *J. Chem. Inf. Comput. Sci.*, **1994**, 34, 1094.
- [10] Kubinyi, H. *QSAR: Hansch analysis and related approaches*; VCH: Weinheim, **1993**.
- [11] Gibson, S.; McGuire, R.; Rees, D.C. *J. Med. Chem.*, **1996**, 39, 4065.
- [12] Joliffe, I.T.; Morgan, B.J.T. *Stat. Methods Med. Res.*, **1992**, 1, 69.
- [13] Langer, T.; Hoffman, R.D. *Quant. Struct. -Act. Relat.*, **1998**, 17, 211.
- [14] Eriksson, L.; Johansson, E. *Chemom. Intell. Lab. Sys.*, **1996**, 34, 1.
- [15] Oprea, T.; Gottfries, T., ChemGPS, a chemical space navigation tool. 13th European Symposium on QSAR; Düsseldorf, **2000**.
- [16] Voght, W.; Nagen D.; Sator, H. *Cluster analysis in clinical chemistry: a model*; J Wiley & Sons, Ltd.: Chichester, **1987**.
- [17] Sheridan, R.P., SanFeliciano, S.G.; Kearsley, S.K.; *J. Mol. Graph.*, **2000**, 18, 320.
- [18] Gardiner, E. J.; Holliday, J.D.; Willet, P.; Wilton, D.J.; Artymiuk, P.J. *Quant. Struct.-Act. Relat.*, **1998**, 17, 232.
- [19] Hudson, B. D.; Hyde, R.M.; Rahr, E.; Wood, J. *Quant. Struct.-Act. Relat.*, **1996**, 15, 285.
- [20] Späth, H. *Cluster-Analyse-Algorithm zur Objektklassifizierung und Datenreduction*, 2nd revised Ed., R. Oldenbourg Verlag: Munich, **1977**.
- [21] Marengo, E.; Todeschini, R. *Chemom. Intell. Lab. Sys.*, **1992**, 16, 37.
- [22] Gasteiger, J.; Rudolph, C.; Sadowski, J. *Tetrahedron Comp. Method.*, **1990**, 3, 537.
- [23] Sadowski, J.; Gasteiger, J.; Klebe, G., *J. Chem. Inf. Comput. Sci.*, **1994**, 34, 1000.
- [24] Hawkins, G.D.; Giesen, D.J.; Lynch, G.C.; Chambers, C.C.; Rossi, I.; Storer, J.W.; Li, J.; Rinaldi, D.; Liotard, D.A.; Cramer C.J.; Truhlar D. G., AMSOL-version 6.5.2, University of Minnesota, Minneapolis, **1997**.
- [25] Chambers, C.C.; Hawkins, G.D.; Cramer, C.J.; Truhlar, D. G. *J. Org. Chem.*, **1996**, 61, 8720.
- [26] Q2 4.5. Multivariate Infometric Analysis srl. Perugia, Italy, **1999**.
- [27] Hassan, M.; Bielawski, J.P.; Hempel, J.C.; Waldman, M. *Mol. Divers.*, **1996**, 2, 64.
- [28] Cummins, D. J.; Andrews, C.W.; Bentley, J.A.; Cory, M. *J. Chem. Inf. Comput. Sci.*, **1996**, 36, 750.
- [29] Patterson, D. E.; Cramer, D.E.; Ferguson, A.M.; Clark, R.D.; Weinberger, L.E. *J. Med. Chem.*, **1996**, 39, 3049.
- [30] Bailey, D.S.; Furness, L.M.; Dean, P.M. *Immunol. Today*, **1999**, 2(Suppl. 1), 6.