A Novel Approach Using Hierarchical Clustering To Select Industrial Chemicals for Environmental Impact Assessment

Stefan Rännar and Patrik L. Andersson*

Department of Chemistry, Umeå University, SE-901 87 Umeå, Sweden

Received August 31, 2009

We propose a four-step strategy, based on principal component analysis and hierarchical clustering, for selecting structurally dissimilar organic substances from a list of commercial, high volume production chemicals. The selection strategy also presents alternative structures with similar characteristics, so that practical aspects of future testing can be easily addressed. The selected compounds listed in this Article are intended for further study regarding their environmental impact and as potential pollutants.

INTRODUCTION

Selecting subsets of substances from large chemical libraries is a common task undertaken in various fields of chemical research, for example, drug discovery,[1,2] model development in quantitative structure−activity or property relationship (QSAR/QSPR)[3−5] analyses, and environmental research.[6−8] Although diverse strategies are used for this purpose, two distinctly different approaches are generally employed when screening libraries: similarity and dissimilarity. The similarity approach[6,9,10] is based on the assumption that chemically similar molecules are more likely than dissimilar molecules to have similar biological activity. This approach can be used when looking for new active compounds that resemble a known active substance and is useful for lead-optimization and in the development of predictive structure−activity models. By contrast, the dissimilarity approach is used when it is necessary to search a large region of chemical space to find lead molecules with similar activities, but which are structurally varied.[2] It is also used (in the context of interest here) when it is important to ensure that the selected substances are representative of a targeted chemical domain.

In scientific projects aimed at developing predictive models of chemicals' attributes, it is important to ensure that both the training sets and the test sets of substances to be used in the development and validation of the models are representative of the domain of interest.[11−13] However, due to the costs and difficulties involved in measuring relevant responses in drug and environmental research, model calibration and testing is often performed on a limited number of compounds. Nevertheless, it is still very important to select these few compounds according to criteria that ensure they are appropriately diverse and represent the domain of interest.[8,14−16] In addition to the relevance of structural diversity and representativeness in the development of predictive models, it is equally important to select test substances used in comparisons of analytical procedures in appropriate ways to ensure that the developed method or protocol can be properly verified.[17]

Achieving a good selection of substances from a given set of chemicals can be accomplished using various techniques,[12−14,18−20] but multivariate characterization of some of the molecules' physicochemical characteristics is usually involved. The list of possible candidate substances often comprises thousands of chemicals, for which some relevant chemical and physical attributes are not known. In fact, very little is known in particular about the biological and environmental properties, such as biodegradability and toxicity, of a large proportion of commercially available chemicals. To avoid restricting our selection of substances from those that have already been extensively studied, we, and many other researchers, have to rely on calculated rather than measured values for the properties we deem to be important. Even though these calculated values may not be as accurate as measured values, they do provide us with a set of comparable numbers for many substances within a reasonable time. However, merely having relevant data is insufficient; to draw meaningful information from the resulting matrices of data on diverse chemical and physical attributes of large numbers of compounds, analysis by appropriate multivariate tools such as principal component analysis (PCA)[21] or self organizing maps (SOM)[22] is required.

PCA has often been used in multivariate chemical characterizations, for various purposes,[19,23] to elucidate a few independent, latent variables that summarize the information contained in calculated descriptor variables. These latent variables can then be used, not only in the interpretation and elucidation of chemical variation, but also as an excellent basis for selecting a representative set of chemicals when using manual or automated approaches, often based on some sort of statistical design. The selection strategy is usually followed by an investigation of the chemical reactivity or biological activity of a selected subset of chemicals. Strategies may differ between situations, but typical approaches used for selecting diverse chemicals include factorial designs,[24] D-optimal designs,[25] and D-optimal Onion designs,[26,27] each of which has been applied by various authors in a number of applications.[8,16,23,28−31]

The selected sets generated by many commonly applied statistical design tools often include substances that are

* Corresponding author phone: +46907865266; fax: +46907867655; e-mail: patrik.andersson@chem.umu.se.

10.1021/ci9003255 © 2010 American Chemical Society
Published on Web 01/06/2010

extreme in terms of their chemical properties, and sometimes unsuitable substances are suggested for further study. These inappropriate results can arise from the selection algorithm that has been used (e.g., D-optimal) because, mathematically, the substances that best cover a defined chemical domain are likely to be positioned furthest from each other within the chemical space. A means to avoid this obstacle is to employ a modified and more elaborate polynomial in the selection procedure, which, however, leads to a larger training set and thus a need to test more compounds. One can also avoid selecting the most extreme substances by trimming, for example, the D-optimal candidate set based on the score vector values. Alternatively, one can apply a selection made by space-filling designs, such as the Onion design, which, in addition to substances with extreme attributes, will also suggest substances from the interior of the domain covered by the PC-model. Even here it can be advantageous to remove the most extreme substances before deploying the selection algorithm. A feature common to most design strategies is that they present the researcher with only one substance from each section of the chemical space. However, this is often insufficient because of various constraints, such as toxicity, interfering with the studied system, analytical difficulties, or commercial availability. If these additional constraints disqualify suggested compounds from selection, one or more alternative substances are necessary for each selection. The model space then needs to be searched for these alternatives, in the vicinity of the candidate substances that have already been suggested, by using some sort of similarity measurement;[32] a search that often has to be done manually. It can be argued that substances disqualified by the constraints mentioned should be removed from the candidate set before design selection. Yet this is often not practical because these properties typically need to be researched manually for each compound and therefore will be too labor intensive for the complete candidate list. Our solution to this is to first make a well design selection (with alternatives) and then investigate this reduced list thoroughly.

We present here a novel strategy to select training set compounds for future environmental impact studies. The methodology is intended to identify chemically diverse and representative substances for specific purposes and to present alternatives for each region of the chemical domain. This is accomplished using a four-step approach that identifies representative (but not too extreme) substances and alternatives.

The four steps in the approach are step 1, chemical mapping; step 2, clustering; step 3, cluster modeling and core selection; and step 4, final substance selection.

This approach has recently been utilized in an ongoing project funded by the Swedish Environmental Protection Agency entitled “Organic Chemicals Emitted from Technosphere Articles” (Chemitecs),[33] which focuses on estimating the amounts, and the environmental impact, of organic substances emitted from technosphere articles. In an initial phase of the project, case-studies have been selected for use in the development and testing of models, ranging from emission estimation to the human perception of exposure. A case-study in this project is not defined as simply an individual substance, but rather a substance in a specific object, in combination with the normal use pattern of that object. Thus, for the case-study selection, not only are the chemical aspects of the individual compounds of interest, but also a number of contextual aspects of the selected objects. This is typically the case where constraints on substance selection are added later, when many candidate substances from each part of the chemical space will be required. In the present Article, we describe only how the chemical aspects of the selection were addressed.

Because the aim of the “Chemitecs” project is to examine emissions into the environment of organic chemicals from goods and objects used by society, an obvious starting point is to compile a candidate list of common, commercially used organic chemicals. It was therefore decided that the starting list of candidate compounds would consist of chemicals that are used in appreciable quantities in production industries. Substances of lower production volumes can, of course, also cause environmental problems, but most of the environmentally problematic chemicals have, at some time, been produced in high quantities.[34] In the present study, we have therefore developed and applied a methodology for selecting sets of chemicals that are defined in Europe as high production volume chemicals (HPVCs) and low production volume chemicals (LPVCs).

MATERIALS AND METHODS

List of Substances. The starting list of candidate substances was comprised of chemicals defined by the European Chemicals Agency as being HPVC (>1000 tonnes/year) or LPVC (10−1000 tonnes/year),[35] where volumes include total production within, and imports into, the European Union.

The lists, extracted in March 2008, initially generated a total of 10 614 chemicals. However, because chemical descriptors can only be calculated from well-defined substances, the list was reduced by the removal of mixtures, polymers, and other chemicals that are not unique substances. All inorganic substances were also removed, and organic salts (salts with Na, K, Ca) were converted to their corresponding neutral forms. The final list was comprised of 1341 HPVC and 5316 LPVC unique organic compounds.

Chemical Descriptors. All substances were characterized by their chemical properties calculated using MOE software.[36] A total of 68 descriptors, based on the substances’ 2D-structures, were computed and selected according to their relevance and interpretability. These descriptors are identical to those used by Stenberg et al.,[7] and a complete list, with brief explanations, is given in the Supporting Information (Table S1). The descriptors include traditional QSAR descriptors (e.g., log K_{ow}, solubility, polarizability, van der Waals volume) in combination with selected flexibility, shape, and connectivity indices. Molecular surface characteristics were represented by 16 “partial equalization or orbital electronegativity” (PEOE) descriptors. Counts of atoms, and single and aromatic bonds, were also used together with some count ratios.

All descriptors (except those already log-transformed) were log-transformed prior to analysis to normalize their distributions and minimize the influence of extreme values. Because of the particular shape of its distribution curve, the descriptor “PEOE_PC-” was transformed using a negative logarithmic transformation.

To describe the different clusters, descriptors reflecting environmentally interesting properties, for example, biodegradability and atmospheric oxidation half-life, were calculated using the BIOWIN software included in the EPI-suite.[37]

Principal Component Analysis (PCA). PCA is a latent variable method that compresses all of the information in a data matrix into a few orthogonal vectors to summarize the variation and correlation patterns in the data. It is a very useful tool for visualizing and interpreting large data matrices. Typically, two to five principal components (PCs) are sufficient to represent most of the structured information in a multivariate table. Each PC consists of two vectors: a score-vector, which graphically displays similarities among the substances (or observations), and a loading-vector, which facilitates interpretation of the trends and correlations among the descriptors (variables).

Traditionally, both score- and loading-vectors are presented as plots with two components in each; for example, score-vector 1 (t1) is plotted versus score-vector 2 (t2) and loading-vector 1 (p1) versus loading-vector 2 (p2). In the present study, the multivariate analysis was calculated in SIMCA-P+ v12[38] software.

Hierarchical Clustering. Clustering algorithms are used to group similar objects based on a similarity, or distance, criterion set by the user. The algorithm used here is based on Ward clustering,[39] which analyzes the increase in error when two similar objects or clusters are merged. In SIMCA-P+ v12, the algorithm is applied to the score-vectors from a PC-model, and the error function calculates the difference in the sum of sum of squares around the mean of each cluster, before and after merging two clusters. The number of clusters has to be defined by the user, either based on a cluster distance threshold or, as in this study, a practical limit. In the present study, we decided that eight substances was a reasonable number of objects for further study in the "Chemitecs" project, and therefore we arbitrarily set the number of clusters to eight.

RESULTS AND DISCUSSION

Step 1: Chemical Mapping. The descriptor table for all HPVCs and LPVCs (totaling 6657 chemicals and 68 descriptors) was subjected to PCA to compress the data into a small number of latent variables that were then used to map and visualize the chemical space. Four chemically interpretable components were extracted, which together explained 75% of the total variance and resulted in a homogeneous swarm of observations in the score-plots. The first two components in the score-plot are shown in Figure 1, in which each substance is represented by a black triangle, and substances that are close together in the score-plot are deemed to be similar in terms of their descriptor values.

The first component, explaining 41% of the variance, is interpreted as a "molecular-size" vector with strong loadings for descriptors such as van der Waal's volume and area, molecular weight, and connectivity indices. Substances found to the right in Figure 1 are the largest, with typically more than 30−40 carbon atoms, often with large aromatic systems and molecular weights of more than 700 Da. The substances with the smallest molecular structures, for example, ethylene and methylamine, are found to the left in the figure.

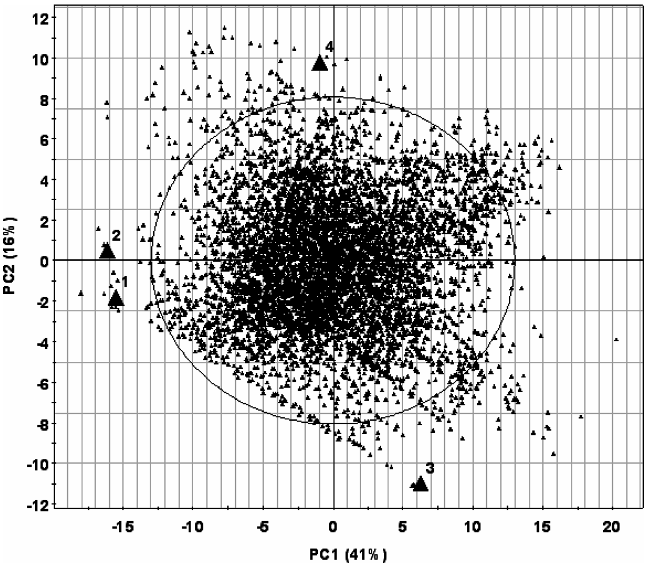


Figure 1. The chemical space formed by PCA of commercial, high volume, organic substances: PC1 versus PC2 score plot (accounting for 41% and 16% of the total variation in the data set, respectively). Each triangle corresponds to a chemical in the original list. Four molecules mentioned in the text are indicated with larger symbols and numbers: (1) ethylene, (2) methylamine, (3) dioctadecyl disulfide, and (4) nitrilotrimethylenetrisphosphonic acid.

The descriptors with the strongest loadings in the second component, explaining 16% of the variance, are various measures of polarity, for example, PEOE descriptors describing hydrophobic surface area and polar surface area, and relative numbers of hydrogen bond donors and acceptors (Table S4). In addition, hydrophobicity (log K_{ow}), with high relevance for, for example, environmental fate, has a large influence on the second component. In Figure 1, hydrophobic compounds like dioctadecyl disulfide, with a calculated log K_{ow} of 17, are found in the lower part of the plot, while hydrophilic substances like nitrilotrimethylenetrisphosphonic acid (log K_{ow}, −4.7) have high score-values in the second component and are found in the upper part of the figure.

The third component, explaining 12% of the variance, describes aromaticity (i.e., the molecules' numbers of aromatic bonds and atoms), and the fourth component, explaining 6.0% of the variance, is related to the degree of halogenation, where substances with most halogens have the highest loading. Notably among the 6657 substances, approximately 1500 are halogenated and 204 have more than two halogens. Many of these, for example, perfluorooctane and hexachloroethane, have high values in the fourth score-vector. Decabromodiphenyl ether is a typical example of a substance with high values in both the third and the fourth components because it is a highly halogenated aromatic compound.

Step 2: Clustering. The next step, following the PC-modeling, is to cluster the observations into a number of groups on the basis of their closeness in the chemical map (and hence similarity in terms of the descriptors). Applying a clustering algorithm to the four score-vectors, the substances considered here were divided into eight cluster groups based on practical aspects of the "Chemitecs" project. Had the aim of the clustering been to differentiate between chemical classes, or to classify new molecules, the number of clusters and their statistical separation would have been of great importance. However, for the purpose of selecting a diverse set of chemicals, the number of clusters depends more on practical aspects such as the available budget for conducting laboratory tests. Our selection of clusters therefore resulted in the selection of eight candidate substances for further study.

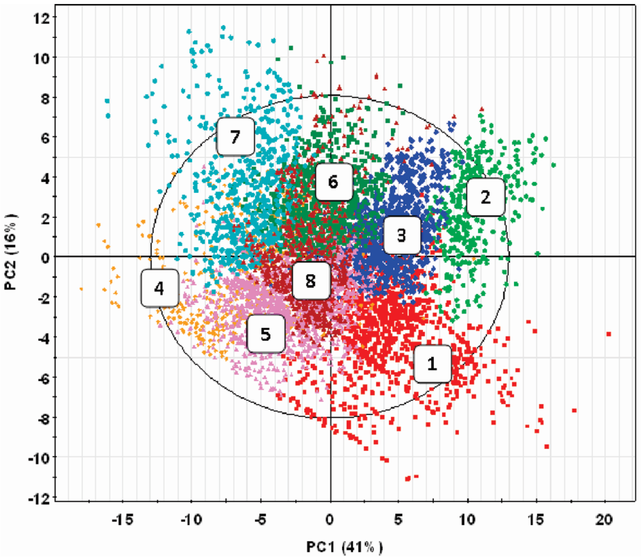


Figure 2. Score-plot (PC1 versus PC2) of the chemical space with the eight clusters marked in different colors and indicated with their respective cluster numbers.

The eight clusters varied in size between cluster 2, comprising 412 substances, and cluster 5 with 1395 substances. Figure 2 depicts the clusters in the score-plot of the first two dimensions.

For environmental studies, there are a number of properties describing biological fate and persistence that can be calculated using the EPI Suite program. Using BIOWIN3 in EPI Suite, which estimates the time-scale of ultimate biodegradation of chemicals, it was found that 319 (77%) of the 412 substances in cluster 2 were predicted to be persistent. The criterion used was a numerical value of less than 2, which corresponds to ultimate biodegradation in a time-scale of months or longer. A high percentage (42%) of potentially persistent molecules was also present in cluster 3. Eighty-nine percent of all substances with calculated log K_{ow} values greater than 5 (a common criterion for assessing bioaccumulation potential) are found in clusters 1, 2, and 3. Many persistent molecules are halogenated, another environmentally important chemical descriptor. In cluster 4, 79% of the 614 substances contain at least one halogen, and 51% have two or more halogens. Cluster 4, which is largely obscured by clusters 5 and 8 in Figure 2, is dominated by PC4, which is strongly influenced by the presence of halogens.

From the mapping of the calculated values of environmentally relevant properties, it seems reasonable to assume that many chemicals, already known to be environmentally problematic, would be positioned in one of the first four clusters. To test this assumption, 12 chemicals listed by the United Nations Environment Program (UNEP) as persistent organic pollutants (POPs) were classified into the eight clusters. The probability of cluster membership was calculated for all substances using the distance between the substance and each cluster's principal component model. This distance to model (DModX) should be interpreted as the molecular similarity between a test substance and the group of substances constituting the cluster. Using a 95% probability threshold for membership, six of them (Toxaphene, DDT, Hexachlorobenzene, PCB-153, 2378-TCDF, 2378-TCDD) were assigned membership of cluster 4, and the remaining six (Aldrin, Mirex, Chlordane, Dieldrin, Endrin, and Heptachlor) were not classified; that is, these are not found to be statistically similar to any of the clusters. However, looking at the calculated distance, they were found closest to cluster 4. All of the UNEP-chemicals are fairly hydrophobic and have at least five halogens, which explain their similarity to other substances found in cluster 4. One important structural feature that separates the nonclassified from the classified UNEP chemicals is aromatic rings. All classified, except Toxaphene, have aromatic rings, while the nonclassified have not. Notably, Toxaphene has the lowest cluster 4 membership probability among the classified substances (0.2 as compared to at least 0.9 for the others).

Step 3: Cluster Models and Core Selection. Once all of the observations in the eight clusters had been identified, the data set was divided into separate tables, and the substances closest to the center of each cluster were selected. This can be done manually, but it is difficult when more than two PCs have to be considered simultaneously. A better approach is to calculate individual cluster PC-models (which also have four components) and then to use the Euclidean distance from the model origin to select the innermost, or core, observations. These core substances will be similar to each other, and dissimilar to core substances from all other clusters. In Figure 3a, the 10 substances chosen to represent cluster 5, that is, the most central, or typical substances of this cluster, are indicated in the score-plot for the cluster model. The molecular structures are given in Figure 3b.

The researcher decides the exact number of core substances to be selected from each cluster. However, it should not exceed approximately 10%−20% of the total cluster count because it is desirable to have sufficient distance between each set of core observations and those of adjacent clusters to avoid any overlapping of selections. In the present Article, we do not provide details of the selected core compounds, but 10 candidates from each cluster (and hence 80 compounds in total) are listed in the Supporting Information (Table S2). Acquiring such a list of core observations is the objective of the approach presented in the present Article, while the final selection (step 4) requires more knowledge than that related merely to the chemical variation and representativeness of substances, which forms the general part of the strategy.

To illustrate the variation within and between clusters, we have visualized two molecular structures from each cluster (Table 1). From these examples, it is clear that the two substances drawn from each cluster are structurally more similar to each other than to those drawn from other clusters. The two substances representing cluster 1 both have long aliphatic chains, a property reflected in the average log K_{ow} of the cluster, which is larger than that of cluster 8. Large molecules with an average molecular weight greater than 800, and several aromatic rings, dominate cluster 2. These are usually sold as salts. It should be remembered, however, that these organic salts were converted to their protonated form before calculations and analyses were performed. Cluster 4 contains relatively small molecules, most of which have an aromatic ring and one or more halogens (chlorine being the most abundant), here exemplified by α-chloro-4-fluorotoluene and 2-chlorobenzoyl chloride. Cluster 7 comprises small, nonaromatic molecules, often containing some oxygen and nitrogen atoms, typified in the selected examples

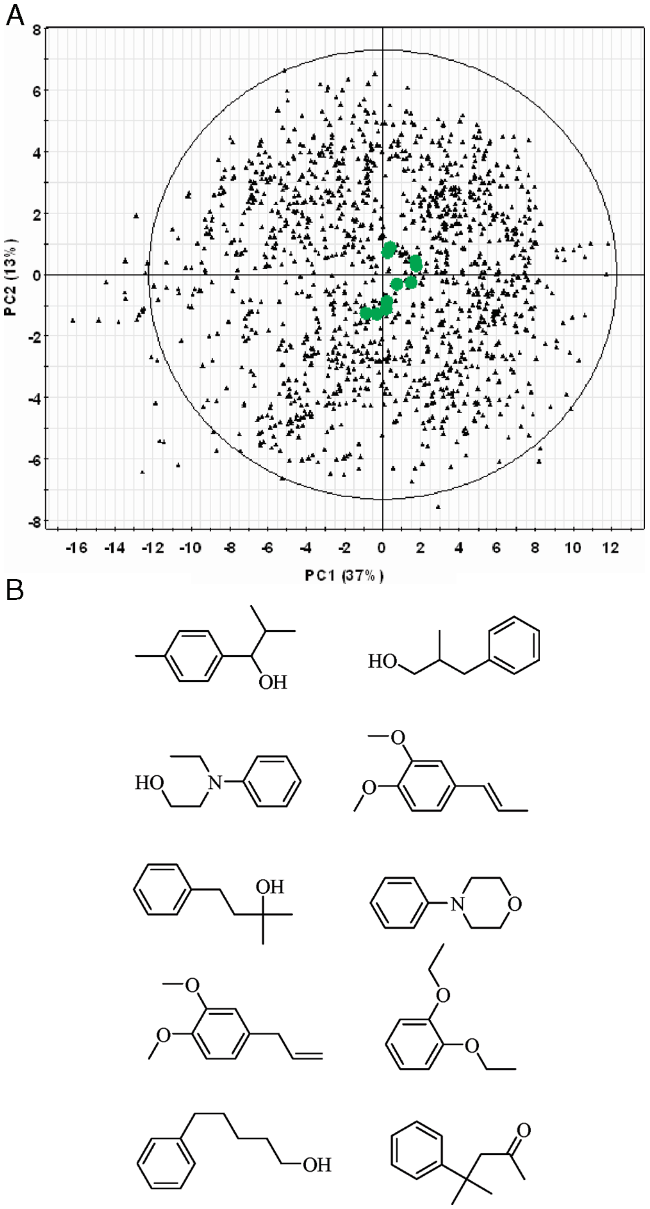
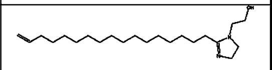
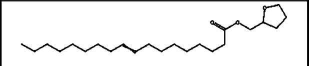
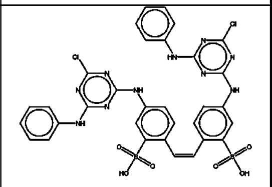
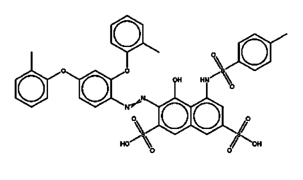
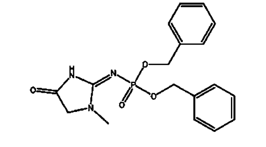
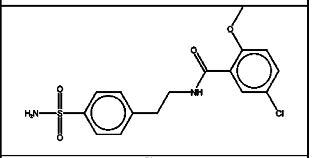
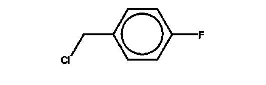
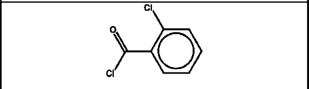
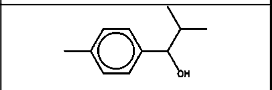
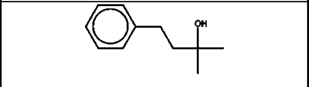
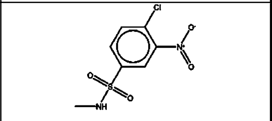
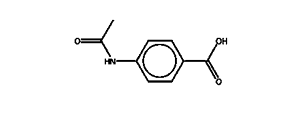
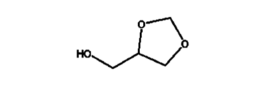
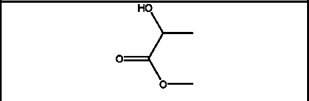
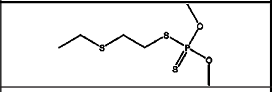
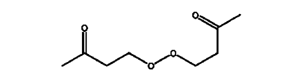


Figure 3. (A) Score-plot (PC1 versus PC2 score plot, accounting for 37% and 13% of the total variation, respectively) of the cluster 5 data set, with the 10 selected core observations indicated by green dots. (B) Molecular structures of the 10 alternative core substances from cluster 5.

from this cluster by 1,3-dioxolan-4-ylmethanol and methyl (*S*)-(−)-lactate.

Step 4: Final Substance Selection. Equipped with the list of core observations from step 3, and knowledge of what the selected training set is to be used for, the final selection (which should include at least one candidate substance from each cluster) can be made. Because the listed core observations from each cluster are considered to be similar in their multivariate chemical properties, any substance from the list will be suitable. This last step in the procedure is also where aspects other than those already captured by the chemical descriptors are introduced. Typical aspects include commercial availability, analytical procedures, and other practical aspects of what the training set is to be used for. In addition, selecting more than one substance from each cluster in the testing procedure can facilitate evaluation of the results by adding a measure of repeatability to the investigation; several candidate substances from each cluster can be treated as replicate objects of the cluster. The extra costs involved in performing this procedure should, however, be weighed against the benefits of adding clusters.

Table 1. Two Examples of Subset Selections Where All Eight Clusters Are Represented

Cluster	Example 1	Example 2
1		
2		
3		
4		
5		
6		
7		
8		

The clustering method presented in this Article is only one of many possible data-sectioning techniques available for dividing a list of substances into groups, but the use of PCA should be regarded as fundamental to the strategy. However, an alternative method of performing data-sectioning is to use factorial design. For a PC-model with four components, this would mean dividing the chemical space into 16 sections, one for each “quadrant”, and assigning all of the observations to one of these quadrants based on their score-values. As mentioned above, in the present project, 16 substances were considered to be too many, so this number would necessarily have been reduced to eight. When using the statistical design approach, a fractional factorial design would have to be used, which will lead to lack of representation of some quadrants in the final selection.

To our best knowledge, there are no means to quantitatively assess the success of various training set selection tools in terms of chemical diversity and representativity. In general, we believe that any strategic protocol would enhance the outcome of screening procedures. A careful selection forms the basis for development and in-depth understanding from structure−property relationships and ultimately QSARs. Nevertheless, we have here compared the novel cluster-based approach with the application of D-optimal design. For

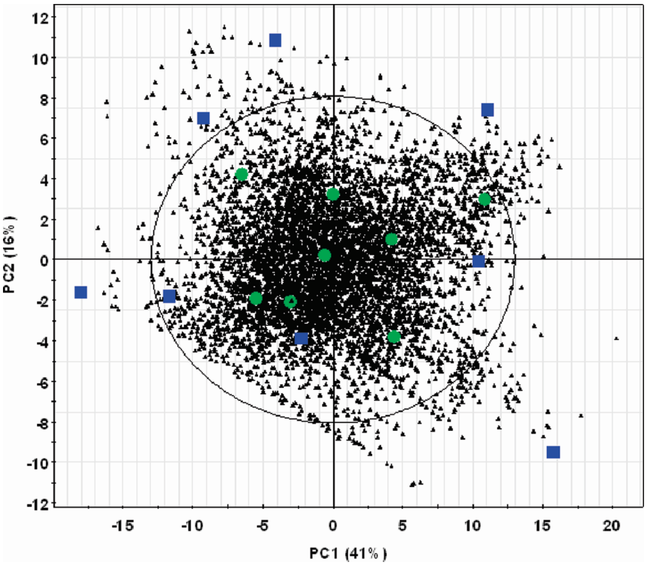


Figure 4. Score-plot based on the first two PCs including all HPVC and LPVC chemicals, in which chemicals selected by the cluster-based methodology (example 1 molecules from Table 1) are indicated by green dots, and eight substances selected using a D-optimal algorithm are indicated by blue squares.

comparison, the molecules listed in "example 1" in Table 1 were selected as candidates of the cluster-based approach. As seen in Figure 4, the chemical domains spanned by the two approaches clearly differ from each other; the substances selected by D-optimal design are located on the edge of the chemical space, while the cluster-based selection is found near to the center, although it still covers a large portion of the domain of interest (note that the selection was done in four dimensions and the results are consistent over all dimensions). This means that the D-optimal approach may yield exceptional chemicals with extreme chemical properties, while the cluster-based approach seems to result in a more representative set of chemicals (note the discussion on complexity of D-optimal algorithms as given in the Introduction). In addition, the region covered by the cluster-based method will be virtually the same regardless of which of the alternative compounds are selected from each cluster.

CONCLUSIONS

In this study, we have presented a novel approach for selecting representative substances from a large set of possible candidates. By using the suggested procedure, two very important issues are addressed: the most structurally extreme substances are avoided, and possible alternative substances are presented. In chemical and biological testing, it is important to have alternative substances to choose between, because a number of aspects of the substances always have to be considered, and extreme substances may be difficult to test and evaluate comprehensively because of their extreme chemical attributes.

The approach is based on PCA and resembles approaches that apply statistical design to latent variables, but our approach differs because in cases where traditional designs suggest a single substance, the proposed procedure gives a group of candidate substances. Substance selection is accomplished by first dividing the compounds into clusters, and then selecting the core compounds from individual group PC models. The variation between clusters ensures that the interpretations and models developed from a training set selected in this fashion will be valid for a wide variety of substances. By applying the four-step strategy (PC modeling, data clustering, PC cluster modeling, and substance selection), the suggested set of substances will not be, in mathematical terms, the best set, but it will be sufficient for spanning a large share of the chemical space and will be more likely to be of practical use.

The clustering approach is particularly appealing, not only because it covers all sections of the mapped, chemical space, but because it also works well when the mapped space is not spherical.

In addition to presenting a novel tool for selecting training set compounds, we propose 80 candidates from a highly relevant group of chemicals for future mapping and testing of their environmental impact. Data describing a large proportion of the most frequently used chemicals are currently lacking, and systematic testing of their chemical and biological attributes is warranted to increase our understanding of properties such as persistence, bioaccumulation, and toxicity. We recommend selecting substances from each cluster of chemicals for in-depth investigations, laboratory tests, screening, and as a basis for future QSAR/QSPR model development.

ACKNOWLEDGMENT

The Swedish Environmental Protection Agency is acknowledged for their funding of the Chemitecs project within which the presented research has been performed.

Supporting Information Available: Table S1: List of calculated chemical descriptors.

Table S2: List of the 80 selected substances.

Table S3: List of explained variances for principal component model.

Table S4: List of the 10 most influential factors in each principal component.

Table S5: List of HPVC/LPVC substances used in the investigation.

This material is available free of charge via the Internet at http://pubs.acs.org.

REFERENCES AND NOTES

(1) Shoichet, B. K. Virtual screening of chemical libraries. *Nature* **2004**, *432*, 862–865.

(2) Taylor, R. Simulation analysis of experimental design strategies for screening random compounds as potential new drugs and agrochemicals. *J. Chem. Inf. Comput. Sci.* **1995**, *35*, 59–67.

(3) Golbraikh, A.; Tropsha, A. Predictive QSAR modeling based on diversity sampling of experimental datasets for the training and test set selection. *J. Comput.-Aided Mol. Des.* **2002**, *16*, 357–369.

(4) Gramatica, P.; Giani, E.; Papa, E. Statistical external validation and consensus modeling: A QSPR case study for K_{oc} prediction. *J. Mol. Graphics Modell.* **2007**, *25*, 755–766.

(5) Öberg, T. A QSAR for Tth hydroxyl radical reaction rate constant: validation, domain of application, and prediction. *Atmos. Environ.* **2005**, *39*, 2189–2200.

(6) Brown, T. N.; Wania, F. Screening chemicals for the potential to be persistent organic pollutants: A case study of arctic contaminants. *Environ. Sci. Technol.* **2008**, *42*, 5202–5209.

(7) Stenberg, M.; Linusson, A.; Tysklind, M.; Andersson, P. L. A multivariate chemical map of industrial chemicals - Assessment of various protocols for identification of chemicals of potential concern. *Chemosphere* **2009**, *76*, 878–884.

(8) Knekta, E.; Andersson, P. L.; Johansson, M.; Tysklind, M. An overview of OSPAR priority compounds and selection of a representative training set. *Chemosphere* **2004**, *57*, 1495–1503.

(9) Willett, P.; Barnard, J. M.; Downs, G. M. Chemical similarity searching. *J. Chem. Inf. Model.* **1998**, *38*, 983–996.

(10) Auer, J.; Bajorath, J. Emerging chemical patterns: A new methodology for molecular classification and compound selection. *J. Chem. Inf. Model.* **2006**, *46*, 2502–2514.

(11) Golbraikh, A.; Shen, M.; Xiao, Z.; Xiao, Y.-D.; Lee, K.-H.; Tropsha, A. Rational selection of training and test sets for the development of validated QSAR models. *J. Comput.-Aided Mol. Des.* **2003**, *17*, 241–253.

(12) Pötter, T.; Matter, H. Random or rational design? Evaluation of diverse compound subsets from chemical structure databases. *J. Med. Chem.* **1998**, *41*, 478–488.

(13) Tropsha, A.; Gramatica, P.; Gombar, V. K. The importance of being earnest: Validation is the absolute essential for successful application and interpretation of QSPR models. *QSAR Comb. Sci.* **2003**, *22*, 69–77.

(14) Schultz, T. W.; Netzeva, T. I.; Cronin, M. T. D. Selection of data sets for QSARs: Analyses of *Tetrahymena* toxicity from aromatic compounds. *SAR QSAR Environ. Res.* **2003**, *14*, 59–81.

(15) Öberg, T. Virtual screening for environmental pollutants: structure-activity relationships applied to a database of industrial chemicals. *Environ. Sci. Technol.* **2006**, *25*, 1178–1183.

(16) Stenberg, M.; Andersson, P. L. Selection of non-dioxin-like PCBs for *in vitro* testing on the basis of environmental abundance and molecular structure. *Chemosphere* **2008**, *71*, 1909–1915.

(17) Sköld, C.; Winiwarter, S.; Wernevik, J.; Bergström, F.; Engström, L.; Allen, R.; Box, K.; Comer, J.; Mole, J.; Hallberg, A.; Lennernäs, H.; Lundstedt, T.; Ungell, A.-L.; Karlén, A. Presentation of a structurally divers and commercially available drug data set for correlation and benchmarking studies. *J. Med. Chem.* **2006**, *49*, 6660–6671.

(18) Wu, W.; Walczak, B.; Massart, D. L.; Heuerding, S.; Erni, F.; Last, I. R.; Prebble, K. A. Artificial neural networks in classification of NIR spectral data: Design of the training set. *Chemom. Intell. Lab.* **1996**, *33*, 35–46.

(19) Eriksson, L.; Andersson, P. L.; Johansson, E.; Tysklind, M. Megavariate analysis of environmental QSAR data. Part I - A basic framework founded on principal component analysis (PCA), partial least squares (PLS), and statistical molecular design (SMD). *Mol. Diversity* **2006**, *10*, 169–186.

(20) Snarey, M.; Terrett, N. K.; Willet, P.; Wilton, D. J. Comparison of algorithms for dissimilarity-based compound selection. *J. Mol. Graphics Modell.* **1997**, *15*, 372–385.

(21) Jackson, J. E. *A User's Guide to Principal Components*; John Wiley & Sons, Inc.: New York, NY, 1991.

(22) Kohonen, T. *Self-Organizing Maps*; Springer: New York, NY, 1995.

(23) Harju, M.; Andersson, P. L.; Haglund, P.; Tysklind, M. Multivariate physicochemical characterisation and quantitative structure-property relationship modeling of polybrominated diphenyl ethers. *Chemosphere* **2002**, *47*, 375–384.

(24) Box, G. E. P.; Hunter, W. G.; Hunter, J. S. *Statistics for Experimenters*; John Wiley & Sons, Inc.: New York, NY, 1978.

(25) De Aguiar, P. F.; Bourguignon, B.; Khots, M. S.; Massart, D. L.; Phan-Than-Luu, R. D-optimal designs. *Chemom. Intell. Lab.* **1995**, *30*, 199–210.

(26) Olsson, I. M.; Gottfries, J.; Wold, S. D-optimal onion designs in statistical molecular design. *Chemom. Intell. Lab.* **2004**, *73*, 37–46.

(27) Olsson, I. M.; Gottfries, J.; Wold, S. Controlling coverage of D-optimal onion designs and selections. *J. Chemom.* **2004**, *18*, 548–557.

(28) Harju, M.; Hamers, T.; Kamstra, J. H.; Sonneveld, E.; Boon, J. P.; Tysklind, M.; Andersson, P. L. Quantitative structure-activity relationship modeling on *in vitro* endocrine effects and metabolic stability involving 26 selected brominated flame retardants. *Environ. Toxicol. Chem.* **2007**, *26*, 816–826.

(29) Papa, E.; Fick, J.; Lindberg, R.; Johansson, M.; Gramatica, P.; Andersson, P. L. Multivariate chemical mapping of antibiotics and identification of structurally representative substances. *Environ. Sci. Technol.* **2007**, *41*, 1653–1661.

(30) Andersson, P. L.; Öberg, K.; Örn, U. Chemical characterization of brominated flame retardants and identification of structurally representative compounds. *Environ. Toxicol. Chem.* **2006**, *25*, 1275–1282.

(31) Eriksson, L.; Johansson, E.; Müller, M.; Wold, S. On the selection of the training set in environmental QSAR analysis when compounds are clustered. *J. Chemom.* **2000**, *14*, 599–616.

(32) Nikolova, N.; Jaworska, J. Approaches to measure chemical similarity - a review. *QSAR Comb. Sci.* **2003**, *22*, 1006–1026.

(33) Chemitecs. Organic Chemicals Emitted from Technosphere Articles. http://www.chemitecs.se (accessed Aug 3, 2009).

(34) Muir, D. C. G.; Howard, P. H. Are there other persistent organic pollutants? A challenge for environmental chemists. *Environ. Sci. Technol.* **2006**, *40*, 7157–7166.

(35) European Commission, Joint Research Centre. ESIS: European chemical Substancese Information System http://ecb.jrc.ec.europa.eu/esis/ (accessed March 17, 2008).

(36) *Molecular Operating Environment (MOE), 2006.08*; Chemical Computing Group: Quebec, Canada, 2006.

(37) *Exposure Assessment Tools and Models, Estimation Program Interface (EPI) Suite, version 3.20*; U.S. Environmental Protection Agency. Exposure Assessment Branch: Washington, DC, 2008.

(38) *SIMCA-P+, version 12.0*; Umetrics AB: Umeå, Sweden, 2008.

(39) Ward, J. H. Hierarchical grouping to optimize an objective function. *J. Am. Stat. Assoc.* **1963**, *58*, 236–244.

CI9003255