A Nonlinear Map of Substituent Constants for Selecting Test Series and Deriving Structure–Activity Relationships. 1. Aromatic Series

Daniel Domine,[†,‡] James Devillers,*[,†] and Maurice Chastrette[‡]

CTIS, 21 rue de la Bannière, 69003 Lyon, and Laboratoire de Chimie Organique Physique, U.R.A. CNRS 463, Université Lyon-I, 43 Bd du 11 Novembre 1918, 69622 Villeurbanne CEDEX, France

Received August 23, 1993•

A nonlinear mapping (NLM) analysis was performed on a set of 166 aromatic substituents described by six variables encoding hydrophobic (π), steric (MR), and electronic effects (HBA, HBD, F, and R). NLM allowed to easily summarize the main information contained in the original data table. By means of collections of graphs, it was possible to relate the structure of the substituents to their π, MR, HBA, HBD, F, and R values. The proposed approach provides a useful and easy tool for the selection of test series and for deriving structure–activity relationships.

Introduction

For a chemical to engender a biological response when administered to a living organism, a number of processes must occur. Briefly, these deal with dissolution in body fluids, transport to a site of action, binding to a receptor, and initiation of a biological action. It is well known that these processes are governed by the physicochemical properties of the molecules.[1-3] In quantitative structure–activity relationship (QSAR) studies, to relate the physicochemical properties of aromatics and aliphatics to a biological activity, many parameters describing the hydrophobic, steric, and electronic effects of their substituents have been derived.[2-6] Among them, the most widely used are the π contribution of Hansch which depicts the lipophilic character of the substituents,[7] the Hammett σ constants which are used to account for electronic processes,[5,8] the Swain and Lupton F and R parameters derived from the σ constants which separate the inductive and resonance effects of the substituents,[9] and the molar refractivity (MR) used to describe the steric bulk of substituents.[10] For a comprehensive account of the parameters used in QSAR studies, one should refer to valuable previously published reviews.[2,11] The above substituent constants have been widely used, and others are still being developed. As a result, numerous data compilations of substituent constants have been elaborated.[1,3,12-14]

In drug design, it is essential to select test series with high information content in order to reduce the costs in research by maximizing the information content obtained from each molecular probe in a set of congeners.[15] A lot of works have been directed toward this aim, and many authors have proposed different methods. Historically, the first selection strategy was presented by Craig[16] who proposed use of 2-D plots with uncorrelated physicochemical properties (e.g., π vs σ) in order to select substituents covering a broad spectrum of physicochemical properties. In 1972, Topliss[17] proposed the so-called "decision tree" which consists in a stepwise synthesis of compounds taking into account the physicochemical properties supposed to influence the activity and the ease for synthesis. In 1973, Hansch *et al.*[18] introduced multivariate data analysis for solving this problem and used hierarchical cluster analysis (HCA) to derive ideal test series. In 1974, Darvas[19] published a procedure based on the simplex method, which has been successfully applied by Guilliom *et al.*[20] In 1975, Wootton *et al.*[21] introduced the multidimensional mapping (MM). The name of the method may be misleading since no map is produced. It consists in a "blind" walk in the n-dimensional space defined by the n physicochemical parameters chosen. This method was successfully applied to the selection of test series[22] and improved in 1983.[23] Goodford *et al.*[22] used nonlinear mapping (NLM) to visualize their results, and it appeared that the selected substituents were widely spread on the map. Streich *et al.*[24] combined the method of Wootton *et al.*[21] with principal components analysis (PCA) to obtain the so-called PCMM method. With PCMM, like with MM, synthetic chemists may feel unhappy with the thought that something even better may be hidden in this black box and that not enough room is left for the chemical intuition. Indeed, these methods do not offer a global vision of all the possible substituents and directly propose a series of substituents.

To solve this problem, Dove *et al.*[25] introduced the notion of mapping by the use of spectral mapping and stressed its advantages. With this method, a good test series with high data variance and low collinearity was always obtained if substituents distant from each other were selected in such a way that the whole space is systematically covered. This could simply be done by inspection of the map by eye. Since it was possible to obtain different test series on the same map, synthetic feasibility could always adequately be taken into account. However, they also underlined that in some cases the information content of the map could be too low. Their example only carried 65% of the information. In the same way, Alunni *et al.*[26] used PCA to derive clusters of substituents. They underlined four classes that were alkyls, donors, acceptors, and halogens. Although this method is interesting since it makes use of a graphical representation of the results, it suffers from the same problem as the method of Dove *et al.*[25] Furthermore, the first factorial plane (i.e., defined by the first two principal components) and the four classes may not be precise enough in terms of chemical information for the selection of representative test series. A similar approach was also used by van de Waterbeemd *et al.*[27] who derived five chemical classes from 59 substituents. It must be noted that the above list of techniques only reflects the main axes of research in the field. Due to the importance of the problem, other attempts have been made

* Author to whom all correspondence should be addressed.
† CTIS.
‡ Université Lyon-I.
• Abstract published in *Advance ACS Abstracts*, March 1, 1994.

0022-2623/94/1837-0973$04.50/0

to derive new approaches or modify some existing techniques.[15,28-32] For a comprehensive review, one should refer to a paper of Pleiss and Unger[33] dedicated to this topic and containing 245 bibliographical references. However, it must be pointed out that, from a practical point of view, none of these approaches are completely satisfactory.

In order to solve this problem, we propose the use of an original graphical approach based on the nonlinear mapping method. Briefly, NLM was designed by Sammon[34] and introduced in chemistry by Kowalski and Bender.[35,36] It is aimed at representing the points of an n-dimensional space in a lower d-dimensional space, preserving interpoint distances. As recently underlined,[37,38] NLM is well suited for structure–property and structure–activity relationship (SPR and SAR) studies since it allows to summarize the information contained in large data tables. Furthermore, the maps derived can be interpreted in terms of SAR by plotting various relevant qualitative and quantitative information on them.[38] Under these conditions, this study is aimed at providing a method allowing the selection of test series in order to easily derive SAR from a nonlinear map of substituent constants.

Nonlinear Mapping

On the basis of a concept similar to the classical multidimensional scaling (MDS),[39-41] NLM was designed by Sammon[34] to represent a set of points defined in an n-dimensional space by a human-perceivable configuration of the data in a lower d-dimensional space (d = 2 or 3). NLM tries to preserve distances between points in the display space as similar as possible to the actual distances in the original space. The procedure for performing this transformation can be summarized as follows. (i) Interpoint distances in the original space are computed. (ii) An initial configuration (generally random) of the points in the display space is chosen. (iii) A mapping error (E) is calculated from the distances in the two spaces. (iv) Coordinates of points in the display space are iteratively modified by means of a nonlinear procedure so as to minimize the mapping error. The algorithm terminates when no significant decrease in the mapping error is obtained over the course of several iterations.[38,42,43]

In the Sammon's algorithm used in this study, the minimization process is the steepest descent procedure, which is performed as follows. Suppose the interpoint distances $d_{ij}(m)$ between points i and j at the mth configuration described by the Euclidean distance as shown below:

$$d_{ij}(m) = [\sum_{k=1}^{d}(x_{ik}(m) - x_{jk}(m))^2]^{1/2}$$

and the corresponding error $E(m)$ as defined by Sammon,[34]

$$E(m) = \frac{1}{\sum_{i<j}^{N} d_{ij}^*} \sum_{i<j}^{N} \frac{[d_{ij}^* - d_{ij}(m)]^2}{d_{ij}^*}$$

then the steepest descent procedure proceeds as shown below. The coordinates in the $(m + 1)$th configuration are given by:

$$x_{pq}(m+1) = x_{pq}(m) - (\mathrm{MF} \cdot \Delta_{pq}(m))$$

where

$$\Delta_{pq}(m) = \frac{\partial E(m)}{\partial x_{pq}(m)} \Big/ \left|\frac{\partial^2 E(m)}{\partial x_{pq}(m)^2}\right|$$

and MF is a magic factor empirically determined as 0.3 or 0.4.[34]

This process is carried out iteratively until a threshold fixed by the user is attained (i.e., minimal error or minimal difference between the error at step $m - 1$ and step m in the iteration process). Precautions must be taken to prevent any two points in the d-dimensional space from becoming identical to avoid problems in the calculation of the partial derivatives. Points in the n-dimensional space must also be different.

Additional information on the practical aspects of the NLM method and a review of its uses in QSAR studies can be found in a previous paper.[38]

Experimental Section

A nonlinear mapping analysis[38] was performed on a set of 166 aromatic substituents[14] described by six substituent constants encoding their hydrophobic, steric, and electronic effects. These parameters were respectively the π constant, the molar refractivity (MR), the H-bonding acceptor (HBA) and donor (HBD) abilities, and the inductive and resonance parameters of Swain and Lupton[9] F and R. Data used in this study can be found in Hansch and Leo.[14] All inductive and resonance field constants F and R were recalculated from the σ_m and σ_p constants of Hammett[9] with equations of Swain and Lupton,[9] and corrections were made when values obtained were different from those reported in the compilation of substituent constants.[14] It is obvious that quantitative values for the H-bonding abilities[44,45] would have been better, but available data were still too scarce to handle all the substituents of our data set. For the NLM analysis, π, MR, F, and R were centered (i.e., zero mean) and reduced (i.e., unit variance). For HBA and HBD, the 1's were replaced by a value yielding a unit variance. Note that, due to the fact that calculations were performed on the distance matrix, the results were not affected by the centering. It was only performed to improve the visualization of the data when they were reported on the map. The results obtained were interpreted in terms of SPR and SAR by plotting various qualitative and quantitative information on the nonlinear map derived as recently described.[38,46-49] The map was interpreted taking care of its statistical significance, viz., inspecting the total mapping error and the goodness of fit of each point.[38] The NLM analysis was performed with the STATQSAR package[50] and the graphical analysis with GraphMu.[51]

Results and Discussion

Figure 1.1 shows the nonlinear map of the 166 aromatic substituents described by six substituent constants. It is noteworthy that is was impossible with PCA on standardized data (i.e., unit variance and zero mean) to summarize on a sole plane the information contained in the original data matrix. Indeed, the first two factors only explained 58% of the total variance. At the opposite, with a low mapping error of 6.4e − 2 obtained with NLM, we can advance that the main information contained in the original data matrix is summarized on the nonlinear map.[34,38] For interpreting the nonlinear map, it is necessary to inspect each individual on the map and compare its location to parent substituents and its neighbors. Therefore, it is necessary to have an estimation of the individual goodness of fit for all substituents. Indeed, even if the mapping error is low, the error carried by some points can still be important. For this purpose, the individual mapping error[38] of each point has been plotted on Figure 1.1 by means of squares proportional to the

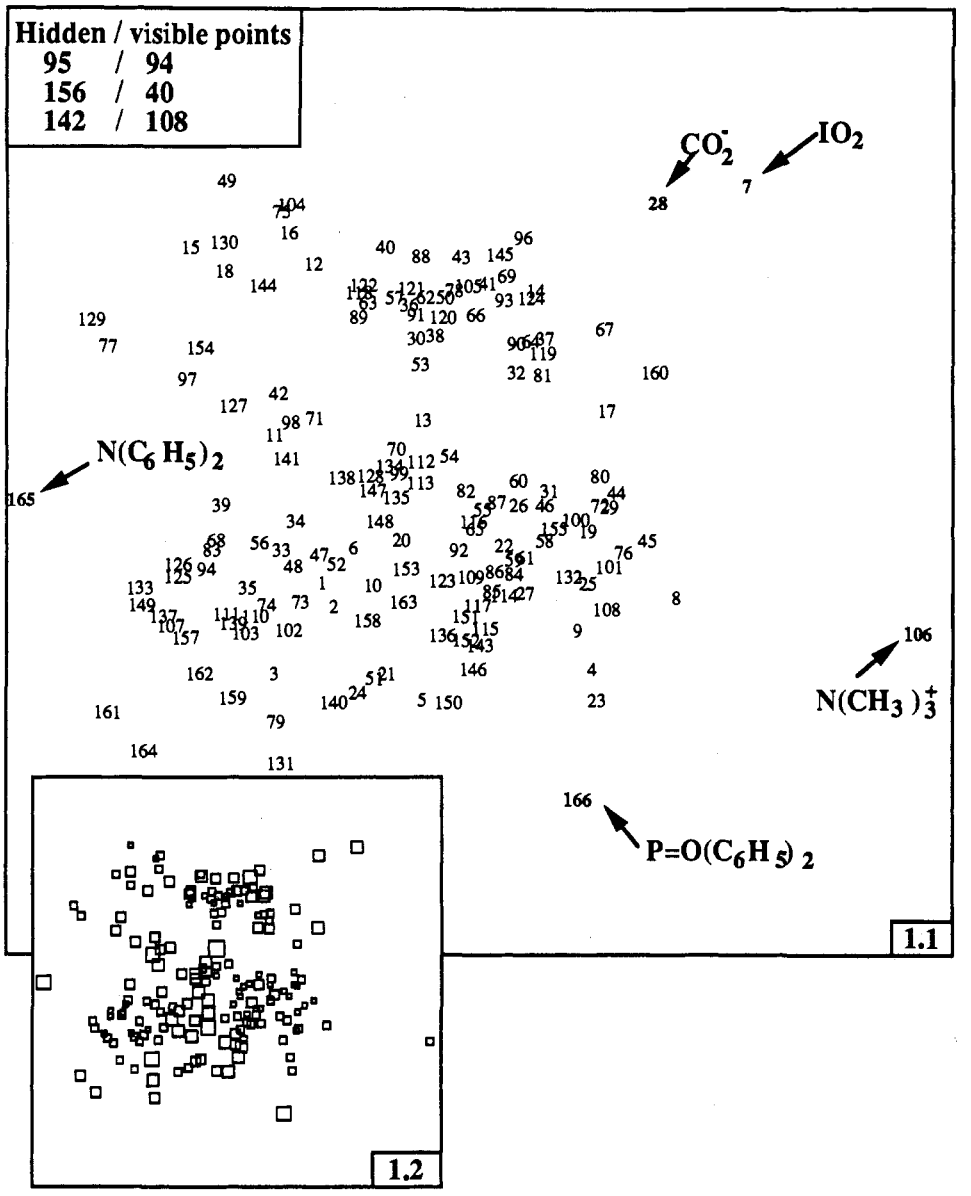

Figure 1. (1.1) Nonlinear map of the 166 aromatic substituents described by six substituent constants (π, HBA, HBD, MR, F, and R). (1.2) Plot of the individual mapping errors on each substituent of the nonlinear map. Squares are proportional in size to the magnitude of the errors. 1, Br; 2, Cl; 3, F; 4, SO_2F; 5, SF_5; 6, I; 7, IO_2; 8, NO; 9, NO_2; 10, NNN; 11, H; 12, OH; 13, SH; 14, $B(OH)_2$; 15, NH_2; 16, NHOH; 17, SO_2NH_2; 18, $NHNH_2$; 19, 5-Cl-1-tetrazolyl; 20, $N{=}CCl_2$; 21, CF_3; 22, OCF_3; 23, SO_2CF_3; 24, SCF_3; 25, CN; 26, NCS; 27, SCN; 28, CO_2^-; 29, 1-tetrazolyl; 30, NHCN; 31, CHO; 32, CO_2H; 33, CH_2Br; 34, CH_2Cl; 35, CH_2I; 36, NHCHO; 37, $CONH_2$; 38, CH=NOH; 39, CH_3; 40, $NHCONH_2$; 41, $NHC{=}S(NH_2)$; 42, OCH_3; 43, CH_2OH; 44, $SOCH_3$; 45, SO_2CH_3; 46, OSO_2CH_3; 47, SCH_3; 48, $SeCH_3$; 49, $NHCH_3$; 50, $NHSO_2CH_3$; 51, CF_2CF_3; 52, C≡CH; 53, $NHCOCF_3$; 54, CH_2CN; 55, $CH{=}CHNO_2$ (trans); 56, $CH{=}CH_2$; 57, $NHC{=}O(CH_2Cl)$; 58, $COCH_3$; 59, $SCOCH_3$; 60, $OCOCH_3$; 61, CO_2CH_3; 62, $NHCOCH_3$; 63, $NHCO_2CH_3$; 64, $C{=}O(NHCH_3)$; 65, $CH{=}NOCH_3$; 66, $NHC{=}S(CH_3)$; 67, $CH{=}NNHC{=}S(NH_2)$; 68, CH_2CH_3; 69, $CH{=}NNHCONHNH_2$; 70, CH_2OCH_3; 71, OCH_2CH_3; 72, SOC_2H_5; 73, SC_2H_5; 74, SeC_2H_5; 75, NHC_2H_5; 76, $SO_2C_2H_5$; 77, $N(CH_3)_2$; 78, $NHSO_2C_2H_5$; 79, $P(CH_3)_2$; 80, $PO(OCH_3)_2$; 81, $C(OH)(CF_3)_2$; 82, CH=CHCN; 83, cyclopropyl; 84, COC_2H_5; 85, $SCOC_2H_5$; 86, $CO_2C_2H_5$; 87, $OCOC_2H_5$; 88, $CH_2CH_2CO_2H$; 89, $NHCO_2C_2H_5$; 90, $CONHC_2H_5$; 91, $NHCOC_2H_5$; 92, $CH{=}NOC_2H_5$; 93, $NHC{=}S(C_2H_5)$; 94, $CH(CH_3)_2$; 95, C_3H_7; 96, $NHC{=}S(NHC_2H_5)$; 97, $OCH(CH_3)_2$; 98, OC_3H_7; 99, $CH_2OC_2H_5$; 100, SOC_3H_7; 101, $SO_2C_3H_7$; 102, SC_3H_7; 103, SeC_3H_7; 104, NHC_3H_7; 105, $NHSO_2C_3H_7$; 106, $N(CH_3)_3^+$; 107, $Si(CH_3)_3$; 108, $CH{=}C(CN)_2$; 109, 1-pyrryl; 110, 2-thienyl; 111, 3-thienyl; 112, $CH{=}CHCOCH_3$; 113, $CH{=}CHCO_2CH_3$; 114, COC_3H_7; 115, $SCOC_3H_7$; 116, $OCOC_3H_7$; 117, $CO_2C_3H_7$; 118, $(CH_2)_3CO_2H$; 119, $CONHC_3H_7$; 120, $NHCOC_3H_7$; 121, $NHC{=}OCH(CH_3)_2$; 122, $NHCO_2C_3H_7$; 123, $CH{=}NOC_3H_7$; 124, $NHC{=}S(C_3H_7)$; 125, C_4H_9; 126, $C(CH_3)_3$; 127, OC_4H_9; 128, $CH_2OC_3H_7$; 129, $N(C_2H_5)_2$; 130, NHC_4H_9; 131, $P(C_2H_5)_2$; 132, $PO(OC_2H_5)_2$; 133, $CH_2Si(CH_3)_3$; 134, $CH{=}CHCOC_2H_5$; 135, $CH{=}CHCO_2C_2H_5$; 136; $CH{=}NOC_4H_9$; 137, C_5H_{11}; 138, $CH_2OC_4H_9$; 139, C_6H_5; 140, $N{=}NC_6H_5$; 141, OC_6H_5; 142, $SO_2C_6H_5$; 143, $OSO_2C_6H_5$; 144, NHC_6H_5; 145, $NHSO_2C_6H_5$; 146, 2,5-di-Me-1-pyrryl; 147, $CH{=}CHCOC_3H_7$; 148, $CH{=}CHCO_2C_3H_7$; 149, cyclohexyl; 150, 2-benzthiazolyl; 151, COC_6H_5; 152, $CO_2C_6H_5$; 153, $OCOC_6H_5$; 154, $N{=}CHC_6H_5$; 155, $CH{=}NC_6H_5$; 156, $NHCOC_6H_5$; 157, $CH_2C_6H_5$; 158, $CH_2OC_6H_5$; 159, $C{\equiv}CC_6H_5$; 160, $CH{=}NNHCOC_6H_5$; 161, $CH_2Si(C_2H_5)_3$; 162, $CH{=}CHC_6H_5$ (trans); 163, $CH{=}CHCOC_6H_5$; 164, ferrocenyl; 165, $N(C_6H_5)_2$; 166, $P{=}O(C_6H_5)_2$.

magnitude of the individual errors (Figure 1.2). Figure 1.2 shows that variations in individual errors are low. Therefore, as the total mapping error is also low (i.e., $6.4e-2$), it is not necessary to pay particular attention to any point when interpreting the map. A rapid glimpse of Figure 1.1 allows to stress the atypical locations of substituents $n°$ 7 (IO_2), 28 (CO_2^-), 106 ($N(CH_3)_3^+$), 165 ($N(C_6H_5)_2$), and 166 ($P{=}O(C_6H_5)_2$). The atypical location of substituents $n°$ 7, 28, and 106 may be attributed to the fact that these groups have very low π values.[14] Furthermore, substituents $n°$ 28 and 106 are the sole groups bearing a charge, and substituent $n°$ 7 is the most bulky

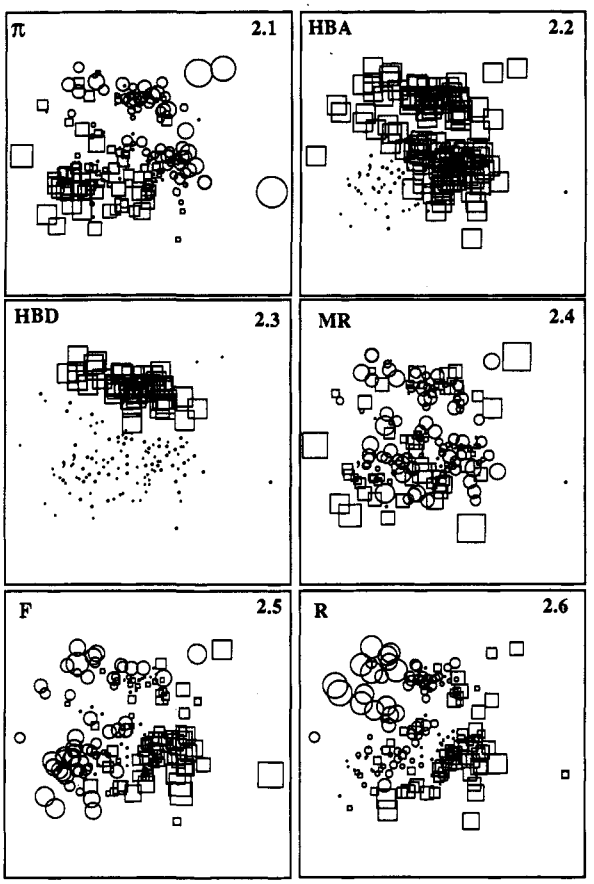

Figure 2. Plot of the scaled values of the six parameters on each substituent of the nonlinear map. Squares (positive values) and circles (negative values) are proportional in size to the magnitude of the parameters. In Figure 2.2,.3, the dots indicate the substituents which do not have the ability to accept and donate H-bonds, respectively.

group of the set. This may also explain the atypical locations of substituents $n°$ 165 and 166 which also have very high values of MR compared to those of all the other substituents.

I. Interpretation of the Nonlinear Map in Terms of Structure–Property Relationships. A. Representation of the Data on the Nonlinear Map. Figure 1.1 could be directly interpreted in terms of SPR, but this would require to dart back and forth between the original data and the structural features of the substituents. To facilitate this work, the values used for the NLM analysis (i.e., centered and reduced for π, MR, F and R; reduced for HBA and HBD) have been plotted on the nonlinear map by means of squares (positive values) and circles (negative values) whose sizes are proportional to the magnitude of the studied parameters (Figure 2). Briefly, the larger the square, the larger the value, and the larger the circle, the smaller the value. Figure 2 shows that the substituents are distributed and clustered on the nonlinear map according to their substituent constants. Indeed, for all parameters except MR (Figure 2.4), gradients or clusters can be observed. Thus, π values of the substituents decrease along an axis running from the bottom left to the top right-hand side of Figure 2.1. In the same way, there is an obvious clustering of H-bond acceptor and donor substituents in Figure 2.2,.3. Last, Figure 2.5,.6 reveals that gradients are observable for F and R values. F values increase from left to right, and R values increase from the top left-hand corner to the bottom right-hand corner. In order to underline structure–property relationships for

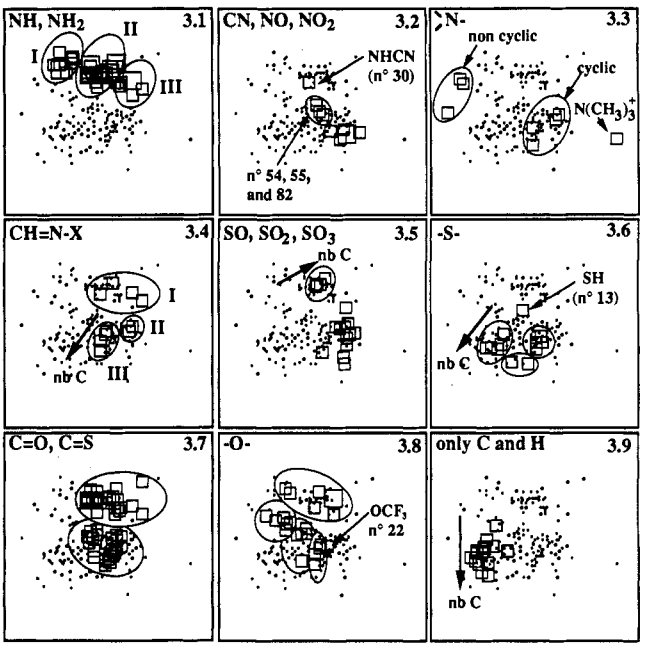

Figure 3. Plot of the presence or frequency of some functional groups or skeleton similarities in the aromatic substituents. Squares are proportional in size to the number of groups. The absence of a functional group is represented by a dot.

the 166 substituents under study and demonstrate the coherence of the nonlinear map (Figure 1.1) with regards to the structure of the substituents, various structural information (i.e., presence of functional groups and/or skeleton similarities) has been reported on the nonlinear map (see Figure 3).

B. Projection on the Nonlinear Map of the Presence or Frequency of Functional Groups. Figure 3 is aimed at giving a full description of the nonlinear map (Figure 1.1) in terms of chemical structures. Figure 3.1 shows that substituents containing primary or secondary amine groups cluster at the top of the nonlinear map. Comparison with Figure 2 reveals that this is indeed associated with HBA and HBD abilities (Figure 2.2,.3) but it is also generally associated with low π values (Figure 2.1). A more precise inspection of these substituents also reveals that the amine cluster can be divided into three subclusters. The first, located on the left-hand side (subcluster I = $n°$ 15, 16, 18, 49, 75, 104, 130, and 144), contains substituents with formula NHR (R = H, OH, NH_2, alkyl, and phenyl). In the middle are found substituents in which the NH group is bound to a C=O, C=S, CN, CO_2, or SO_2 group (including their derivatives) and to the substitution site (subcluster II = $n°$ 30, 36, 40, 41, 50, 53, 57, 62, 63, 66, 78, 89, 91, 93, 96, 105, 120, 121, 122, 124, 145, and 156). Subcluster III ($n°$ 17, 37, 64, 67, 90, 119, and 160) consists of the substituents for which the amine group is not the group bound to the derivatives (e.g., $CONHCH_3$). The only exception is substituent $n°$ 69 which is located near subcluster II. Comparison with Figure 2 shows that this repartition of the three subclusters is due to increased R and F values.

In Figure 3.2, we have represented the number of CN, NO, and NO_2. All these substituents are located in the same region of the map and therefore have similar physicochemical properties. The outlier observed is substituent $n°$ 30 (NHCN) which has an amine group and therefore clusters with the other substituents containing this group (Figure 3.1). Comparison of Figure 3.2 with

Figure 2.1–.6 shows that the location of these substituents is characterized by relatively large F and R values (Figure 2.5,.6). This effect is less for the substituents $n°$ 54, 55, and 82.

On Figure 3.3 is represented the presence of tertiary amine groups. This map indicates that these substituents form two clusters. It is noteworthy that the noncyclic ones are found in the same cluster in the left-hand side of the map ($n°$ 77, 129, and 165), while the cyclic ones ($n°$ 19, 29, 109, and 146) are found in the bottom right-hand cluster. Comparison with Figure 2 shows that they principally differ by their F and R values. The quaternary amine $N(CH_3)_3^+$ (substituent $n°$ 106) has also been represented on Figure 3.3. It appears as an outlier due to its very low π value and high F value. Another difference is that it cannot accept H-bonds unlike tertiary amines.

Figure 3.4 reveals that the substituents containing the group CH=NX with X = N or O form three clusters on the nonlinear map. Cluster I ($n°$ 38, 67, 69, and 160) contains the considered group bound to amido or thioamido groups (except $n°$ 38: alcohol). Cluster II ($n°$ 19 and 29) consists of the tetrazolyl substituents. Cluster III ($n°$ 65, 92, 123, and 136) contains the considered group bound to an alkoxy group. Examination of Figure 2.2,.3 reveals that cluster I differs from clusters II and III by the fact that the former can accept and donate H-bonds while the latter group can only accept H-bonds. For cluster III, it is noteworthy that a gradient linked to the number of carbon atoms in the alkyl chain of the alkoxy group running through substituents $n°$ 65, 92, 123, and 136 can be observed (nb C in Figure 3.4).

Figure 3.5 shows that substituents containing SO, SO_2, or SO_3 groups are preferentially located in the right-hand side of the cloud of points displayed on the map, indicating that their π values are generally low (Figure 2.1) and that they have the ability to accept H-bonds (Figure 2.2). Two clusters can be easily identified among these substituents. The first, at the top of the figure, contains the substituents with the general formula $NHSO_2R$ with R = alkyl or phenyl ($n°$ 50, 78, 105, and 145). A closer inspection of this cluster reveals that a gradient linked to the number of carbon atoms of the group R can be drawn. Indeed, running from left to right, we find a methyl ($n°$ 50), an ethyl ($n°$ 78), a propyl ($n°$ 105), and, last, a phenyl ($n°$ 145) group bound to the $NHSO_2$ group. The second cluster consists of the same type of series for SOR, SO_2R, and SO_3R (Figure 3.5) plus the two fluorinated substituents ($n°$ 4 and 23). They differ from the elements of the first cluster by higher F and R values (Figure 2.5,.6) and higher π values (Figure 2.1) for the substituents $n°$ 4 and 23. Another fundamental difference is that the elements of the first cluster can accept and donate H–bonds while those of the second cluster can only accept H-bonds (Figure 2.2,.3). Between these two clusters, SO_2NH_2 (substituent $n°$ 17) occupies an intermediate location with HBA and HBD abilities but rather high values for F and R and a low value of π.

Figure 3.6 shows that the substituents containing an -S- group are found at the bottom of the cloud of points on the map. The thiol group ($n°$ 13) has also been represented on this figure and appears as an outlier above the -S- substituents. This is due to its HBD ability (Figure 2.3) and, also, its rather low π value (Figure 2.1). A finer examination of this figure shows that it is possible to separate this cluster into three subclusters. Indeed, on the left-hand side are found the substituents with the general formula -SR, R being an alkyl group ($n°$ 47, 73, and 102) and the thienyl substituents ($n°$ 110 and 111), among which an axis representing the number of carbon atoms can be drawn from the top to the bottom of the subcluster. On the right-hand side are found the thioethers bound to a carbonyl or cyano group ($n°$ 27, 59, 85, and 115) which tend to have higher F and R values (Figure 2.5,.6) and have the ability to accept H-bonds (Figure 2.2). The substituents are distributed according to the number of carbon atoms they contain. In the middle, at the bottom of the figure, are found two atypical substituents ($n°$ 24 and 150) which possess particular properties.

On Figure 3.7 has been plotted the presence of C=O or C=S groups. These substituents form two clusters which are related to their association to groups giving or not giving H-bonds. Indeed, the top cluster is characterized by HBA and HBD abilities (Figure 2.2,.3), while the substituents of the cluster located below only have the ability to accept H-bonds. The difference also lies in the lower π values of the substituents belonging to the former cluster (Figure 2.1).

A more detailed inspection of this figure could reveal gradients depending on the number of carbon atoms for each chemical family running in the vertical direction to the bottom of the map. Thus, for example, if we consider the esters with the general formula CO_2R with R = CH_3 ($n°$ 61), C_2H_5 ($n°$ 86), C_3H_7 ($n°$ 117), and C_6H_5 ($n°$ 152), it can be seen in Figure 1.1 that they are distributed along an axis linked to the number of carbon atoms they contain.

Inspection of Figure 3.8 reveals that the substituents containing an -O- group can have very different physicochemical properties according to the functional group in which they are involved but, also, their position (i.e., bound to the substitution site or not). Note that acids and esters have not been represented on this figure. A close inspection of Figure 3.8 reveals that four clusters can be isolated. First, at the top of the figure are found all the substituents containing an OH group ($n°$ 12, 14, 16, 38, 43, and 81). They can donate H-bonds (Figure 2.3) and have lower π values (Figure 2.1) except substituent $n°$ 81. The three remaining clusters cannot give H-bonds. They are, first, on the left-hand side of the map, substituents for which the oxygen atom is involved in an ether functional group and is bound to the substitution site ($n°$ 42, 71, 97, 98, 127, and 141). This results in low R values (Figure 2.6). The second cluster located in the middle of the map consists of the substituents with the general formula CH_2OR with R = CH_3 ($n°$ 70), C_2H_5 ($n°$ 99), C_3H_7 ($n°$ 128), C_4H_9 ($n°$ 138), and C_6H_5 ($n°$ 158). These substituents have higher R values (Figure 2.6). The third cluster contains the substituents CH=NOR with R = CH_3 ($n°$ 65), C_2H_5 ($n°$ 92), C_3H_7 ($n°$ 123), and C_4H_9 ($n°$ 136). For these last two clusters, gradients depending on the number of carbon atoms can be observed. Last, on the right-hand side of these clusters is found the substituent $n°$ 22 (OCF_3) which occupies an atypical location, compared to the OR substituents, due to its higher F and R values.

Figure 3.9 shows that substituents containing only C and H atoms cluster in the bottom left-hand corner of the map. Furthermore, a closer inspection of the map reveals that a gradient depending on the number of carbon atoms can be drawn up.

It is obvious that all possible groups were not represented in Figure 3, but our approach is flexible enough to allow

the projection of any other information. For SAR purposes, the observation of all the maps allows to determine substituents different in nature but having similar properties and, also, the relative influence of the functional groups on the properties of the substituents. The nonlinear map (Figure 1.1) is coherent in terms of chemical information, since there are clusters for each functional group and gradients are observed inside each cluster.

II. Selection of Test Series. Since only a tiny fraction of the almost infinite number of possibilities can be studied in drug modification, we cannot afford redundancy.[14] Testing two congeners that have essentially the same physicochemical parameters (i.e., very close to each other on Figure 1.1) is most likely to be less valuable than testing two with different properties.[14] For selecting test series with high information content, the use of NLM coupled to graphical tools can be very useful since NLM presents the same advantages as the linear methods using the representation on a plane of the individuals,[25-27] but, in addition, it is more likely to get more information on the map, as has been shown here and in previous papers.[38,46] In this study, the information contained in the large data table of 166 substituents is summarized on Figure 1.1. Furthermore, Figures 2 and 3 give a full description of the map in terms of structural information. Selection of test series is performed by a simple inspection of the map by eye. In addition to structural information, it is, of course, possible to adequately consider all other available information (e.g., synthetic feasibility or previous knowledge on the activity). This can be achieved by means of the graphical tools presented in this paper. NLM coupled to graphical tools is therefore a very simple and straightforward representation of the results in line with classical chemical thinking, which should be attractive to synthetic chemists. Furthermore, the 2-D map gives a full picture of the data structure in the starting population which cannot be obtained with MM[21] and is only partly represented by HCA.[14,18] The within-cluster position of substituents is now known, and no *a priori* decision as in other methods (except for the parameter space to be considered) is necessary to establish that map.

The map presented here is restricted to monosubstitutions, but other maps obtained from the compilation of the physicochemical parameters for polysubstitutions could be easily derived following the procedure presented in this paper.

For comparison, a possible "ideal" test series[14] obtained from a hierarchical cluster analysis (HCA) has been represented on Figure 1.1 (Figure 4). Figure 4 shows that the NLM would have been well suited for the selection of such a test series since the points selected by the authors[14] are widely spread on our map. Squares represent the test series selected and circles previously selected substituents not retained due to various constraints.[14] On Figure 4, the arrows link the substituents not retained to those chosen for replacing them. These arrows show that some replacements are made between substituents having rather different substituent constants. For example, the selection of H ($n°$ 11) which was in another cluster to replace $N(CH_3)_3^+$ ($n°$ 106) reveals a gap in the HCA approach. Indeed, the replacement of $N(CH_3)_3^+$ ($n°$ 106) was logical since it is an outlier, but the selection of H ($n°$ 11) may not be the best choice since it is located too close to two other selected substituents. With HCA, the replacement by another substituent in a different cluster is performed

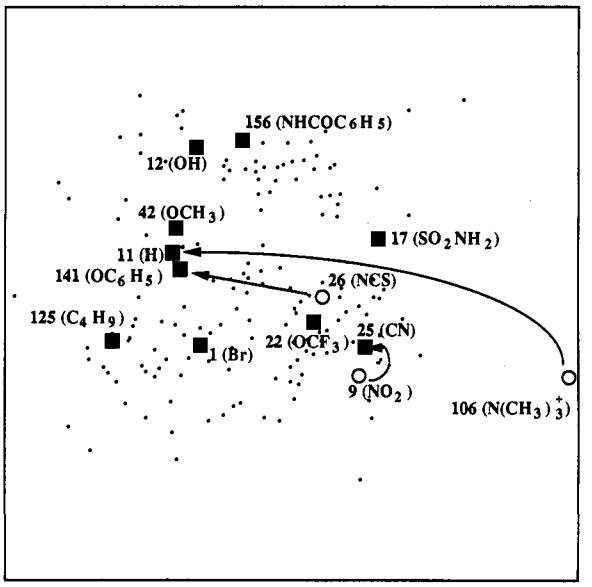

Figure 4. Representation of a selected test series.[14] For captions, see text.

in a blind manner unless we return to the original data table. With NLM, it can be made more easily by selecting a substituent in the vicinity of the undesirable substituent.

III. Deriving Structure–Activity Relationships from the Nonlinear Map. Our graphical approach is particularly suitable in SAR studies since it underlines relationships between chemical structures and biological responses. To briefly illustrate this point, one example dealing with experimental results depicting the activity of the aniline mustards action against the solid tumor B-16 melanoma is presented below.[52] These data were retrieved from a publication of Panthananickal and co-workers[52] who performed a QSAR analysis on a set of 22 substituted di(2-chloroethyl)anilines. For the projection of the biological data on Figure 1.1, we only kept the sixteen 4-substituted derivatives (Figure 5). For a better visualization, data given as $\log(1/C)$ were transformed as C in mmol kg^{-1}. Therefore, the larger the squares, the smaller the activity. Figure 5 shows that the necessary concentration for having a 25% increase in the life span of the mice increases along an axis running from the top left-hand corner to the bottom right-hand corner of the map. Comparison with Figure 2 confirms the results of Panthananickal *et al.*,[52] since it appears that the activity seems to depend on both the electronic and lipophilic characters of the substituents. Only CN ($n°$ 25) and OCH_2CH_3 ($n°$ 71) are poorly fit on the map since a small square is observed in a region of the map where a larger square would have been expected. Figure 5 shows that our map allows to qualitatively predict the activity of a derivative and matches the experimental results. Furthermore, we can stress that the selection of substituents for this study was rational since the selected points are well spread on the nonlinear map. The physicochemical information contained in this figure summarizes well the actual influence of the substituent constants on the biological action of drugs.

The NLM analysis of the large aromatic substituent constants data table (166 × 6) shows that it is possible to summarize the main information contained in this table. The graphical approach used in this study allows to derive and represent SPR on the nonlinear map. These collections of graphs provide useful and easy tools for the

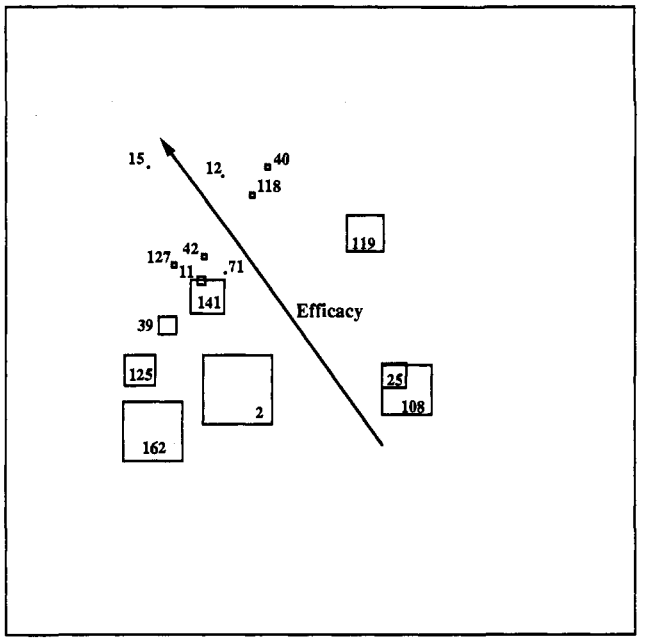

Figure 5. Plot of the antitumor activity of 16 aniline mustards.[52] Squares are proportional to the concentration (mmol kg^{-1}) inducing a 25% increase in the life span of mice. Substituent numbers are given for easy cross-reference with Figure 1.1.

selection of substituents for the design of test series. Indeed, to ensure a broad spectrum of substituent constants, substituents are simply selected in the different regions of Figure 1.1. The plot of the data used to derive the map and the structural information (Figures 2 and 3) gives a full description of the nonlinear map to help in the selection of a test series. This method is open, and any information susceptible to help in the selection of the test series (e.g., synthetic feasibility, previous knowledge on the biological activity) can be represented. When substituents very different in nature are located in the same region of the map, these may all be assayed in order to test if there can be different mechanisms of action or whatever specifically linked to the substituents considered. For particular cases such as multiple substitutions or larger differences in the structures of the chemicals to be studied, it may be useful to recalculate a map with the variables suspected to influence the activity under study and to perform a selection as described in this paper. Last, once the biological tests are performed with a test series, the nonlinear map can be useful to stress relationships between the structures and properties of the substituents and their biological activities. Due to the encouraging results obtained in this study, future work will be directed toward the study of aliphatic substituents.

References

(1) Norrington, F. E.; Hyde, R. M.; Williams, S. G.; Wootton, R. Physicochemical-Activity Relations in Practice. 1. A Rational and Self-Consistent Data Bank. *J. Med. Chem.* **1975**, *18*, 604–607.

(2) Dearden, J. C. Physico-chemical Descriptors. In *Practical Applications of Quantitative Structure-Activity Relationships (QSAR) in Environmental Chemistry and Toxicology*; Karcher, W., Devillers, J., Eds.; Kluwer Academic Publishers: Dordrecht, 1990; pp 25–59.

(3) Hansch, C.; Leo, A.; Taft, R. W. A Survey of Hammett Substituent Constants and Resonance and Field Parameters. *Chem. Rev.* **1991**, *91*, 165–195.

(4) Taft, R. W. The General Nature of the Proportionality of Polar Effects of Substituent Groups in Organic Chemistry. *J. Am. Chem. Soc.* **1953**, *75*, 4231–4238.

(5) Bowden, K. Electronic Effects in Drugs. In *Comprehensive Medicinal Chemistry*; Ramsden, C. A., Ed.; Pergamon Press: Oxford, 1990; Vol. 4, pp 205–239.

(6) Silipo, C.; Vittoria, A. *QSAR: Rational Approaches to the Design of Bioactive Compounds*; Elsevier: Amsterdam, 1991; p 575.

(7) Hansch, C.; Fujita, T. ρ-σ-π Analysis. A Method for the Correlation of Biological Activity and Chemical Structure. *J. Am. Chem. Soc.* **1964**, *86*, 1616–1626.

(8) Ludwig, M.; Wold, S.; Exner, O. The Role of *meta* and *para* Benzene Derivatives in the Evaluation of Substituent Effects: a Multivariate Data Analysis. *Acta Chem. Scand.* **1992**, *46*, 549–554.

(9) Swain, C. G.; Lupton, E. C. Field and Resonance Components of Substituent Effects. *J. Am. Chem. Soc.* **1968**, *90*, 4328–4337.

(10) Dunn, W. J. Molar Refractivity as an Independent Variable in Quantitative Structure-Activity Studies. *Eur. J. Med. Chem.* **1977**, *12*, 109–112.

(11) Livingstone, D. J. Quantitative Structure-Activity Relationships. In *Similarity Models in Organic Chemistry, Biochemistry and Related Fields*; Zalewski, R. I., Krygowski, T. M., Shorter, J., Eds.; Elsevier: Amsterdam, 1991; pp 557–627.

(12) Hansch, C.; Leo, A.; Unger, S. H.; Kim, K. H.; Nikaitani, D.; Lien, E. J. "Aromatic" Substituent Constants for Structure-Activity Correlations. *J. Med. Chem.* **1973**, *16*, 1207–1216.

(13) Rekker, R. F.; de Kort, H. M. The Hydrophobic Fragmental Constant; An Extension to a 1000 Data Point Set. *Eur. J. Med. Chem.* **1979**, *14*, 479–488.

(14) Hansch, C.; Leo, A. *Substituent Constants for Correlation Analysis in Chemistry and Biology*; John Wiley & Sons: New York, 1979.

(15) Schaper, K. J. Rational Selection of Test Series for QSAR Analysis. *Quant. Struct. Act. Relat.* **1983**, *2*, 111–120.

(16) Craig, P. N. Interdependence between Physical Parameters and Selection of Substituent Groups for Correlation Studies. *J. Med. Chem.* **1971**, *14*, 680–684.

(17) Topliss, J. G. Utilization of Operational Schemes for Analog Synthesis in Drug Design. *J. Med. Chem.* **1972**, *15*, 1006–1011.

(18) Hansch, C.; Unger, S. H.; Forsythe, A. B. Strategy in Drug Design. Cluster Analysis as an Aid in the Selection of Substituents. *J. Med. Chem.* **1973**, *16*, 1217–1222.

(19) Darvas, F. Application of the Sequential Simplex Method in Designing Drug Analogs. *J. Med. Chem.* **1974**, *17*, 799–804.

(20) Guilliom, R. D.; Purcell, W. P.; Bosin, T. R. Sequential Simplex Optimization Applied to Drug Design in the Indole, 1-Methylindole, and Benzo[b]thiophene Series. *Eur. J. Med. Chem.* **1977**, *12*, 187–192.

(21) Wootton, R.; Cranfield, R.; Sheppey, G. C.; Goodford, P. J. Physicochemical-Activity Relationships in Practice. 2. Rational Selection of Benzenoid Substituents. *J. Med. Chem.* **1975**, *18*, 607–613.

(22) Goodford, P. J.; Hudson, A. T.; Sheppey, G. C.; Wootton, R.; Black, M. H.; Sutherland, G. J.; Wickham, J. C. Physicochemical-Activity Relationships in Asymmetrical Analogues of Methoxychlor. *J. Med. Chem.* **1976**, *19*, 1239–1247.

(23) Wootton, R. Selection of Test Series by a Modified Multidimensional Mapping Method. *J. Med. Chem.* **1983**, *26*, 275–277.

(24) Streich, W. J.; Dove, S.; Franke, R. On the Rational Selection of Test Series. 1. Principal Component Method Combined with Multidimensional Mapping. *J. Med. Chem.* **1980**, *23*, 1452–1456.

(25) Dove, S.; Streich, W. J.; Franke, R. On the Rational Selection of Test Series. 2. Two-Dimensional Mapping of Intraclass Correlation Matrices. *J. Med. Chem.* **1980**, *23*, 1456–1459.

(26) Alunni, S.; Clementi, S.; Edlund, U.; Johnels, D.; Hellberg, S.; Sjöström, M.; Wold, S. Multivariate Data Analysis for Substituent Descriptors. *Acta Chem. Scand.* **1983**, *B37*, 47–53.

(27) van de Waterbeemd, H.; El Tayar, N.; Carrupt, P. A.; Testa, B. Pattern Recognition Study of QSAR Substituent Descriptors. *J. Comput.-Aided Mol. Des.* **1989**, *3*, 111–132.

(28) Martin, Y. C.; Panas, H. N. Mathematical Considerations in Series Design. *J. Med. Chem.* **1979**, *22*, 784–791.

(29) Wooldridge, K. R. H. A Rational Substituent Set for Structure-Activity Studies. *Eur. J. Med. Chem.* **1980**, *15*, 63–66.

(30) Austel, V. Selection of Test Compounds from a Basic Set of Chemical Structures. *Eur. J. Med. Chem.* **1982**, *17*, 339–347.

(31) de Winter, M. L. Computer Pre-selection of Compounds for Pharmacological Screening. Prediction by Fragment Description. *Eur. J. Med. Chem.* **1985**, *20*, 175–179.

(32) Boyd, D. B.; Seward, C. M. The Substituent Parameter Database: A Powerful Tool for QSAR Analysis. In *QSAR: Rational Approaches to the Design of Bioactive Compounds*; Silipo, C., Vittoria, A., Eds.; Elsevier: Amsterdam, 1991; pp 167–170.

(33) Pleiss, M. A.; Unger, S. H. The Design of Test Series and the Significance of QSAR Relationships. In *Comprehensive Medicinal Chemistry*; Ramsden, C. A., Ed.; Pergamon Press: Oxford, 1990; Vol. 4, pp 561–587.

(34) Sammon, J. W. A Nonlinear Mapping for Data Structure Analysis. *IEEE Trans. Comput.* **1969**, *C-18*, 401–409.

(35) Kowalski, B. R.; Bender, C. F. Pattern Recognition. A Powerful Approach to Interpreting Chemical Data. *J. Am. Chem. Soc.* **1972**, *94*, 5632–5639.

(36) Kowalski, B. R.; Bender, C. F. Pattern Recognition. II. Linear and Nonlinear Methods for Displaying Chemical Data. *J. Am. Chem. Soc.* **1973**, *95*, 686–692.

(37) Hyde, R. M.; Livingstone, D. J. Perspectives in QSAR: Computer Chemistry and Pattern Recognition. *J. Comput.-Aided Mol. Des.* **1988**, *2*, 145–155.
(38) Domine, D.; Devillers, J.; Chastrette, M.; Karcher, W. Non-linear Mapping for Structure-Activity and Structure-Property Modelling. *J. Chemom.* **1993**, *7*, 227–242.
(39) Kruskal, J. B. Multidimensional Scaling by Optimizing Goodness of Fit to a Nonmetric Hypothesis. *Psychometrika* **1964**, *29*, 1–27.
(40) Kruskal, J. B. Comments on a "Nonlinear Mapping for Data Structure Analysis". *IEEE Trans. Comput.* **1971**, *C-20*, 1614.
(41) Wish, M.; Carroll, J. D. Multidimensional scaling and its applications. In *Handbook of Statistics*; Krishnaiah, P. R., Kanal, L. N., Eds.; North-Holland Publishing Company: Amsterdam, 1982; Vol. 2, pp 317–345.
(42) Klein, R. W.; Dubes, R. C. Experiments in Projection and Clustering by Simulated Annealing. *Pattern Recognit.* **1989**, *22*, 213–220.
(43) Valko, K.; Cserhati, T.; Forgacs, E. Comparative Investigations of the Retention Behaviour of Nucleoside Derivatives on Alumina Stationary Phases in Thin-Layer Chromatography and High-Performance Liquid Chromatography. *J. Chromatogr.* **1991**, *550*, 667–675.
(44) Leahy, D. E.; Morris, J. J.; Taylor, P. J.; Wait, A. R. Membranes and their Models: Towards a Rational Choice of Partitioning System. In *QSAR: Rational Approaches to the Design of Bioactive Compounds*; Silipo, C., Vittoria, A., Eds.; Elsevier: Amsterdam, 1991; pp 75–82.
(45) Abraham, M. H.; Duce, P. P.; Prior, D. V.; Barratt, D. G.; Morris, J. J.; Taylor, P. J. Hydrogen Bonding. Part 9. Solute Proton Donor and Proton Acceptor Scales for Use in Drug Design. *J. Chem. Soc., Perkin Trans. 2* **1989**, 1355–1375.
(46) Domine, D.; Devillers, J.; Garrigues, P.; Budzinski, H.; Chastrette, M.; Karcher, W. Chemometrical Evaluation of the PAH Contamination in the Sediments of the Gulf of Lion (France). *Sci. Total Environ.* **1994**, in press.
(47) Devillers, J.; Thioulouse, J.; Domine, D.; Chastrette, M.; Karcher, W. Multivariate Analysis of the Input and Output Data in the Fugacity Model Level I. In *Applied Multivariate Analysis in SAR and Environmental Studies*; Devillers, J., Karcher, W., Eds.; Kluwer Academic Publishers: Dordrecht, 1991; pp 281–345.
(48) Domine, D.; Devillers, J.; Chastrette, M.; Karcher, W. Multivariate Structure-Property Relationships (MSPR) of Pesticides. *Pestic. Sci.* **1992**, *35*, 73–82.
(49) Devillers, J.; Karcher, W.; Chastrette, M.; Domine, D. Multivariate Structure-Environmental Fate Relationships for Chlorinated Chemicals. *J. Chim. Phys.* **1992**, *89*, 1703–1708.
(50) STATQSAR Package, CTIS, Lyon, France, 1993.
(51) Thioulouse, J. MacMul and GraphMu: Two Macintosh Programs for the Display and Analysis of Multivariate Data. *Comput. Geosci.* **1990**, *16*, 1235–1240.
(52) Panthananickal, A.; Hansch, C.; Leo, A. Structure-Activity Relationship of Aniline Mustards Acting against B-16 Melanoma in Mice. *J. Med. Chem.* **1979**, *22*, 1267–1269.