

Molecular Dataset Diversity Indices and Their Applications to Comparison of Chemical Databases and QSAR Analysis

Alexander Golbraikh*

Latvian Institute of Organic Synthesis, 21 Aizkraukles Street, Riga LV-1006, Latvia

Received July 12, 1999

A new mutual molecular dataset diversity index (MMDDI), individual molecular dataset diversity index (IMDDI), and volume ratio (VR) are proposed to assess molecular dataset diversity. MMDDI and IMDDI can serve as valuable instruments for selecting monomer pools for combinatorial synthesis and in decision making about acquiring new databases. MMDDI can also be used as one of the criteria to estimate the quality of quantitative structure–activity relationship (QSAR) models aimed at the prediction of biological activities. The indices can be calculated directly from molecular descriptor values. The procedures applied for MMDDI and IMDDI calculations allow one to automatically compile lists of compounds, which can simplify molecular diversity analyses and database searching. The information can also be used for forming training and test sets in QSAR analysis.

1. INTRODUCTION

One of the most important problems in chemical database analysis and quantitative structure–activity relationship (QSAR) is molecular diversity of databases or datasets. Contemporary methods used in searching for new biologically active compounds make use of combinatorial libraries and high-throughput screening. The goal is to ensure there is as wide a range of structures as possible involved in the studies.^{1,2} Computer-based approaches were developed recently for selecting pools of starting monomers for combinatorial synthesis.^{1–4} If for a broad screening project³ several possible pools are selected, it is necessary to estimate their molecular diversity.

As outlined elsewhere,^{5,6} pharmaceutical companies search large chemical libraries to find new lead compounds. Acquiring a new database can be very costly, so a preliminary analysis of molecular diversity between this database and ones already possessed by a company is desired.

Another area where the problem of molecular diversity can arise is QSAR analysis. A QSAR model to predict biological activities for a series of compounds is developed by using a training set. To examine the robustness of the model, a test set of compounds is used.⁷ The wider the class of compounds, for which the model can correctly predict their activities, the higher the quality of the model is. Therefore, an estimation of molecular diversity of the test set with respect to the training set is important.

A method recently applied to molecular database diversity studies⁵ is the self-organizing map (SOM),^{8,9} also known as the Kohonen neural network. SOM allows visualizing the distribution of compound representative points in a multidimensional descriptor space by projecting them onto a two-dimensional (2D) map while preserving the orders of distances between them.^{8,9} This approach allowed comparison

of two databases containing only 1% of identical active compounds and explained the distribution of representative points on the map by structural differences of the molecules.⁵ The method has been shown to be useful in molecular database diversity studies and helpful in decision making about purchasing new chemical databases.⁵

As with any projecting technique, SOM implementation leads to some distortions in the relative positions of representative points on a 2D map in comparison with the original points in a multidimensional descriptor space. Additional distortions are brought in because SOM is a nonlinear method.^{10,11} It does not provide also a *quantitative* criterion of molecular database diversity, which could be, for instance, more suitable in deciding to acquire new databases.

Another method for comparing two datasets could be based on the cellular approach proposed by Pearlman and Smith.^{12,13} Suppose there are two datasets of compounds: dataset 1 (*reference dataset*) and dataset 2. Each compound is represented by a point in a multidimensional descriptor space. Each dimension in the descriptor space is divided into a certain number of bins. Thus, the whole descriptor space area occupied by representative points of compounds is divided into cells. Let N_2 be the total number of compounds in dataset 2, and N_{diff} be the number of representative points of dataset 2 compounds in the cells where are no representative points of dataset 1 compounds. Then the diversity of dataset 2 with respect to dataset 1 could be defined as the ratio N_{diff}/N_2 . The drawbacks of the cell-based approach are as follows. (i) It can work only for large datasets and descriptor spaces of low dimensionalities, since dividing each metric into n bins results in n^K number of cells, where K is the space dimensionality. So if $n \sim 10$, K must not exceed 5 or 6.^{12,13} (ii) Sometimes, points that fall into different (adjacent) cells can represent similar compounds. (iii) The cubic cell diagonal in K -dimensional space is equal to $K^{1/2}$. Therefore, the representative points separated by a distance almost equal

* Current address: School of Pharmacy, University of North Carolina, Chapel Hill, NC 27599-7360.

to $K^{1/2}$ can belong to one cell (thus the corresponding compounds are being treated as similar), or they can belong to different cells, even nonadjacent ones.

Another possibility for defining molecular diversity of dataset 2 with respect to dataset 1 could be based on clustering. Again, let N_2 be the total number of compounds in dataset 2 and N_{diff} be the number of compounds of dataset 2 in the clusters in which there are no representative points of dataset 1 compounds. Then the diversity of dataset 2 with respect to dataset 1 can be defined as N_{diff}/N_2 . The main drawback of this approach is that, generally, all clustering procedures give clusters occupying nonequal volumes in the multidimensional descriptor space or the borders between clusters do not correspond to the "natural" borders between occupied areas. The representative points of a big cluster (i.e. occupying a large volume) do not obligatorily correspond to similar compounds (i.e. the distances between representative points of which are small).

In this paper, a *quantitative* criterion of molecular diversity of one molecular dataset with respect to another (reference dataset), free of the drawbacks of the cell-based and cluster approaches, is proposed. It is referred to as a mutual molecular dataset diversity index (MMDDI). The index can be calculated directly using the descriptor values or principal component analysis (PCA) scores.

A series of molecular diversity measures were defined for one dataset. In ref 14 Daylight database fingerprints were used to compare diversity among combinatorial libraries. Other approaches are based on similarity matrices.¹⁴ They were further developed in ref 15, where the diversity of a combinatorial library was obtained by the calculation of the mean pairwise intermolecular dissimilarity. Pairwise similarity measures can be based on BCUTs,^{12,13} or fingerprints,¹⁴ or cosine coefficients,¹⁵ or sets of topological indices,^{16,17} or molecular physicochemical properties,^{18,19} etc. While this method can be used for estimating molecular dataset diversity, it tells nothing about the clustering of compounds within it and voids in the descriptor space. To address this shortcoming, another criterion of molecular dataset diversity is proposed, which can be used in addition to the above-mentioned diversity measures. It is referred to as the individual molecular dataset diversity index (IMDDI).

Lists of compounds automatically created by the procedures elaborated for indices calculations can be used for database searching and in QSAR analysis. Applications of indices to database diversity studies and QSAR analysis will be briefly considered.

One more criterion of molecular diversity of one dataset with respect to another, the reference dataset, is introduced. It is a ratio of the volume in a multidimensional descriptor space occupied by representative points of the reference dataset to the total volume occupied by the representative points of both datasets. This index is referred to as a volume ratio (VR). To calculate VR more precisely and to improve the precision of the developed approaches, thorough investigation of the multidimensional descriptor space is necessary. Possible direct approaches for obtaining a good estimate of multidimensional volume occupied by the representative points will be also considered.

Some of the techniques developed in this work could also be useful for many other types of databases, where the objects

are characterized by a series of quantitative variables (properties): medicinal, physiological, sociological, etc., databases.

2. MOLECULAR DATASET DIVERSITY INDICES

A new molecular dataset diversity index is proposed to compare molecular datasets. A mutual molecular dataset diversity index (MMDDI) of dataset 2 with respect to dataset 1 (reference dataset) is defined as follows.

Let N_1 and N_2 be the number of compounds in datasets 1 and 2, correspondingly, so the total number of compounds in both datasets is $N = N_1 + N_2$. K molecular descriptors X_{ij} , $j = 1, \dots, K$, are calculated for all compounds $i = 1, \dots, N$ and scaled in some manner so that $X_{ij} \in [X_{\min,j}; X_{\max,j}]$, where $X_{\min,j}$ and $X_{\max,j}$ are the minimum and maximum values of the j th descriptor. The total volume occupied by the representative points of compounds in the descriptor K -dimensional space can be estimated as the product of lengths of intervals $[X_{\min,j}; X_{\max,j}]$, $j = 1, \dots, K$

$$V = \prod_{j=1}^K (X_{\max,j} - X_{\min,j}) \quad (1)$$

Equation 1 is not the exact formula, since the distribution of representative points is not uniform within this volume. For instance, there may be empty areas within it. Approaches to improve the estimate of the occupied volume will be discussed in sections 6, 10, and 11. The procedures proposed in those sections can be, however, rather time-consuming.

The defined K -dimensional region is, in fact, a K -dimensional rectangular parallelepiped. If the normalization is performed according to formula 6a (see section 3), all $X_{\min,j} = 0$ and all $X_{\max,j} = 1$, it will be referred to as a K -dimensional descriptor cube. In some calculations principal components (PCs) were used instead of descriptors. In these cases, the multidimensional region will be referred to as a PC parallelepiped. The average volume corresponding to one point is $L^K = V/N$ (if one imagines it as a K -dimensional cube, the length of its side would be L). For each point representing a compound from dataset 2, a K -dimensional sphere with a radius

$$R = cL = c(V/N)^{1/K} \quad (2)$$

is constructed with the center in this point (Figure 1), where c is a coefficient (it is referred to as a "dissimilarity level", DL). A compound from dataset 2 is considered as being dissimilar from all the compounds of dataset 1 with the DL equal to c , if within the corresponding sphere there are no representative points of dataset 1 compounds. If the total number of compounds of dataset 2 dissimilar from all the compounds of dataset 1 is N_{diff} , the MMDDI is

$$\text{MMDDI}_{2,1}(c) = N_{\text{diff}}/N_2 \quad (3)$$

It must be emphasized that MMDDI is nonsymmetrical: generally, the diversity of some dataset 2 with respect to dataset 1 is not equal to the diversity of dataset 1 with respect to dataset 2. It can be easily understood from an example shown in Figure 2. Dataset 2 contains compounds identical or similar to those from dataset 1 as well as the dissimilar ones. In this case, $\text{MMDDI}_{2,1} > 0$, while $\text{MMDDI}_{1,2}$ can be

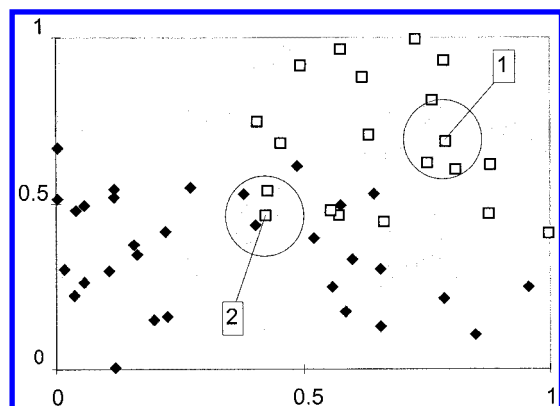


Figure 1. Mutual molecular dataset diversity index (MMDDI) definition. Let the descriptor space be two-dimensional. Compounds belonging to dataset 1 and dataset 2 are represented by black diamonds and white squares, respectively. Around each point representing a compound of dataset 2 a hypersphere with radius R (see section 2) is constructed (two of these hyperspheres which in this case are circles are shown). Compound 1 of dataset 2 is dissimilar from all the compounds of dataset 1, because there are no points representing compounds from dataset 1 within the circle with the center in point 1. At the same time, compound 2 of dataset 2 is similar to some compounds of dataset 1. If the total number of compounds in dataset 2 is N_2 , and the number of compounds in dataset 2 dissimilar from all of the compounds of dataset 1 is N_{diff} , then $\text{MMDDI}_{2,1} = N_{\text{diff}}/N_2$.

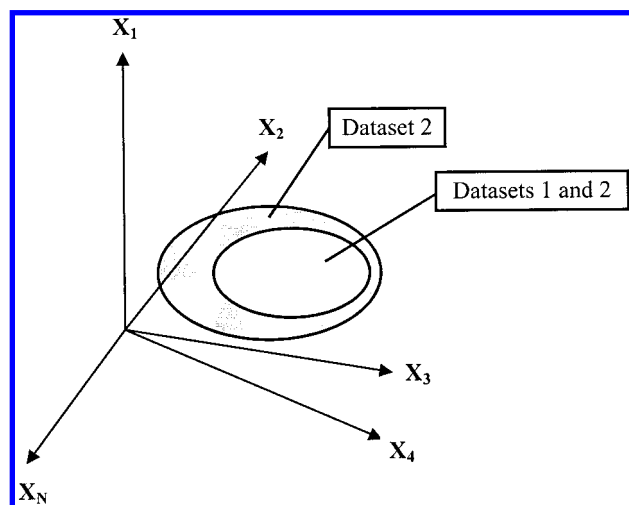


Figure 2. Nonsymmetry of MMDDI. The light gray area in a multidimensional descriptor space (X_1, X_2, \dots, X_N are descriptors) is occupied by points representing compounds belonging to both datasets 1 and 2, and the dark gray area, only by points representing dataset 2 compounds. If R (see section 2) is smaller than the maximum distance between the boundaries of both areas, $\text{MMDDI}_{2,1} > 0$. If at the same time no one compound of dataset 2 represented in the light gray area is dissimilar from all of the compounds of dataset 1, $\text{MMDDI}_{1,2}$ is equal to zero.

equal to zero. If $\text{MMDDI}_{1,2} > \text{MMDDI}_{2,1}$, one can say that the molecular diversity of dataset 1 with respect to dataset 2 is higher than that of dataset 2 with respect to dataset 1. If structures contained in datasets 1 and 2 are, for instance, possible building blocks for combinatorial synthesis, the first one in terms of molecular diversity is preferable. By calculating MMDDI, all compounds in dataset 2 dissimilar from those in dataset 1 can be automatically found. For each compound of dataset 2, a list of compounds of dataset 1 similar to it can also be automatically obtained. These data are important for further molecular diversity analysis.

MMDDI is a mutual index, dependent on two datasets. Another useful index is defined for one dataset. The individual molecular dataset diversity index (IMDDI) is defined as the part of compounds dissimilar (with a predefined DL) to all the other compounds in a dataset. Let n be the number of compounds in a dataset. For all representative points, K -dimensional spheres with centers in these points are constructed, their radius being calculated according to eq 2, in which n instead of N is used. A compound is considered dissimilar from all the other compounds, if within the corresponding sphere there are no points representing other compounds. If the number of dissimilar compounds is n_{diff} ,

$$\text{IMDDI}_1(c) = n_{\text{diff}}/n \quad (4a)$$

If DL is too small, and all compounds appear to be dissimilar from each other, formula 4b is to be used (see below). It is possible to find clusters of two, three, etc., compounds dissimilar from all of the other compounds in a dataset and define the corresponding IMDDI_2 , IMDDI_3 , etc. using formula 4a. If all clusters contain an equal number of compounds, formula 4b is to be used. In these cases, n_{diff} will be the number of clusters containing two, three, etc., representative points. By definition, every cluster of dissimilar compounds makes a contribution of $1/n$ to IMDDI. Therefore, IMDDI can be defined as follows

$$\text{IMDDI}(c) = (n_{\text{clus}} - 1)/n \quad (4b)$$

where n_{clus} is the total number of clusters. One was subtracted from n_{clus} to make IMDDI be equal to zero, if there is only one cluster of representative points; i.e., all points constitute one cluster. Equation 4a is a particular case of eq 4b when only "clusters" containing one point are considered. A clustering procedure applied in this work will be described in section 6.

For each compound of a dataset, the procedure applied for IMDDI calculation can create a list of similar to it compounds. Lists of compounds belonging to each cluster can also be obtained. Thus, unique structures can be found in a dataset as well as the underrepresented ones. As far as acquiring a new database is concerned, a high IMDDI value obtained with a high DL could be a drawback of a database. It indicates that there are "holes" in the area of the descriptor space occupied by representative points. These holes can correspond to physically impossible structures, as they can suggest that some structures are missing in a database. As to QSAR analysis, the lists of compounds can be useful in forming training and test sets and in finding possible outliers. It could be interesting also to perform this diversity analysis for compounds contributing to MMDDI, i.e., compounds of dataset 2 dissimilar from all of the reference dataset 1 compounds (see above in this section).

Another value characterizing the diversity of database 2 with respect to database 1 is a volume ratio (VR). It is computed according to the formula

$$\text{VR}_1^K = V^{(1)}/V \quad (5a)$$

where V is the total volume occupied by points representing both datasets in the multidimensional descriptor space and $V^{(1)}$ is the volume occupied by points representing dataset

1. K stands for the dimensionality of the descriptor space. In this form, the index is dependent on dimensionality of the descriptor space. So the formula

$$VR_1 = (V^{(1)}/V)^{1/K} \quad (5b)$$

is supposed to be preferable. Of course, the VR can be calculated by using formula 1. The VR value thus obtained will be used in section 8. This method of calculation of VR is, however, rather inexact. Actually, if one or only a small part of compounds in a reference dataset have some of the descriptors much higher or lower than all the other compounds, the VR_1 value can be high, and at the same time, the MMDDI value of some other dataset with respect to this one can also be high. For use of this criterion, better estimates of the volume occupied by the representative points than those given by formula 1 are necessary. Approaches to obtain these estimates will be discussed in sections 10 and 11.

3. MOLECULAR DESCRIPTORS

Calculations of MMDDI, IMDDI, and RV can be performed for any set of numerical descriptors. Ideally, they must be carefully selected for each particular case. There were several methods developed to address this problem.^{20–25} It is, however, out of the scope of this paper. So the same set of descriptors was used for all the examples considered. These descriptors take into account topological, structural, and global molecular features. At the same time, many of these indices are highly correlated, as it follows already from their definition (see below). High correlation between descriptors is the general property of almost all real cases. The descriptors applied for calculations of indices introduced in this study recently have been successfully used for molecular diversity analysis of databases containing organophosphorus compounds.⁵ As in ref 5, all descriptor weights were equal to 1. The goal of this study is to demonstrate some possible applications of the proposed molecular diversity criteria and methods developed. The indices introduced were calculated for model databases obtained from examples considered in ref 5 and for training and test sets of some QSAR models developed elsewhere,^{26–29} as well as for construction of the Kohonen maps for these sets. The following 60 molecular descriptors were used: 20 Kier-Hall molecular connectivity indices³⁰ ($^0\chi$, $^1\chi$, $^2\chi$, $^3\chi_C$, $^3\chi_P$, $^4\chi_P$, $^4\chi_{PC}$, $^5\chi_P$, $^5\chi_C$, $^6\chi_P$, $^0\chi^v$, $^1\chi^v$, $^2\chi^v$, $^3\chi_C^v$, $^3\chi_P^v$, $^4\chi_P^v$, $^4\chi_{PC}^v$, $^5\chi_P^v$, $^5\chi_C^v$, $^6\chi_P^v$); numbers of paths and vertices with degrees 1–4, the half-sum of the off-diagonal elements of the distance matrix, the sum of degrees (I), the sum of squares of degrees (M1), the logarithm of the product of degrees, combination of I and M1 ($0.5M1^2 - I + 1$); Gutman³¹ and Platt³² indices; a series of information indices (IC^0 , SIC^0 , CIC^0 , IC^1 , SIC^1 , CIC^1 , IDW , IDW mean);^{33,34} the number of nitrogen, oxygen, and sulfur atoms in a molecule and a sum of paths between them, as well as a set of physicochemical characteristics (most of them were calculated using the structure additive approach^{35,36}) (molecular weight, molecular (van der Waals) volume, molecular refractivity, parachor, octanol–water partition coefficient,³⁷ molar volume at a normal boiling point, normal boiling point divided by critical temperature, enthalpy of atomization, heat capacity of liquid); and a set of electronegativity parameters (of a molecule by Sander-

son,³⁸ mean, variance, sum of mean square and variance, variation coefficient, and maximum value of electronegativity of atoms). The following formulas were used for the normalization of descriptors

$$X_{ij}^n = \frac{X_{ij} - X_{j,\min}}{X_{j,\max} - X_{j,\min}} \quad (6a)$$

$$X_{ij}^n = \frac{X_{ij} - \langle X_j \rangle}{\sigma_j} \quad (6b)$$

where X_{ij} and X_{ij}^n are the nonnormalized and normalized j th descriptor values for compound i ; correspondingly, $X_{j,\min}$ and $X_{j,\max}$ are the minimum and maximum values for the j th descriptor, $\langle X_j \rangle$ is the mean of the j th descriptor, and σ_j is its standard deviation:

$$\sigma_j = \left[\frac{\sum_{i=1}^n (X_{ij} - \langle X_j \rangle)^2}{n} \right]^{1/2}$$

4. SELF-ORGANIZING MAP

Kohonen SOM^{8,9} is a nonlinear method of projecting points from a multidimensional space onto a two-dimensional map. The main properties which make this method very valuable in molecular diversity and QSAR studies are preserving the order of distances between the representative points (points close to each other in the multidimensional descriptor space are projected onto points which are close to each other, or even one point) and high resolution (map is “afraid of void”, so the representative clusters of points occupy most of the map).⁵

SOM consists of the input and two-dimensional output (Kohonen) layers. The number of input layer nodes is equal to the number of descriptors. The number of the Kohonen layer nodes is equal to MN , where M and N are defined by a user. All the input layer nodes are connected to that of the Kohonen layer. The network is initialized by a random selection of weights W_{ij} ($i = 1, \dots, MN$; $j = 1, \dots, K$) for all the output layer nodes. The starting value of learning rate α and a size of a node neighborhood are also initialized (see below). The training process consists of the following steps.^{4,5}

- Each time descriptors X_{ij}^n (i is a compound number, $j = 1, \dots, K$) for one compound are presented to the input layer.
- The winning node is defined as that for which $D_i = \sum_{j=1}^K (W_{ij} - X_{ij}^n)^2$ is a minimum.
- Weights of the winning node and the neighborhood nodes are adjusted according to the formula:

$$W_{ij}(t+1) = (1 - \alpha)W_{ij}(t) + \alpha X_{ij}^n$$

where t is the step number. The process is continued until convergence is reached. α is reduced during the process. The neighborhood size is also usually reduced.

This algorithm was modified slightly by adding the so-called “conscience mechanism” to avoid a problem consisting of the fact that some nodes of the Kohonen layer can never win.^{9,38} This mechanism consists of a slight decreasing of a probability to win for the nodes winning too frequently and increasing it for the nodes winning too rarely.

To increase the resolution of the map, another improvement of the algorithm, namely, an interpolation option, was introduced.⁷ If it is used, the location of a representative point in the map is calculated by weighted averaging the coordinates of the three nodes closest to the original point (i.e. with the smallest D_i values), D_i values serving as weights.

5. PRINCIPAL COMPONENT ANALYSIS^{39,40}

It is well-known that molecular descriptors, or variables, may be highly correlated with each other. A principal component analysis^{39,40} (PCA) (in its linear form) is a method of obtaining noncorrelated linear combinations of descriptors. The method consists of obtaining eigenvalues and eigenvectors of the correlation matrix. The eigenvectors are the principal components (PCs). They constitute a new orthonormalized basis in the multidimensional descriptor space (the old basis is defined by the descriptor coordinates). Each representative point in the multidimensional descriptor space is then projected onto these eigenvectors. The coordinates thus obtained are the PCA scores. The ratios of eigenvalues to the total sum of them multiplied by 100 are equal to the percentages of total dispersion explained by the corresponding PCs. Thus, the method provides the first PC Y_1 being defined as a normalized linear combination of descriptors, which accounts for the maximum dispersion of representative points. The second one, Y_2 , is defined by the maximum dispersion among all normalized linear combinations of descriptors, not correlated with Y_1 . The third one, Y_3 , is defined by the maximum dispersion among all normalized linear combinations of descriptors, not correlated with Y_1 and Y_2 , etc. If the number of descriptors is lower than the number of compounds, the total number of PCs is generally equal to the number of descriptors. Otherwise, it is equal to the number of compounds minus 1. Usually a much lower number of PCs accounts for a major part (typically, 95%) of variable dispersion, and only these ones are taken into account. In this paper, PCA was applied to reduce the number of descriptors and to manage a correlation problem: MMDDI values for some of the examples considered were also calculated using the representative point distribution in PC parallelepiped.

6. CLUSTERING

Clustering of compounds was based on a similarity matrix, the elements of which D_{ij} were equal to the Euclidean distances between the corresponding representative points in the multidimensional descriptor space:

$$D_{ij} = \left[\sum_{m=1}^K (X_{im} - X_{jm})^2 \right]^{1/2}$$

where X_{im} and X_{jm} ($i = 1, \dots, n$; $j = 1, \dots, n$; $m = 1, \dots, K$) are the m th normalized descriptor values or PCA scores for the i th and j th compounds, K is the number of descriptors or PCs, and n is the number of compounds. (Other frequently used similarity matrix definitions can be found in ref 41.) The distance between two subsets of points is defined as the minimum distance between two points, one point belonging to one subset and the other point to the other subset. The following cluster definition was used in this work: a cluster is a set of points, the minimum distance

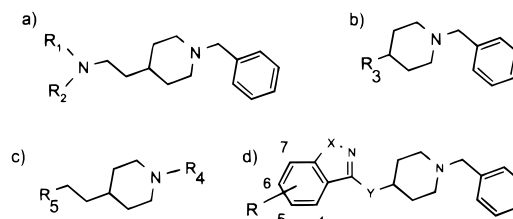


Figure 3. Acetylcholinesterase inhibitors used for the development of QSAR models.^{26,27} R_1 contains a benzoyl moiety, R_3 contains phthalimide or a similar group, and R , R_2 , R_4 , and R_5 are different other substituents. In the first model, compounds with structures *a*, *b*, and *c* constituted the training set, and compounds with structures *d* constituted the test set. In the second model, both the training and test sets included compounds from groups *a* and *b*.

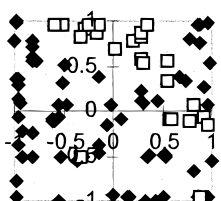
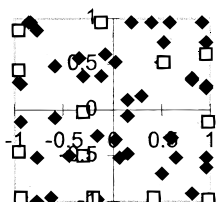
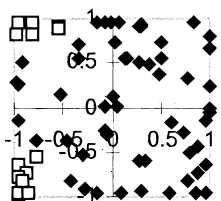
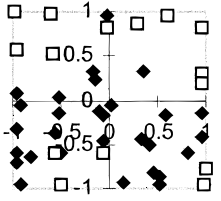
between any two subsets of which is not exceeding a predefined value. This predefined value R was calculated according to eq 2. Practically, the clusters were obtained by union of lists of compounds similar to each compound in a dataset (see section 2). In fact, this clustering procedure gives the results equivalent to a single linkage clustering.⁴⁰ The advantage of this procedure is that for each compound the lists of similar compounds are formed. Clustering was used in this work to calculate IMDDI values. It can also be used to obtain a better estimate of a total volume occupied by representative points. For bigger clusters, eq 1 can be used, while for the smaller ones a method described in section 11 can be applied.

Single linkage clustering can assign points that are far away to one cluster, if there is a chain of close to each other points connecting them.⁴⁰ Thus, one occupied area in the descriptor space corresponds to one cluster. High distances between clusters indicate that there are voids in the area of the descriptor space occupied by representative points. These voids can correspond to physically impossible structures, since they can be evidence that some structures are missing in a database. Therefore, in some cases, the high number of clusters (and high IMDDI value) in combination with the high distances between them can be a database drawback.

7. APPLICATIONS OF MMDDI TO QSAR ANALYSIS

Four examples concerning molecular diversity of training and test sets selected for a QSAR model development were studied. In the first example,²⁶ all the training set compounds belonged to 1-benzyl-4-[2-(*N*-benzoylamino)ethyl]piperidine derivatives, or contained a phthalimide group instead of a benzoyl one, while all the test set compounds were *N*-benzylpiperidine-benzisoxazole derivatives (Figure 3). So the training and test set compounds have rather different chemical structures, and it is not surprising that the areas in the Kohonen map occupied by the corresponding representative points are well-separated (Table 1). At the same time, in example 2,²⁷ the test set included every fifth compound of the first 66 ones (groups *a* and *b* in Figure 3) from the training set,²⁶ while the remaining 53 compounds constituted the training set. In this case, the molecules belonging to the same series of compounds were included in both the training and the test sets, so the corresponding areas in the Kohonen map were not separated (Table 1). In example 3, substance P inhibitors bound to Neurokinin-1 receptors were considered.²⁸ A very diverse set of compounds was selected as the training set. Test set compounds belonging to two other series

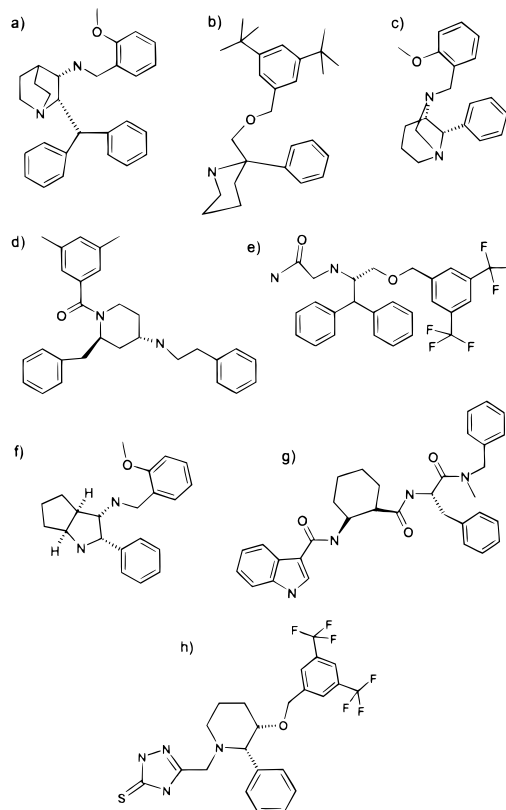
Table 1. Mutual Molecular Database Diversity Index (MMDDI) Values Calculated for Test Sets with Respect to Training Sets Used in QSAR Models Development^a

no.	comps in the training set	comps in the test set	Kohonen map	MMDDI	ref
1	82	29		0.83	26
2	53	13		0.15	27
3	73	18		0.94	28
4	33	16		0.69	29

^a Coordinates of representative points in the Kohonen maps were used. In all calculations the dissimilarity level equals 1. Black diamonds and white squares denote representative points of training and test set compounds, respectively.

appeared to be dissimilar to those included in the training set (Figure 4). It explains the excellent separation between areas in the Kohonen map occupied by the representative points of compounds belonging to the training and test sets (Table 1). In the last example, a QSAR model for inhibitory activity prediction for a series of HIV protease inhibitors was considered.²⁹ The training and test set compounds had rather different chemical structures (some of the molecules are shown in Figure 5), and, as it was expected, the corresponding areas in the Kohonen map were relatively good separated (Table 1).

Several kinds of calculations were performed. First, MMDDI values were obtained by using coordinates of the representative points in the Kohonen maps. SOM was applied to molecular descriptors normalized according to eq 6a. In these calculations, DL = 1. The results are presented in Table 1. The results correlate with what is seen in the maps: the better the areas occupied by representative points of compounds belonging to the training and test sets are separated, the higher the MMDDI value is. Then, MMDDI values were calculated directly using molecular descriptors (DL was also equal to 1) normalized according to (6a). MMDDI values for these cases appeared to be different from those in the first calculations. Thus, in examples 1 and 2, they were equal

**Figure 4.** Some substance P antagonists used for the development of a QSAR model.²⁸ Some structures included into the training set (a–f) and test set (g and h) are presented.

to 0.14 and 0.15, respectively. So, in example 1, for which there is a high molecular diversity of the test set with respect to the training set, the MMDDI value appeared to be lower than that for example 2, for which the molecular diversity is low. Incorrect results were also obtained for examples 3 and 4, the MMDDI values being 0.56 and 0.63, correspondingly. Therefore, such a direct use of MMDDI can lead to incorrect results. It can be partially explained by the dependence of MMDDI on DL that, in turn, is dependent on the dimensionality of the descriptor space. If the conclusions from calculations with a certain DL are correct for a space with one dimensionality, it does not yet mean they will be correct for a space with a different dimensionality. For instance, for a given number of compounds, the radius of the probe sphere with DL = 1 obtained according to eq 2 may be only slightly lower than the length of the multidimensional descriptor cube side, if the dimensionality of the descriptor space is big enough. Thus, in MMDDI calculations in high dimensional spaces lower DL values must be used. In the case of the 49 HIV-1 protease inhibitors the probe sphere radius calculated with DL = 1 was 0.937, and for 111 *N*-benzylpiperidines it was 0.925. Therefore, a more sophisticated procedure was developed. It consists of varying the DL with a small step from a small number to that one for which the MMDDI value is zero. Then some high (for instance, 0.9) MMDDI value is selected, and for each case the DL at which it is reached is determined. The higher the DL is, the higher the diversity of a dataset under study with respect to the corresponding reference dataset is. Results of this procedure implemented for each of the four cases are presented in Figure 6. Another way is to select some small DL value at which all MMDDI values are

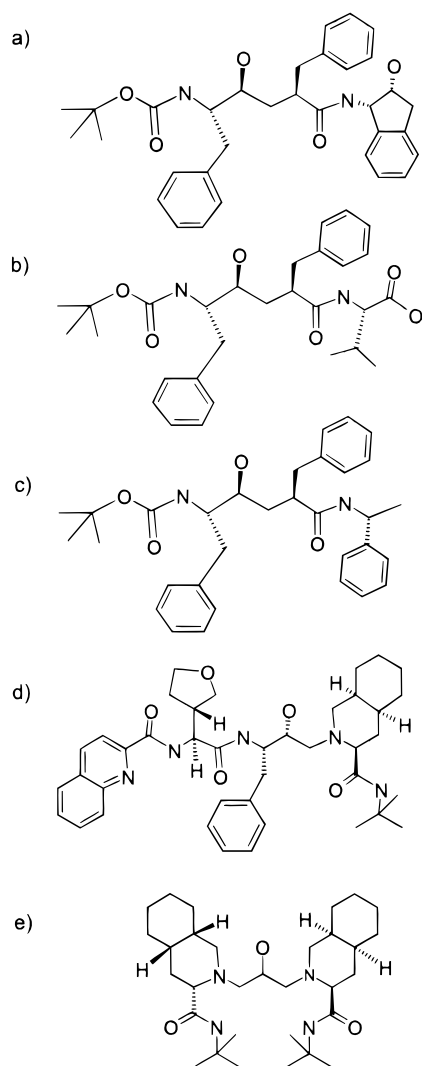


Figure 5. Some HIV-1 protease inhibitors used for the development of a QSAR model.²⁹ Some typical compounds included into the training set (a–c) and into the test set (d and e) are shown.

different. The higher the MMDDI is, the higher the molecular diversity of one dataset with respect to the corresponding reference dataset is. Since for all of the examples considered the similar K -dimensional sphere radii were obtained, the use of the same DL level for all of them was justified. Using this approach, the results were obtained which appeared to coincide with the expected ones. Nevertheless, this approach is also not completely consistent. If, for instance, one or a small part of the compounds had some of the descriptors much higher or lower than those of all of the other compounds, all the other representative points would occupy only a small part of the descriptor parallelepiped. Thus, if it is the case, the comparison of MMDDI values calculated for different datasets (test sets) with respect to different reference datasets (training sets) with the same DL values will be incorrect. Thus, it is not always possible to compare different QSAR models, using this approach directly. Therefore, in this case additional improvement to the whole procedure is necessary. Namely, the real volume occupied by the representative points must be found before MMDDI calculations: eq 1 is not exact enough. Possible procedures will be theoretically considered in detail in sections 10 and 11. The simple approach considered here is valid, if different datasets are compared with the same reference dataset. For

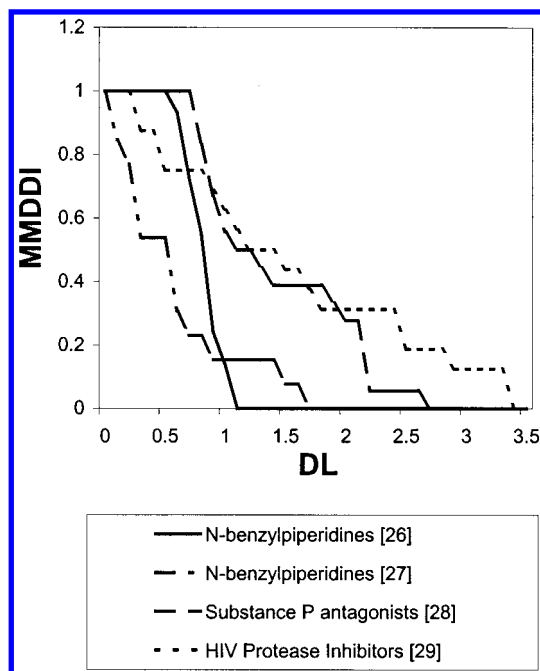


Figure 6. MMDDI values for test sets with respect to training sets used for the development of QSAR models (section 7). Calculations were performed using 60 descriptors (see section 3).

instance, in example 2 (Table 1) the training and test sets contained 53 and 13 *N*-benzylpiperidines, respectively. The MMDDI value was 0.54 (for $DL = 0.3$). Comparison of the test set consisting of 29 *N*-benzylpiperidines from example 1 (Table 1) with the same training set of 53 compounds gave MMDDI values equal to 1 until DL was equal to 0.5. Thus, much higher molecular diversity of the last test set in comparison with the first one, with respect to the same training set from example 2, was confirmed. All 91 substance P inhibitors²⁸ (training set and test set compounds together) and all 49 HIV-1 protease inhibitors²⁹ (training and test set compounds together) were also compared with these 53 *N*-benzylpiperidines. MMDDI values up to $DL = 0.4$ and $DL = 1.2$, respectively, were equal to 1. Obviously, there is high molecular diversity of these sets with respect to 53 *N*-benzylpiperidines, and these results confirmed it. The calculations with the descriptors normalized according to (6b) gave similar results for all the examples considered.

Another reason it was impossible to correctly estimate the DL value in the K -dimensional space is the correlation between descriptors: in fact, the representative points occupy only a subspace of a lower than K dimensionality. For all the examples, it was found that more than 95% of the total variance of descriptor values could be explained by only six PCs. So in fact the points are distributed in a six-dimensional PC space. The small derivations of the absolute majority of the representative points out of this subspace can be neglected. The outliers, if any, can be easily found. So to manage the descriptor correlation problem, PCA can be used prior to MMDDI calculations, and MMDDI could then be obtained by using the representative point coordinates in this PC space (the PCA scores). These calculations were also performed for the examples considered above. Six principal components were taken into account. For instance, test sets containing 29 and 13 compounds (examples 1 and 2 in Table 1) were compared with the same training set of 53 compounds (example 2 in Table 1). MMDDI values were 0.97

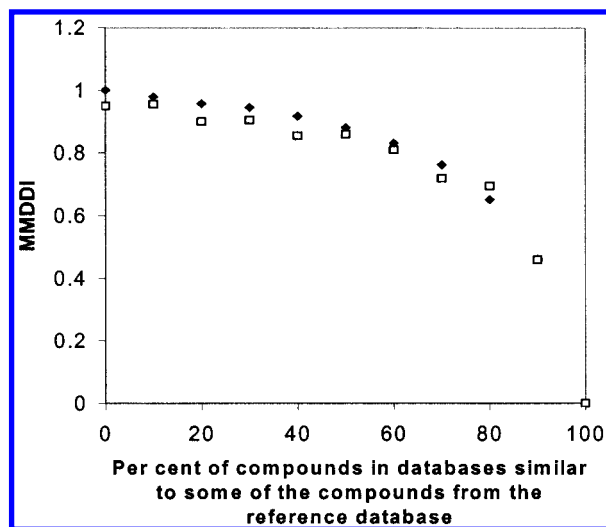


Figure 7. MMDDI values for 11 databases containing a different number of compounds (black diamonds) and 11 databases containing the same number of compounds (white squares) calculated with respect to the same reference database (see section 8).

and 0.39, respectively, for $DL = 0.3$. Again, these results coincide with the previous findings.

Nevertheless, the problem of unevenness of point distribution within a PC parallelepiped remains. It will be discussed in detail in sections 10 and 11.

8. APPLICATIONS OF MMDDI TO DATABASES

To consider the applications of MMDDI to molecular diversity of databases, the following examples were constructed by using 1251 organophosphorus compounds,⁵ 98 compounds of which were commercial pesticides, and the remaining 1153 compounds were taken from the other source.⁵ At first, 98 compounds were considered as a reference database. The structures of the 1153 compounds were selected, which appeared to be similar to at least one compound of the reference database with $DL = 0.3$. Then 11 database models were constructed out of the set of 1153 compounds. In the first one the structures similar to some of the 98 compounds were included. In the second to the tenth databases, the percentage of the similar compounds to some of the 98 ones was decreasing from 90 to 10% (the compounds for these databases were randomly selected). At last, for the eleventh database, only dissimilar structures to that of 98 were included. In all cases, DL was equal to 0.3. For these databases, MMDDI values were calculated, the 98 compounds being used as the reference database. The results are presented in Figure 7. They are as expected. The lower the part of similar compounds is to those in the reference database, the higher the MMDDI value is. Another set of 11 databases was created by using the random selection of 100 compounds from each of the first 11 databases. For this new set of databases, MMDDI values were also computed, and the results are presented in Figure 7. It can be seen that, in both cases, the results are similar. This means that MMDDI values are almost independent of the size of a database, if compounds of all main subgroups are represented in it. In the examples considered, the presentation of all main groups in the corresponding databases is guaranteed by the random selection. (If, for instance, there are 1000 compounds in a database, then a subgroup of 50 compounds (that

comprises 5% of all the compounds) localized in a small part of a descriptor parallelepiped can be considered as significant. The probability that no one compound of these 50 will be selected in 100 trials is equal to 0.0045.) Thus, it is possible to compare MMDDI values obtained for databases containing a different number of compounds. It is worth noting that the descriptors in these calculations were not renormalized for each particular case. (The normalization was performed for all of the 1251 molecules.) Different results were obtained, if before MMDDI calculations they were renormalized again according to (6a). It can be explained by the changes of distances between the representative points. Thus, the VR for the reference database containing 98 compounds without renormalization was equal to 0.326, while in the case of renormalization together with the database containing 100 compounds similar to those of the reference database it was 0.972. So it is not surprising the MMDDI was changed from zero to 0.99 in this case. By the way, such a low VR of the above-mentioned database containing 1153 compounds with respect to the reference database also corroborates a high molecular diversity of this database with respect to the reference one. Therefore, in this case, the VR appears to serve as another quantitative criterion of molecular diversity. In fact, it cannot work in all cases. If, as was mentioned above, one or only a small part of compounds in a reference database have some of the descriptors much higher or lower than all of the other compounds, the VR value can be high, and the MMDDI value of some other database with respect to this one can also be high. So a more precise estimate of the volume occupied by representative points in a K -dimensional descriptor or PC parallelepiped is necessary to use this criterion. By using this volume value, DL in MMDDI and IMDDI (see below) calculations could also be estimated in advance. This problem will be theoretically considered in sections 10 and 11.

An additional, but not less important, preference of these methods of molecular dataset diversity investigations is the possibility of automatically obtaining the list of all the compounds in a database under study, which are dissimilar from those in the reference database, and the lists of all the similar ones together with the corresponding reference database compounds. This can significantly simplify the molecular dataset diversity analysis.

Thus, the following procedure for comparing databases with a reference one with regard to molecular diversity is suggested.

1. Compose a dataset containing all databases.
2. Select and calculate the appropriate molecular descriptors that are supposed to correlate with the interested compound activities or properties.
3. Normalize these descriptors in some way, for instance, using eq 6a or 6b. Optionally, perform PCA on the data obtained.
4. For each database under consideration, calculate MMDDI with respect to the reference database, using different DL values, and obtain the number of compounds dissimilar to those contained in the reference database. DL can be varied from a value near zero to the value that makes R from (2) one order of the diagonal length of descriptor parallelepiped. In fact, the software can automatically define the higher limit of DL values. Additionally, obtain the lists of compounds

dissimilar to all of those from the reference database and similar to each of them, if necessary.

5. Draw conclusions out of all these data, for instance, with respect to acquiring a new database. Of course, not only the MMDDI values must be taken into account but also a number of compounds in the examined databases, which are dissimilar to all the compounds of the reference database. Other considerations can concern the structures of dissimilar compounds, but this question is not covered in this paper.

9. IMDDI APPLICATIONS

Equations 4a,b define IMDDI. IMDDI₁ is a fraction of compounds in a dataset, which are dissimilar from all the other compounds (with the given DL). In fact, the distribution of representative points in a multidimensional descriptor space can be very irregular (not uniform). One good example is the same 1251 organophosphorus compounds⁵ already considered above (see Kohonen maps presented in ref 5). In other words, if there is a compound in the vicinity of the representative point of which there are no representative points of other compounds of a dataset, this compound has a unique structure, or the compounds with the similar structures are underrepresented in it. Using this approach, these compounds can be found automatically, thus making the whole molecular diversity analysis easier. Another application of this approach is QSAR analysis. Namely, it can be used in the selection of compounds for the training and test sets and possible outlier searching. For instance, IMDDI values with DL varying from 0.1 to 1.7 were obtained for 66 *N*-benzylpiperidines considered in the previous section. It was found that one compound was dissimilar from all of the other compounds until DL = 1.6. It was found²⁷ that this compound is the only outlier in the test set, for which the model was unable to predict correctly the inhibition activity. The other two clusters of compounds (see sections 2 and 6) at DL = 1.6 contained compounds from both the training and test sets.

10. SEARCH OF DESCRIPTOR SPACE

Additional improvements to the approach in question could be achieved by a more thorough study of the descriptor (or PC) space. The idea is as follows. Normally, within a descriptor or PC parallelepiped the representative points of compounds are not distributed uniformly. There are empty areas within the parallelepiped, which could be taken into consideration by subtracting their volume from the parallelepiped volume found with (1). Thus, the choice of DL used in a *K*-dimensional sphere radius calculation for the procedure described in section 2 can be justified. It will also be possible to correctly estimate VR (see section 2) and compare different QSAR models using the MMDDI values (see section 7). Consider a *K*-dimensional sphere, the radius of which is equal to the average distance between a random point and the closest to it point representing a dataset compound. The procedure is as follows. Many times, place randomly the center of this probe sphere within the descriptor parallelepiped (the center coordinates being distributed uniformly) and check for the existence of representative points within it. If the total number of checks is *N* and the number of times the sphere was not containing any representative point is *N*_{unocc}, the unoccupied volume can be

estimated as

$$V_{\text{unocc}} = (N_{\text{unocc}}/N)V \quad (7)$$

This approach will work, however, only if the radius of this sphere is small in comparison with the parallelepiped sides. In this case, it is possible to neglect the event when a part of the sphere is outside of the parallelepiped. Otherwise, to take into account the "sticking out of the parallelepiped" part of the sphere, the following method can be used. The neighboring to the descriptor or PC parallelepiped part of the space must be laid out with the identical parallelepipeds together with the representative points as a parquet. If a part of the sphere is out of the parallelepiped, the procedure must search also for the corresponding copies of representative points within the sphere. This procedure is computationally impossible, if the dimensionality *K* of the descriptor space exceeds 5 or 6 because the total number of adjacent parallelepipeds to the given one is $3^K - 1$. In some cases, even more additional parallelepipeds must be taken into consideration. (If the lengths of the *K*-dimensional parallelepiped sizes are of different orders, this laying out the space with the copies of them will not work. Consider, for instance, a three-dimensional rectangular parallelepiped with a very large length while the width and height are small and equal to each other. The points within this parallelepiped are distributed, in fact, only in one dimension. So, it is possible to name it as a *quasi-one-dimensional distribution* of points.)

Obviously, the condition when the probe sphere radius is small in comparison with the parallelepiped side is satisfied only if the number of compounds in a database is big enough. We are going to obtain the formula for calculation of the probe sphere radius. Simultaneously, the number of representative points, for which the sphere radius is small enough for not taking into account the adjacent parallelepipeds, will be estimated quantitatively. The volume of a *K*-dimensional sphere with radius *R* can be obtained by direct integration²⁴

$$V_K = U_K R^K \quad (8a)$$

where *U_K* is the volume of the sphere with the radius equal to 1:

$$U_K = \int \int \dots \int_{x_1^2 + x_2^2 + \dots + x_K^2 \leq 1} dx_1 dx_2 \dots dx_K = \frac{\pi^{K/2}}{\Gamma((K/2) + 1)} \quad (8b)$$

where *x_i* (*i* = 1, ..., *K*) are the integration variables and Γ is the gamma function. It can be shown that in the case of even *K* (*K* = 2*m*)

$$U_{2m} = \pi^m / m! \quad (8c)$$

and in the case of odd *K* (*K* = 2*m* + 1)

$$U_{2m+1} = 2 \frac{(2\pi)^m}{(2m+1)!!} \quad (8d)$$

At the same time, the aforementioned average distance between a random point and the closest to it representative point $\langle r \rangle$ could be estimated as follows. Consider a *K*-dimensional descriptor (or PC) space with the uniform

Table 2. Probe Sphere Radii Calculated According to Equation 11^a

no. of comps	dimensionality of the descriptor space															
	3	4	5	6	8	10	12	15	20	25	30	40	50	60	80	100
50	0.150	0.228	0.300	0.367	0.484	0.585	0.675	0.793	0.961	1.104	1.232	1.454	1.646	1.817	2.118	2.381
100	0.119	0.192	0.262	0.327	0.444	0.546	0.637	0.757	0.928	1.074	1.204	1.429	1.623	1.796	2.100	2.365
300	0.083	0.146	0.210	0.273	0.387	0.490	0.581	0.704	0.879	1.028	1.161	1.390	1.588	1.764	2.071	2.339
500	0.070	0.129	0.190	0.250	0.363	0.465	0.557	0.680	0.857	1.008	1.141	1.373	1.572	1.749	2.058	2.327
1 000	0.055	0.108	0.165	0.223	0.333	0.434	0.526	0.650	0.827	0.980	1.115	1.349	1.550	1.729	2.041	2.311
3 000	0.038	0.082	0.133	0.186	0.291	0.389	0.480	0.604	0.783	0.938	1.075	1.313	1.516	1.697	2.013	2.286
5 000	0.032	0.072	0.120	0.171	0.273	0.370	0.460	0.584	0.763	0.919	1.057	1.296	1.501	1.683	2.000	2.274
10 000	0.026	0.061	0.104	0.152	0.250	0.345	0.434	0.557	0.737	0.894	1.033	1.274	1.480	1.664	1.983	2.258
30 000	0.018	0.046	0.084	0.127	0.218	0.309	0.396	0.518	0.698	0.855	0.996	1.239	1.448	1.634	1.956	2.234
50 000	0.015	0.041	0.076	0.116	0.204	0.294	0.380	0.501	0.680	0.838	0.979	1.223	1.433	1.620	1.943	2.222
100 000	0.012	0.034	0.066	0.104	0.187	0.274	0.358	0.478	0.657	0.815	0.956	1.202	1.414	1.601	1.927	2.207

^a Representative points are distributed within a multidimensional descriptor cube with the volume equal to 1.

distribution of points within it with the density equal to n points within a volume equal to that of descriptor (or PC) parallelepiped. Then within each sphere with the volume equal to that of the parallelepiped there will also be n points on average. Radius R of this sphere can be obtained from formulas 8a–d. Namely, in the case of $K = 2m$ and $K = 2m + 1$,

$$R = \frac{(m!V)^{1/(2m)}}{\pi^{1/2}} \text{ and } R = \left[\frac{(2m+1)!!V}{2(2\pi)^m} \right]^{1/(2m+1)} \quad (9)$$

respectively. If, for instance, $V = 1$ and $K = 60$, as is the case in the examples considered above, then $R = 1.958$. A probability that there are no representative points within a sphere with radius $r < R$ and one and only one representative point within a thin layer with thickness dr bordering this sphere is

$$dp = n \left[1 - \left(\frac{r}{R} \right)^K \right]^{n-1} \frac{K}{R^K} r^{K-1} dr \quad (10)$$

A sphere surface area was obtained by differentiating (8a) by radius. Multiplying (10) by r and integrating from zero to R , the following formula was obtained

$$\langle r \rangle = R \frac{\Gamma(n+1) \Gamma(1+(1/K))}{\Gamma(n+1+(1/K))} = R \frac{n!}{(n+(1/K))(n-1+(1/K))\dots(1+(1/K))} \quad (11)$$

From (10), it is seen that in the case of high dimensionality, $\langle r \rangle$ is only slightly lower than R , if only n is not big enough. Using in (11) the Stirling formula for approximation of the gamma function for large x

$$\Gamma(x+1) = [2\pi x]^{1/2} (x/e)^x$$

and taking into account that $\Gamma(1+(1/K)) > 0.885$, it was shown that for large n the following condition is satisfied:

$$n > \left(0.885 \frac{R}{\langle r \rangle} \right)^K \quad (12)$$

A factor slightly exceeding 1 for large n was omitted in the right side of (12). The condition which must be satisfied is $\langle r \rangle \ll R$. If $\langle r \rangle \sim R/10$ is considered sufficiently small (for a descriptor cube it is still almost 0.2 of the cube side), $n \sim 8.85^K = 10^{0.947K}$, etc. In a real database, the number of

compounds can be 10^5 or even more. Thus, condition (12) can be satisfied for a number of descriptors no more than 5. Some results obtained by using eq (11) are presented in Table 2. Radius values were calculated according to (10). These results show that, in many real cases, especially when a number of descriptors used is high, the probe sphere radius is even greater than the descriptor cube side.

11. BETTER ESTIMATION OF THE OCCUPIED VOLUME

In case the cell-based approach^{12,13} is applicable, the occupied volume in the multi-dimensional descriptor space can be estimated by multiplying (1) by the ratio of the number of occupied cells to the total number of cells. The result will depend, however, on the total number of cells.

A general direct method for estimating the occupied volume in the multidimensional descriptor space can be based on the approach developed by Pearlman and Smith,¹² with the only difference being that the probe sphere radius must be calculated according to (9). The procedure is as follows. (i) Center the probe sphere on one of the representative points and mark all other representative points located within this sphere. (ii) Center the probe sphere on another representative point that is not yet marked, and mark all representative points within this sphere. (iii) Repeat the previous step, until all compounds have been marked. (iv) Calculate the total volume of all these spheres. This procedure is simple and fast: it is an $O(N)$ process,¹² where N is the number of compounds in a dataset. This approach has several drawbacks. (i) As in the previous section, if the number of compounds does not satisfy condition (12), the sticking of a part of the probe sphere out of the descriptor or PC parallelepiped must be taken into account. To obtain a better estimate, the neighborhood of the descriptor or PC parallelepiped must be laid out with the copies of this parallelepiped together with the representative points within it, as in the previous section. If a part of the sphere is out of the parallelepiped, the procedure must search also for the copies of representative points within this sphere. This method can be practically realized only if the number of descriptors (or PCs) does not exceed 5 or 6 (see the previous section). (ii) Overlapping of the probe spheres is not taken into account. In fact, it is mathematically difficult. Instead of the probe spheres, the probe cubes are to be preferably used, since it is easier to estimate the volume of overlapping areas of the cubes (see below). (iii) The result depends on the order in which the representative points are selected.

To overcome the drawbacks of the above approach, another method is proposed. It consists of constructing K -dimensional cubes with centers in representative points, the volumes of cubes being equal to V/N and sides parallel to the coordinate axes. (Here V is the volume obtained according to (1), and N is the total number of points.) Let V_{i_1,i_2} be the volume of the intersection of cubes i_1 and i_2 , V_{i_1,i_2,i_3} be the volume of the intersection of cubes i_1 , i_2 , and i_3 , etc. $V_{1,2,\dots,N}$ be the volume of the intersection of all the cubes. Then the total volume occupied can be defined as follows

$$V_{\text{occ}} = V - \sum_{1 \leq i_1 < i_2 \leq N} V_{i_1,i_2} + \sum_{1 \leq i_1 < i_2 < i_3 \leq N} V_{i_1,i_2,i_3} - \sum_{1 \leq i_1 < i_2 < i_3 < i_4 \leq N} V_{i_1,i_2,i_3,i_4} + \dots + (-1)^{N+1} V_{1,2,\dots,N} \quad (13)$$

If the intersection of some cubes is empty, the corresponding volume is equal to zero. The right side of (13) contains totally $2^N - N$ terms, so the direct use of the formula in this form is practically impossible. If the number of representative points is not big enough, as is the case in QSAR studies, it is possible, however, to calculate only the first two sums in (13) and neglect all the following terms. As a result, a better estimation of the total volume can be obtained than that obtained by applying eq 1, since the following sums in (13) represent only the fractions of the preceding sums. Thus, the total volume will be overestimated only by a fraction or the total of the volume represented by the third sum in (13) (the sum of the volumes of the intersections of four cubes). V_{occ} is dependent on V , as do the other terms in (13). So this method is expected to be less precise than that elaborated in the previous section. On the other hand, however, it is simpler and faster since it is not based on the Monte Carlo procedure.

This approach can be simplified (with some additional loss of precision) by preliminary clustering of the representative points, and by applying techniques described in section 2 for bigger clusters and eq 13 for the smaller ones. In the case of cluster overlapping, (13) can also be used for clusters.

12. CONCLUSIONS

A new mutual molecular dataset diversity index (MMDI), individual molecular dataset diversity index (IMDDI), and volume ratio (VR) were introduced. The indices can be calculated using descriptor values of molecules, or principal component analysis (PCA) scores, i.e., the coordinates of representative points of compounds in a multidimensional descriptor or PC space. MMDDI and VR are quantitative criteria of the diversity of one dataset (dataset 2) with respect to the other (reference) dataset (dataset 1). MMDDI is defined as a ratio of the number of compounds of dataset 2, in the vicinity of representative points of which there are no points representing compounds of dataset 1, to the total number of compounds in dataset 2. Neighborhood radius is defined as a root of degree K (where K is the number of descriptors or PCs) of the average volume in the descriptor space corresponding to one representative point, multiplied by a factor named the dissimilarity level (DL). The procedure developed allows one to automatically obtain a list of all the compounds in dataset 2 which contribute to MMDDI (that are dissimilar from all the compounds of the reference

dataset). For each compound of dataset 2, a list of similar to it compounds from dataset 1 can also be obtained. VR is defined as a ratio of the volume in a multidimensional descriptor space occupied by representative points of the reference dataset to the total volume occupied by the representative points of both datasets. IMDDI is defined for one dataset. It is the number of clusters of compounds represented in a dataset minus one divided by the number of compounds. The clustering procedure used for calculation of IMDDI is equivalent to the single linkage clustering. The distance between clusters is defined as the minimum distance between two points, one belonging to one cluster and the other to the other cluster. Distances between clusters are calculated. Additionally, for each compound a list of similar to it compounds can be automatically obtained, as well as lists of compounds included in each cluster.

Methods developed in this work are free from the drawbacks of the cell-based approach.^{12,13} IMDDI can be used together with other molecular dataset diversity criteria developed elsewhere (see, for instance, refs 14 and 15).

The main properties of the proposed indices that make them valuable in the molecular diversity studies are as follows.

1. The indices can be used to select starting pools of monomers for combinatorial synthesis.
2. The indices and the lists of compounds obtained by the procedures elaborated for MMDDI and IMDDI calculations can serve as valuable instruments in decision making about acquiring new databases.
3. IMDDI can be used as a criterion of fullness of structure representation in a database.
4. MMDDI and IMDDI and the lists of compounds obtained can be used for forming training and test sets to develop QSAR models and finding possible outliers.
5. MMDDI and VR can be used to estimate the robustness of QSAR models.

To estimate more precisely the volume occupied by the dataset representative points in the multidimensional descriptor space, several procedures were developed. These procedures can be applied for VR calculation as well as direct estimation of DL value used in MMDDI and IMDDI calculations. Most of these methods are applicable only if the number of descriptors or PCs does not exceed 5 or 6.

The approaches developed could be useful for studies of other kinds of databases, the objects of which are characterized by a series of different quantitative variables (properties). These could be databases used in medicine, physiology, economics, etc.

ACKNOWLEDGMENTS

All calculations of descriptors and Kohonen maps were performed during author's stay in the Laboratory of Chemometrics, Orléans University, France (head Prof. J. R. Chré-tien), in 1998–1999. The author is also grateful to Dr. D. Kireev, currently at Synthe Labo, Strasbourg, France, for a helpful discussion.

REFERENCES AND NOTES

- (1) Willett, P. Computational tools for the analysis of molecular diversity. *Perspect. Drug Discovery Des.* **1997**, 7/8, 1–11.

- (2) Snarey, M.; Terrett, N. K.; Willett, P.; Wilton, D. J. Comparison of algorithms for dissimilarity-based compound selection. *J. Mol. Graphics. Modell.* **1997**, *15*, 372–385.
- (3) Zheng, W.; Cho, S. J.; Tropsha, A. Rational Combinatorial Library Design. 1. Focus-2D: A New Approach to the Design of Targeted Combinatorial Chemical Libraries. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 251–258.
- (4) Cho, S. J.; Zheng, W.; Tropsha, A. Rational Combinatorial Library Design. 2. Rational Design of Targeted Combinatorial Libraries Using Chemical Similarity Probe and the Inverse QSAR Approaches. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 259–268.
- (5) Bernard, P.; Golbraikh, A.; Kireev, D.; Chretien, J. R.; Rozhkova, V. Comparison of Chemical Databases: Analysis of Molecular Diversity with Self-Organising Maps (SOM). *Analyst* **1998**, *26*, 333–341.
- (6) 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.
- (7) Austel, V. Experimental Design in Synthesis Planning and Structure–Property Correlations. In *Chemometrics Methods in Molecular Design*; van de Waterbeemd, H., Ed.; VCH: Weinheim, Germany, 1995; pp 49–62.
- (8) Kohonen, T. *Self-organization and Associative Memory*; Springer-Verlag: Berlin, 1988.
- (9) *Neural Computing*; NeuralWare, Inc.: Pittsburgh, PA, 1995.
- (10) Kireev, D. B.; Ros, F.; Bernard, P.; Chretien, J. R.; Rozhkova, N. In *Computer-Assisted Lead Findings and Optimization*; van de Waterbeemd, H., Testa, B., Folkers, G., Eds.; Verlag Helvetica Chimica Acta and Wiley–VCH: Basel, Switzerland, and Weinheim, Germany, 1997; pp 255–264.
- (11) Kireev, D. B.; Bernard, P.; Chretien, J. R.; Ros, F. Application of Kohonen Neural Networks in Classification of Biologically Active Compounds. *SAR QSAR Environ. Res.* **1997**, *8*, 93–107.
- (12) Pearlman, R. S.; Smith, K. M. Software for chemical diversity in the context of accelerated drug discovery. *Drugs Fut.* **1998**, *23* (8), 885–895.
- (13) Pearlman, R. S.; Smith, K. M. Metric Validation and the Receptor-Relevant Subspace Concept. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 28–35.
- (14) Martin, E. J.; Blaney, J. M.; Siani, M. A.; Spellmeyer, D. C.; Wong, A. K.; Moos, W. H. Measuring Diversity: Experimental Design of Combinatorial Libraries for Drug Discovery. *J. Med. Chem.* **1995**, *38*, 1431–1436.
- (15) Turner, D. B.; Tyrrell, S. M.; Willett, P. Rapid Quantification of Molecular Diversity for Selective Database Acquisition. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 18–22.
- (16) Basak, S. C.; Magnuson, V. R.; Regal, R. R. Determining Structural Similarity of Chemicals Using Graph-Theoretic Indices. *Discrete Appl. Math.* **1988**, *19*, 17–44.
- (17) Basak, S. C.; Grunwald, G. D. Tolerance Space and Molecular Similarity. *SAR QSAR Environ. Res.* **1995**, *3*, 265–277.
- (18) Downs, G. M.; Willett, P.; Fisanick, W. Similarity Searching and Clustering of Chemical Structure Databases Using Molecular Property Data. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 1094–1102.
- (19) Kearsley, S. K.; Sallamack, S.; Fluder, E. M.; Andose, J. D.; Mosley, R. T.; Sheridan, R. P. Chemical Similarity Using Physicochemical Property Descriptors. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 118–127.
- (20) Clementi, S.; Wold, S. How to Choose the Proper Statistical Method. In *Chemometrics Methods in Molecular Design*; van de Waterbeemd, H., VCH: Weinheim, Germany, 1995; pp 319–338.
- (21) Wold, S.; Eriksson, L. Statistical Validation of QSAR Results. In *Chemometrics Methods in Molecular Design*; van de Waterbeemd, H., VCH: Weinheim, Germany, 1995; pp 309–318.
- (22) Tetko, I. V.; Villa, A. E.; Livingstone, D. J. Neural network studies. 2. Variable selection. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 794–803.
- (23) Hoffman, B.; Cho, S. J.; Zheng, W.; Wyrick, S.; Nichols, D. E.; Mailman, R. B.; Tropsha, A. Quantitative structure–activity relationship modeling of dopamine D(1) antagonists using comparative molecular field analysis, genetic algorithms–partial least-squares, and *K* nearest neighbor methods. *J. Med. Chem.* **1999**, *42*, 3217–3226.
- (24) Cho, S. J.; Tropsha, A. Cross-Validated R^2 Guided Region Selection for Comparative Molecular Field Analysis (CoMFA): A Simple Method To Achieve Consistent Results. *J. Med. Chem.* **1995**, *38*, 1060–1066.
- (25) Hasegawa, K.; Kimura, T.; Funatsu, K. GA strategy for variable selection in QSAR studies: Application of GA-based region selection to a 3D-QSAR study of acetylcholinesterase inhibitors. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 112–120.
- (26) Bernard, P.; Kireev, D. B.; Chretien, J. R.; Fortier, P.-L.; Copet, L. Automated Docking of 82 N-benzylpiperidine Derivatives to Mouse Acetylcholinesterase and Comparative Molecular Field Analysis with “Natural” Alignment. *J. Comput.-Aided Mol. Des.*, in press.
- (27) Golbraikh, A.; Bernard, P.; Chretien, J. R. Validation of protein-based alignment in 3D Quantitative Structure–Activity Relationship with help of CoMFA models. *Eur. J. Med. Chem.*, in press.
- (28) Takeuchi, Y.; Shands, E. F. B.; Beusen, D. D.; Marshall, G. R. Derivation of a Three-Dimensional Pharmacophore Model of Substance P Antagonists Bound To The Neurokinin-1 Receptor. *J. Med. Chem.* **1998**, *41*, 3609–3623.
- (29) Pérez, C.; Pastor, M.; Ortiz, A. R.; Gago, F. Comparative Binding Energy Analysis of HIV-1 Protease Inhibitors: Incorporation of Solvent Effects and Validation as a Powerful Tool in Receptor-Based Drug Design. *J. Med. Chem.* **1998**, *41*, 836–852.
- (30) Kier, L. B.; Hall, L. H. *Molecular Connectivity in Structure–Activity Analysis*; Wiley: New York, 1986.
- (31) Gutman, I.; Ruscic, B.; Trinajstić, N.; Wilcox C. F., Jr. Graph theory and molecular orbitals. XII. Acyclic polyenes. *J. Chem. Phys.* **1975**, *62*, 3399–3405.
- (32) Sabljic, A. In *Practical Applications of Quantitative Structure–Activity Relationships (QSAR) in Environmental Chemistry and toxicology*; Karcher, W., Devillers J., Eds.; Kluwer: Dordrecht, The Netherlands, 1990; pp 61–82.
- (33) Basac, S. C. In *Practical Applications of Quantitative Structure–Activity Relationships (QSAR) in Environmental Chemistry and Toxicology*; Karcher, W., Devillers J., Eds.; Kluwer: Dordrecht, The Netherlands, 1990; pp 83–103.
- (34) Hansch, C.; Leo, A. *Substituent Constants for Correlation Analysis in Chemistry and Biology*; Wiley-Interscience: Chichester, U.K., 1979; p 13.
- (35) Kireev, V. A. *Methods of Practical Calculations in Thermodynamics of Chemical Reactions*. Moscow, Chimiya, 1975.
- (36) Stull, D. R.; Westrum, E. F., Jr.; Sinke, G. C. *The Chemical Thermodynamics of Organic Compounds*; Wiley: New York, 1969.
- (37) Sanderson, R. T. *Chemical bonds and bond energy*; Academic Press: New York, 1976.
- (38) DeSieno, D. Adding a Conscience to Competitive Learning. In *IEEE International Conference on Neural Networks*, San Diego, CA July 24–27, 1988, Vol. 1; 1988; pp 117–124.
- (39) Adams, M. J. *Chemometrics in Analytical Spectroscopy*; Royal Society of Chemistry: Cambridge, U.K., 1995.
- (40) *Computational Biometrics*; Nosov, V. N., Ed.; Moscow University Press: Moscow, 1990 (in Russian).
- (41) Willett, P.; Barnard, J. M.; Downs, G. M. Chemical Similarity Searching. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 983–996.
- (42) Fichtengolz, G. M. *A Course of Calculus; Differential and Integral Nauka: Moscow, 1969 (in Russian).*

CI990437U